A Novel Method for Face Track Linking in Videos

Macrina Lobo
Dept. of EEE
IIT Guwahati
Assam 781039, India
mmclobo@gmail.com

Mayank Pratap Singh
Dept. of EEE
IIT Guwahati
Assam 781039, India
mpsingh2604@gmail.com

Raghvendra Kannao
Dept. of EEE
IIT Guwahati
Assam 781039, India
raghvendra@iitg.ernet.in

Prithwijit Guha
Dept. of EEE
IIT Guwahati
Assam 781039, India
pguha@iitg.ernet.in

ABSTRACT

Face based video indexing and discovering co-occurrence patterns of faces are important components of any video analytics system. We propose a "self-supervised" system for identifying durations of scene presence of people by tracking their faces and linking the resulting face tracks. Multiple faces detected using the Viola-Jones detector are tracked across the frames in the TLD framework. Patches extracted from all the tracked face regions are subjected to spherical clustering to form a dictionary of representative patches. Face features are next extracted by concatenating arrays of inner products between face image patches and the dictionary elements. The features obtained from a certain face track are considered as positives (faces in same track belong to the same person) while the ones extracted from another co-occurring track are used as negatives (faces in concurrent tracks must belong to different people) to train SVMs. The faces in a new face track are classified by the trained SVMs and are linked to an existing track if the resulting likelihood for the corresponding SVM exceeds a certain threshold. The main contribution of this work lies in this proposal of SVM likelihood based linking of face tracks in videos. The performance analysis of the proposed system is presented on two news broadcast videos.

Keywords

Face Detection, TLD Tracker, Spherical Clustering, SVM, Track Linking

1. INTRODUCTION

With the large amount of video data at our disposal, using computers to analyze videos has become a necessity. Since people are an integral part of most videos, an analysis of the face occurrence patterns is a necessity in any efficient video analytics system. At a basic level, knowing the actors in a movie or sitcom, the news reporters or celebrities in a news broadcast enables us to index the video by the corresponding faces. Information such as which two faces occur together and for what duration over multiple videos can help us in deriving conclusions on associations of people. Even popularities of electoral candidates can be predicted by analyzing the visual media space occupied by them. If the person's face occurs frequently but relatively uniformly over several days, the person is probably popular or "trending" in the news. Such applications call for efficient methods for extracting faces in videos, tracking them over time and linking the extracted face tracks for analyzing scene presence of people.

Existing face based video indexing systems generally use the Viola-Jones face detector [6] along with mean-shift [5] or particle filter [2, 3] based trackers. Pande et. al. [5] has used the Viola-Jones based frontal face detector along with backward-forward tracking of multiple faces in a case based reasoning framework. The face instances of each track are grouped in an incremental clustering framework to discover the face modes. Two face tracks containing sufficiently similar set of modes are then linked to identify the multiple occurrences of the same person in different time durations. Zhang et al. [10] have proposed a similar work on photographs and used the assumption that no two faces in the same photograph can belong to a same person. This work was extended in [9] to perform face linking in videos. Apart from these, Markov random field based methods are also used in the literature to link multiple face tracks [7, 8].

We have also used the Viola-Jones detector to identify the face regions in images. These face regions are tracked further in the TLD framework [4] to extract the face tracks. TLD based methods have recently shown superior tracking performance compared to the traditional mean-shift or particle filtering based approaches. Patches of size 8×8 are randomly selected from all the extracted face tracks and are subjected to spherical (k-means) clustering to learn a dictionary of representative patches. Concatenated arrays of inner products between spatially ordered and non-overlapping patches with the dictionary elements form the feature vector of each face. Feature vectors collected from faces in the same track are marked as positives while the ones from another concurrent track are labeled as negatives for training SVM based classifiers. SVMs are thus trained over each

ICVGIP '14, December 14-18, 2014, Bangalore, India

http://dx.doi.org/10.1145/2683483.2683551 .

