A Prior-Less Method for Multi-Face Tracking in Unconstrained Videos

Chung-Ching Lin
IBM Research AI
cclin@us.ibm.com

Ying Hung
Rutgers University
yhung@stat.rutgers.edu

Abstract

This paper presents a prior-less method for tracking and clustering an unknown number of human faces and maintaining their individual identities in unconstrained videos. The key challenge is to accurately track faces with partial occlusion and drastic appearance changes in multiple shots resulting from significant variations of makeup, facial expression, head pose and illumination. To address this challenge, we propose a new multi-face tracking and re-identification algorithm, which provides high accuracy in face association in the entire video with automatic cluster number generation, and is robust to outliers. We develop a co-occurrence model of multiple body parts to seamlessly create face tracklets, and recursively link tracklets to construct a graph for extracting clusters. A Gaussian Process model is introduced to compensate the deep feature insufficiency, and is further used to refine the linking results. The advantages of the proposed algorithm are demonstrated using a variety of challenging music videos and newly introduced body-worn camera videos. The proposed method obtains significant improvements over the state of the art [51], while relying less on handling video-specific prior information to achieve high performance.

1. Introduction

The task of Multiple Object Tracking (MOT) or Multiple Target Tracking (MTT) is to recover the trajectories of a varying number of individual targets while the status of targets is estimated at different time steps. Multi-face tracking is one of the important domains enabling high-level video content analysis and understanding, e.g., crowd analysis, semantic analysis, and event detection.

In this paper, our goal is to track an unknown number of human faces and maintain their identities in unconstrained videos (e.g., movies, TV series, music videos [51], body-worn camera videos). Our method does not assume any extra prior knowledge about the videos or require manual efforts (e.g., input underlying number of clusters in videos). Despite having different methods proposed to address this topic, this problem remains challenging due to the inherent unconstrained settings in videos. The videos might contain multiple shots captured by one or multiple moving cameras, irregular camera motion and object movement, arbitrary camera setting and object appearance, and people may move in-and-out camera field of view multiple times. The appearance of faces change drastically owing to significant variations of lighting condition, camera angle, expression, and head pose. Commonly, partial occlusions are caused by accessories and other body parts, such as glasses and hair, as well as hand gestures.

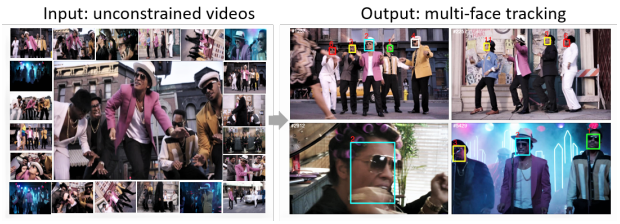

Figure 1: An example of multi-face tracking in unconstrained videos. The Bruno Mars music video shows the task is challenging due to partial occlusion and significant variations of lighting condition, camera angle, expression, and head pose across shots.

This is a difficult task and has a different focus from tracking in constrained videos (e.g., surveillance videos captured by steady or slowly-moving cameras), where the main challenge is to deal with different viewpoints, lighting conditions, and crowded pedestrian crossings. Many methods have been proposed [1, 3, 30, 41, 47, 49, 54]. In those papers, three popular datasets, MOT Challenge [29], PETS [16] and KITTI [17], are usually used to evaluate the performance of MOT methods. The videos, however, do not include multiple shot changes and appearance changes. These MOT methods attempt to solve different challenges and cannot be easily applied to unconstrained videos with large camera movement or multiple abrupt shot changes.

Due to the fast-growing popularity of unconstrained videos, especially on the Web, solutions to this problem are in high demand and have attracted great interest from researchers. The recently proposed methods [22, 23] enable users to track persons in unconstrained videos. These methods focus on the tracking accuracy within each shot,

but the scope does not include persons association across shots: they assign a new ID when a person reappears in the videos.