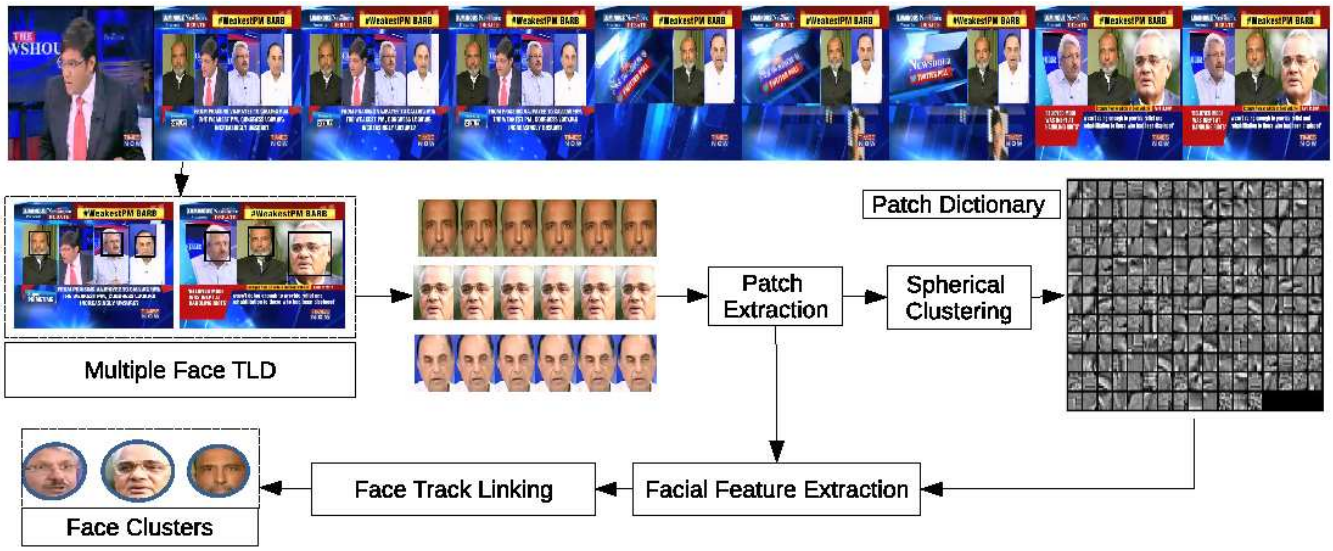

Figure 1: Illustrating the functional block diagram of the proposed system. The face regions detected by the Viola-Jones detector is filtered first to remove the false detections and are subjected to multiple object TLD based tracking. Patches extracted from all the tracked faces are grouped using spherical k-means clustering to form a dictionary of "mean patches". In a new face, spatially ordered non-overlapping patches are used further to represent any face as a spatially ordered collection of representative patches. The strengths of representation form the feature vector of a face. SVM based classifiers trained on feature vectors obtained from a track are used to classify other tracks and hence link the matching ones.

distinct face track and are used to identify the face tracks of the same person. The major contribution in this work is the proposal of the application of the pre-trained SVM likelihoods for linking new face tracks. A functional block diagram of the proposed system is shown in Figure 1.

The rest of the paper is organized in the following manner. Section 2 describes the methodology employed for detecting and extracting multiple face tracks from videos. Section 3 explains the process of dictionary learning using the patches collected from the extracted face tracks. Section 4 discusses the methodology of extracting face descriptors using the learned dictionary elements. Section 5 describes the methodologies of training SVMs over face tracks and the likelihood based identification of reappearance of face tracks. The results of experimentation on two news broadcast videos are presented in Section 6. Finally, we conclude in Section 7 and sketch the future extensions of the present work.

2. EXTRACTION OF FACE TRACKS

The Viola-Jones detector [6] is used to detect the faces in video frames. TLD based face trackers [4] initialized on the first detection of faces are used to track the face regions across the frames. Multiple face tracking is performed over two sets – first, the set of actively tracked faces ($F_a(t-1)$) till the $(t-1)^{th}$ instant and second, the set of faces detected ($F_d(t)$) in the t^{th} instant. The multiple face tracker initializes with an empty set $F_a(0)$ and keep on adding/removing faces as they appear or disappear/exit from the scene. We use the fractional overlap measure $\gamma_o(A,B) = \frac{|A \cap B|}{|A|}$ to estimate the fraction of the region A overlapping with the region B and is used to take decisions on maintaining sets of face trackers. The three different conditions for tracker initialization, continuation and termination are described as follows.

Identifying New Face – A new face $f_d \in F_d(t)$ does not have any overlap with the tracked faces from the last instant i.e. $\forall f_a \in F_a(t-1) \gamma_o(f_d, f_a) < \eta$

Retaining a Tracker – A face tracker is retained if the prediction confidence is high and/or it has overlap with a detected face region i.e. $\exists f_d \in F_d(t) \gamma_o(f_a, f_d) > 1 - \eta$. In this case, the localized region is updated in $F_a(t)$.

Terminating a Tracker – An actively tracked face $f_a \in F_a(t-1)$ is assumed to disappear or exit the scene if the confidence of the corresponding tracker is very low and the predicted region does not have any overlap with any of the detected regions i.e. if $\forall f_d \in F_d(t) \gamma_o(f_a, f_d) < \eta$. In this case, the corresponding face tracker is terminated and the face region is not updated in $F_a(t)$.

For all our experimentations, we have empirically chosen the value of η as 0.2. The above mentioned scheme for TLD based multiple face tracking provides us with face tracks as time indexed sets of face images extracted from the input videos. Face tracks obtained by multiple face tracking on two news broadcast videos and a sitcom season are shown in Figure 2. The face images in these extracted face tracks are further used to learn a dictionary of representative patches. This dictionary is used later to construct the descriptor of any face image. The process for learning the dictionary of patches is described next.

3. DICTIONARY OF PATCHES

The face images from all the face tracks are first converted into monochrome images followed by scaling to a fixed size. Patches R_i of size $R_{size} \times R_{size}$ are drawn randomly from these face images. Every patch R_i is vectorized and then normalized as

Figure 2: Results of multiple face tracking in the TLD framework. Parallel threads of TLD based trackers are employed to extract face tracks from videos. Here, we show a few images from the face tracks obtained from news broadcast videos ($TIMESNOW$ and $NDTV24\times7$) and a season of *Big Bang Theory*. All the images are scaled to the same size for display purposes.

$$\bar{R}_i[k] \leftarrow \frac{\bar{R}_i[k]-\mu_i}{max(\sigma_i, 1.0)} \tag{1}$$

where, μ_i and σ_i are the respective mean and standard deviation of the pixel intensity values of the patch R_i and $k = 1, \ldots R_{size} \times R_{size}$. This normalization of the individual patches enables us to achieve robustness against illumination changes. These vectorized patches are further magnitude normalized to form unit vectors $\hat{R}_i = \frac{\bar{R}_i}{||\bar{R}_i||}$.

The dictionary of patches is created by clustering the normalized patches into d number of clusters. We have used the modified spherical K-means clustering with cosine similarity as the metric for clustering the patches [1]. The choice of cosine similarity ensures that more importance is given to the non-zero values in a particular dimension instead of the overall magnitude; thereby giving more importance to the structure of the patch instead of the actual values of the patch pixels.