Figure 1 gives some sample frames from datasets used in this paper. We test our algorithm on two distinct types of challenging unconstrained video datasets. The first dataset, provided by [51], contains eight edited music videos with significant variation in expression, scale, pose, expression, and illumination in multiple shots. The second dataset is newly introduced in this paper. It includes four highly challenging unedited videos captured by people using body-worn cameras. The videos depict complex events but have limited quality control, and therefore, include severe illumination changes, camera motion, poor lighting, and heavy occlusion. In both datasets, persons are in-and-out of camera fields of view multiple times, and the proposed method is designed to track the faces across shots while maintaining the assigned identities, as shown in Figure 1.

Our framework incorporates three major components to achieve high accuracy face tracking, and it is robust to substantial face rotations, from frontal to profile. First, we develop a co-occurrence model of multiple body parts to create longer face tracklets. We then develop a recursive algorithm to link tracklets with strong associations. Finally, a Gaussian process model is designed to refine the clustering results by detecting and reassigning outliers. The main contributions can be summarized as follows:

1. We propose a prior-less framework that is capable of tracking multiple faces with unified handling of complex unconstrained videos.
2. The proposed method provides a data-driven estimation of the cluster number in an automatic fashion. This is in contrast to existing works that assume face tracklets are given or require manual entry of the underlying cluster number in advance.
3. Our proposed co-occurrence model can continue tracking multiple faces that are only partially visible. Even with information loss, such as with head turning, or when faces are occluded, the proposed algorithm can determine with whose body the partial observation should be matched and continue to track faces throughout.
4. We introduce a new dataset of four highly challenging, realistic, unedited, body-worn camera videos captured by police officers in different incidents. The dataset introduces new challenges to MOT in unconstrained videos.

2. Related Work

Multiple person tracking. There has been extensive research related to multiple person tracking. Many efforts have explored this problem using data association approaches [26, 50], such as Markov decision process [47], event aggregation [20], greedy algorithm [38], and attentional correlation filter [7]. However, these works either explicitly or implicitly assume continuous appearances, positions, and motions. They are thus ineffective for solving shot change problems. Several existing methods [3, 45, 28, 33] explore appearance features to find tracklet associations, which can link tracklets across shots. These methods employ discriminative appearance-based affinity models to help associate persons in tracking tracklets, but they are not directly applicable to videos with significant variations in facial appearance.

CNN-based representation learning. Many areas have gained performance improvement from advances in deep learning. Several CNN-based models for face recognition provide biometrics-solutions: VGG-Face [37], DeepFace[42], and FaceNet[40]. The datasets that are used to train these CNN models are generally chosen from good conditions, e.g., high image resolution, frontal faces, rectified faces, and full faces. However, in an unconstrained video, a face could be profiled, occluded, cropped, or blurry. In these cases, measuring the similarity with extracted deep face features might yield inferior performance.

Person re-identification [18, 19, 52, 32, 31] also gains performance boost using deep learning techniques. Methods include dual mutual learning, deep transfer learning, multi-loss classification, and triplet loss, etc. These papers focus on searching different perspective views of the same person. The subjects are required to be in the same outfit. Our problem has a unique characteristic that distinguishes it from re-identification modeling. The videos used in our problem are more unstructured because persons' appearances and cameras' movements are unconstrained. These changing parameters along with different shot changes allow for more ambiguous conditions.

Unconstrained face tracking. Tracking has been extensively developed in scenarios with multiple faces [9, 10, 15, 36, 24]. Many multi-face tracking works exist for constrained videos with limited camera motion, e.g., web-cam, surveillance. Current studies focus on the analysis of more unstructured, unconstrained videos. Among them, there have been significant efforts at analyzing the fast-growing popularity of internet videos.