The spherical K-means clustering is initialized by choosing d vectors from the set $\{\hat{R}_i; i = 1, 2, \ldots\}$ as the initial cluster means $\hat{M}_j(1)$ $(j = 1, \ldots d)$ for the first iteration. Every vector $\hat{R}_i$ is assigned a label $l_i(t)$ in the t^{th} iteration as

$$l_i(t) = argmax_{j=1\ldots d} \hat{M}_j^T(t-1)\hat{R}_i \tag{2}$$

The cluster means are then re-estimated as,

$$V_j = \hat{M}_j(t-1) + \frac{\sum_i \hat{R}_i \delta(l_i(t)-j)}{\sum_i \delta(l_i(t)-j)} \tag{3}$$

$$\hat{M}_j(t) = \frac{V_j}{||V_j||} \tag{4}$$

where, $\delta(.)$ is the Kronecker Delta function. These steps are repeated for several iterations to obtain an optimal dictionary. For our experimentation we have resized all the face images to a size of 64×64, and extracted 60 patches (from random positions) of size 8×8 ($R_{size} = 8$) from each face image in the face tracks. The extracted patches are clustered into 1024 clusters ($d = 1024$). After clustering we filter out the means having very low magnitudes (almost equal to zero) and very low standard deviation (less than 0.25). This is performed as the low variance centers represent plain regions with uniform pixel intensities and do not correspond to any interesting features of the object. These are likely to be common across most objects and hence must be filtered out. The intuition for computing the inner product or the cosine similarity is that it is a better similarity measure in case of pixel values and more closely models the behavior of a filter acting on an image.

This learned dictionary is used further to compute the proposed face descriptors expressed as concatenated arrays of inner products. The process of facial feature extraction is described next.

4. FEATURE SPACE OF FACES

The facial features are essential for linking the face tracks from each resized face image (F_{mn}) of every face track (m^{th} image from the n^{th} track). First we extract, vectorize and normalize the non-overlapping patches of size $R_{size} \times R_{size}$from image F_{mn}. The number of patches P_{num} extracted from each image are same and depends on the patch size R_{size} and the size of scaled image. Each extracted patch is now represented by a vector of dimension d storing the cosine similarities between the extracted patch and the representative patches from the dictionary. The facial feature for the face image F_{mn} is obtained by concatenating the representative vectors from all the patches. Hence, each face image F_{mn} is represented by a vector of dimension $P_{num} \times d$.

In our experimentation, from the resized images of size 64×64 we got a total of 64 patches of size 8 and thus the size of representative vector is 64×1024. These learned facial features will have very similar values for the images of same person while significantly different values for that of others. This discriminative property of facial features motivates us to further use a SVM based classifier for linking the tracks. The methodology for SVM based face track linking is described next.

5. FACE TRACK LINKING

The learned facial features are expected to be sufficiently distinct for discriminating between the face images of two different people. In our proposed approach for face track linking using SVM, we exploit these discriminative properties of the facial features for linking two face tracks. We assume that all the face images in a particular track belong to the same person and hence, all of them can be linked together. A SVM trained with all face images from a particular face track as positives, can be used to classify a new

images. If most of the images from a face track **B** are classified as positives by a SVM trained on a previous track **A**, then we link the tracks **A** and **B**.

In the proposed framework, we assume that the co-occurring tracks cannot be linked. Hence, the face track linking is started from the longest available track co-occurring with at least one other track. Training the first SVM with longest track will ensure the robustness of the SVM, while the face images from the co-occurring tracks are used as negative samples. Care is taken that the number of positives and negatives in the SVM are approximately equal to prevent biasing (which was seen to degrade performance) by regulating the number of inputs used to train the SVM. Once the first SVM is trained, each new track is tested with the existing (trained) SVMs. If any of the SVM classifies most of the images (above a threshold α) from a particular track as positives then we link new track with that particular SVM. If all the SVMs reject the face track we train a new SVM for that track. While training a new SVM, we consider support vectors from all existing SVMs as negatives. We do not re-train SVM after a significant number of tracks have been grouped in the cluster due to chances of misclassification on account of possibly amplified error. The threshold α on the number of images above which we link two face tracks is determined experimentally and is specified as the fraction of total track length. In our experiments we have empirically set $\alpha = 0.8$.

6. RESULTS

We have experimented with two news broadcast videos (15 minutes long) from $TIMESNOW$ and $NDTV24 \times 7$, each having $22,500$ frames. $TIMESNOW$ had 146 tracks and 15 faces while $NDTV$ had 200 tracks and 21 faces.The dataset covers studio, field and discussion shots.