Wu et al. [46] propose a multi-face tracking method to handle clustering and tracklet linking simultaneously using hidden Markov random field model in a TV series video (Big Bang Theory, BBT). Their method uses appearance affinity and HSV information to measure the similarity of faces. As a result, the approach is constrained to good quality of frontal faces. Another line of work [48, 53] propose methods to learn the appearance representation to cluster the faces tracklets on TV series and movie videos

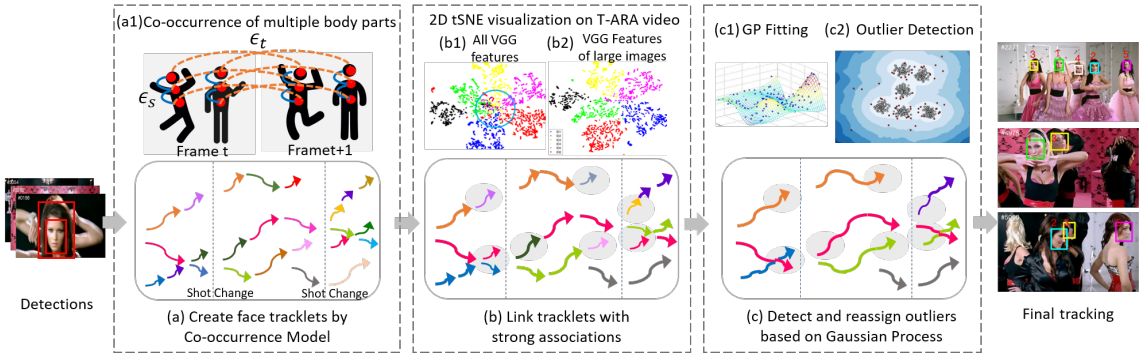

Figure 2: Illustration of three core algorithmic components of the proposed method. (Best viewed in color)

(Buffy, Notting-Hill, and Harry Potter). They assumes face tracklets and the number of clusters are given. In studies [8, 43, 46, 48], the tracklets of false positives are manually removed. Different from face clustering studies starting the problem from given face tracklets, multi-face tracking problem take raw videos as input to perform detection, generate tracklets and link tracklets.

State-of-the-art. Recently, Zhang et al. [51] propose a CNN-based approach of learning video-specific facial discriminative features for multi-face tracking and demonstrate state-of-the-art performance. The main limitation of this method is that it has difficulty in handling videos where many shots contain only one single person. In these cases, the method cannot generate sufficient negative face pairs to train the network, thus different persons might be incorrectly identified as the same person across shots. Additionally, the method requires prior knowledge of the videos to provide actual number of clusters in advance. In reality, the correct and optimal choice of cluster numbers is often ambiguous in application to some videos of minor characters. If cluster numbers are ill-initialized, clustering purity would degrade. Further, an essential prerequisite of this method is to apply an effective shot change detection technique to partition each input video into non-overlapping shots.

In contrast, we propose an algorithm to analyze raw video data and generate final tracking and clustering results automatically in a data-driven fashion. The proposed method seeks to eliminate the sensitivities of handling video-specific prior information.

3. Algorithm

To achieve better tracking results, longer tracklets of each person are desired. The longer the tracklets are, the greater the number of possible facial variations of each person could be captured. However, longer tracklets usually contain more noise, and thus might incur more tracklet linking errors. Considering the pros and cons, we propose a framework, as illustrated in Figure 2, that includes three core algorithmic components: **(1) Create tracklets**. We develop a co-occurrence model of multiple body parts to create longer face tracklets. A face is temporarily missing when a person turns his/her head, or the view of their face is blocked by another object (e.g., hand, or others' head). The model is designed to prevent tracks from being terminated when an image of a face temporarily disappears (Section 3.1). **(2) Link tracklets**. We recursively link tracklets with strong associations. The recursively linked tracklets construct a constrained graph for extracting clusters, and generating initial clustering results (Section 3.2). **(3) Detect and reassign outlier tracklets**. We design a Gaussian Process model to capture the richness of data and compensate the deep feature insufficiency. Our model will detect and re-assign outlier tracklets (Section 3.3).

3.1. Tracking by Co-occurrence Model

Typically, the performance of detectors is greatly affected by pose, occlusion, rotation, size and image resolution. For example, when a person turns his/her head away from the camera or the face is occluded, the face might not be detected. However, the head belonging to that person could be still detected and tracked. We build on an idea that using multiple body parts simultaneously could create longer tracklets. To this end, we developed a co-occurrence model which obtains information of multiple body parts to help continue the tracker during moments when faces are not captured by the camera or not detected by the detector, but the person remains in the video frames.