6.1 Performance Evaluation Of The Multiple Face Tracker

Since the tracker and detector are state-of-the-art methods in themselves, we have not presented a detailed evaluation of them in this work. We have however, presented an analysis of the multiple face TLD in track generation in terms of *Track Purity*. The Track Purity is defined as the ratio of Number of positive face images in a track to total number of images in a track. Out of 146 tracks obtained from $TIMESNOW$, 130 were pure resulting in an average track purity of 89.04%, while for $NDTV24 \times 7$, 177 pure tracks were obtained out of a total of 200 tracks, with an average track purity of 88.5%.

6.2 Performance Evaluation Of The Face Track Linking Scheme

We have evaluated the performance of face track linking method using four metrics viz – Accuracy ($\frac{TP+TN}{(TP+TN+FP+FN)}$) , Recall ($\frac{TP}{TP+FN}$) , Cluster Fragmentation (Ratio of Number of different faces going to a single cluster to total number of clusters) and cluster switch (Number of clusters a single person's face went to). Here, TP, FP, TN and FN are the respective number of true positives, false positives, true negatives and false negatives respectively. The average values obtained have been shown in Table 1.

7. CONCLUSION

We have proposed a novel methodology for linking face tracks extracted from videos. The Viola-Jones detector is used to localize faces in video frames, which are tracked further in a TLD framework leading to the extraction of face tracks. The face images in these tracks are first scaled to a fixed size. Patches drawn randomly from these resized face images are vectorized and normalized first and next subjected to spherical K-means clustering to learn a dictionary of representative vectors. For any face, we extract spatially ordered non-overlapping patches whose inner products with the dictionary vectors are concatenated to form a face descriptor. The face descriptors obtained from the same track are used as positive while the ones acquired from another co-occurring track are considered as negatives. A new face track is subjected to the learned SVMs and the one which successfully classifies most of the images is linked to the new track.

This work was an elementary step towards our broader goal of face analytics in news videos. The present work can be extended with the following improvements. First, an improved face detector needs to be developed as the Viola-Jones detector is only successful in detecting full frontal views. Second, we need to define an exhaustive reasoning scheme for tracking multiple faces with proper handling of occlusion cases. Third, the present work has scalability issues and can not handle large amounts of video data on account of increased search and the growth in the number of SVMs. Thus, within the framework of the present approach, the work can be extended with on-line spherical clustering and incremental learning over the face tracks.

8. REFERENCES

[1] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng. Text detection and character recognition in scene images with unsupervised feature learning. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 440–445. IEEE, 2011.

[2] S. Foucher and L. Gagnon. Automatic detection and clustering of actor faces based on spectral clustering techniques. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*, pages 113–122. IEEE, 2007.

[3] Y. Gao, T. Wang, J. Li, Y. Du, W. Hu, Y. Zhang, and H. Ai. Cast indexing for videos by ncuts and page ranking. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 441–447. ACM, 2007.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas. Face-tld: Tracking-learning-detection applied to faces. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3789–3792. IEEE, 2010.

[5] N. Pande, M. Jain, D. Kapil, and P. Guha. *The video face book*. Springer, 2012.

[6] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[7] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2856–2863. IEEE, 2013.

Table 1: Performance Of Face Track Linking

Videos	Accuracy	Recall	Cluster Fragmentation	Cluster Switch
Video 1	95.98%	73.38%	1.313	2
Video 2	90.1%	69.8%	1.5	4

[8] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji. Constrained clustering and its application to face clustering in videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3507–3514. IEEE, 2013.

[9] T. Zhang, D. Wen, and X. Ding. Person-based video summarization and retrieval by tracking and clustering temporal face sequences. In *IS&T/SPIE Electronic Imaging*, pages 86640O–86640O. International Society for Optics and Photonics, 2013.

[10] T. Zhang, J. Xiao, D. Wen, and X. Ding. Face based image navigation and search. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 597–600. ACM, 2009.