Our starting point is the multiple body parts detections estimated by off-the-shelf body-part detector [5]. Note that the detection method could be replaced by other sophisticated body-part detectors [6, 39]. For each video frame, we extract localization of face, head, torso, and whole body. We denote $\{v^t_{k,\gamma}\}$ a discrete set of outputs of body-part detections in a frame t where $v^t_{k,\gamma} = [c^t_{x,k}, c^t_{y,k}, w^t_k, h^t_k]$, k is the index of the detection; c, w, h are center, width and height of a bounding box; γ denotes the type of body part, such as torso, $\gamma \in \Gamma = \{1, ..., N\}$.

For each body part detection, two thresholds are applied [21]: (1) detection results filtered by a high threshold are

used to create new tracklets; (2) detection results filtered by a low threshold are used to track objects.

We formulate the multi-person tracking problem as a graph structure $G = (v, \epsilon)$ with two types of edges, ϵ_s and ϵ_t, as shown in Figure 2 (1a). Spatial edges ϵ_s denote the connections of different body parts of a candidate within a frame. The spatial edges ϵ_s are used to generate hypothesized states of a candidate. Temporal edges ϵ_t denote the connections of the same body parts over adjacent frames. The state of each individual person in different frames are estimated using temporal edges.

$$\epsilon_s = \{\langle v^t_{k,\gamma}, v^t_{k',\gamma'}\rangle : \gamma \neq \gamma'\},\ \epsilon_t \quad = \{\langle v^t_{k,\gamma}, v^{t-1}_{n,\gamma}\rangle\}. \quad (1)$$

The spatial edges ϵ_s are defined as:

$$\langle v^t_{k,\gamma}, v^t_{k',\gamma'}\rangle = \delta \cdot \phi(v^t_{k,\gamma}, v^t_{k',\gamma'}), \quad (2)$$

where $\phi(v^t_{k,\gamma}, v^t_{k',\gamma'})$ and δ are indicator functions. $\phi(v^t_{k,\gamma}, v^t_{k',\gamma'}) = 1$ when the overlapping area is larger than a threshold ζ. $\delta = 1$ when there is an exclusive connection between two types of body parts in one frame. This constraint ensures that two body parts are associated to the same person only if the connection is not considered as ambiguous, such as two different face detections connected to the same head.

After all the ϵ_s are built, the connected components are used to generate $G^{t,i}_p$ as a hypothesis of a person ξ_i at frame t. Ideally, $G^{t,i}_p$ consists of all detected body parts that belong to the same person ξ_i.

Consider estimation of the current state of a person ξ^t_i given the observations Z^t from frame 0 to t. The problem can be formulated as maximization of the likelihood function given the previous state of the person ξ^{t-1}_i :

$$p(\xi^t_i|Z^t) = \max_j f(G^{t,j}_p|\xi^{t-1}_i), \quad (3)$$

The likelihood $f(G^{t,j}_p|\xi^{t-1}_i)$ can be viewed as a method to evaluate how well a candidate hypothesis matches the previous state. We define the likelihood $f(G^{t,j}_p|\xi^{t-1}_i)$ as the probability of a candidate hypothesis $G^{t,j}_p$ given the previous state ξ^{t-1}_i, where its value is given by the maximum transition probability from one of the body parts among $G^{t,j}_p$:

$$\begin{aligned} f(G^{t,j}_p|\xi^{t-1}_i) &= p(G^{t,j}_p|G^{t-1,i}_p) \\ &= \max\{p(v^{t,j}_\gamma|v^{t-1,i}_\gamma),\ \forall\gamma \in G^{t,j}_p\}. \end{aligned} \quad (4)$$

The body-part transition probability $p(v^{t,j}_\gamma|v^{t-1,i}_\gamma)$ is defined as: $p(v^{t,j}_\gamma|v^{t-1,i}_\gamma) = \eta(v^{t,j}_\gamma, v^{t-1,i}_\gamma)$, where the potential function $\eta(v^{t,j}_\gamma, v^{t-1,i}_\gamma)$ is defined as the overlapping ratio of the bounding boxes.

If a body part gives higher likelihood than the likelihood of another body part, then it has better representation of the candidate. Equations 3 and 4 ensure that if a face is temporally missing, then the body part which collects the most information of a candidate person is still used to track.

When no corresponding body parts coexist, we use KLT [44] and the sum of absolute difference (SAD) to predict the hypotheses of each body parts. After obtaining the current state of a person, we build temporal edges ϵ_t by connecting the same type of body part among $G^{t,i}_p$ and $G^{t-1,i}_p$. Next, we generate face tracklets using face bounding boxes from each individual person's tracklets and extract facial features for clustering in the next session.

3.2. Recursive Constrained Tracklet Linking

After face tracklets are generated, each face tracklet is taken as a node T_i, which includes various face poses of a person with extracted feature $\{f^i_k\}^{n_i}_{k=1}$ and frame indexes $\{t^i_k\}^{n_i}_{k=1}$. We aim to infer the underlying pairwise similarity between nodes to construct meaningful affinity graphs for face clustering. Specifically, we use the VGG-face descriptors [37] to extract features. In contrast to [51], we do not fine-tune the feature extraction network for any video as it brings high computational cost. We design a unified and generalized linking framework based on how the VGG-face network was trained to avoid less informative features by measuring between-node proximity. We further construct similarity graphs that better express the underlying face features in clusters.

We first use face bounding boxes of every tracklet to obtain face images and extract 4096-dimension VGG-face features from the FC7 layer. The extracted deep facial features are normalized for comparisons. Given that the VGG-face network is trained with high-resolution images by triplet loss objective function, our key idea is that higher resolution images and relative distance between nodes would provide more meaningful information in a model exploiting extracted VGG-face features. We build two types of links: $\{L_l\}$ and $\{L_c\}$. $\{L_l\}$ and $\{L_c\}$ are built by the properties of image resolution and relative distance between tracklets respectively.

Figure 2(b) shows 2D tSNE [34] visualizations of extracted VGG features on the T-ara video. It shows that, compared to all features (b1), features of large images (b2) are more discriminative. Thus, we start to build the linkages $\{L_l\}$ using tracklets with higher image resolutions as they could construct strong associations.

We measure the pairwise similarity between two tracklets to build linkings. $M_{ll}(T_i, T_j)$ is used to measure the similarity between tracklets T_i and T_j, taking account of appearance affinity and resolution constraint.

$$M_{ll}(T_i, T_j) = \Lambda^a(T_i, T_j)\Lambda^s(T_i)\Lambda^s(T_j), \quad (5)$$

where $\Lambda^a(T_i, T_j)$ is to evaluate the appearance similarity using the distance between tracklets $D(T_i, T_j)$.

$$\Lambda^a(T_i, T_j) = \begin{cases} 1 - D(T_i, T_j), & \text{if } D(T_i, T_j) < \varphi \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $D(T_i, T_j)$ is measured by the distance between VGG-face features $d(f_k, f_h)$. All linkages are built when $D(T_i, T_j)$ is smaller than a threshold φ.

$$D(T_i, T_j) = \min_{f_k \in T_i, f_h \in T_j} d(f_k, f_h), \quad (7)$$

where $d(f_k, f_h)$ is the Euclidean distance between f_k and f_h.

$\Lambda^s(T_i)$ enforces the resolution constraint and builds linkages among tracklets that have larger image size. We defined $\Lambda^s(T_i)$ as:

$$\Lambda^s(T_i) = \begin{cases} 1, & \text{if } T_i \in \Psi_L \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

We apply k-means method to separate all tracklets based on the average image size of each tracklet and obtain group Ψ_L, which consists of tracklets with larger image size.

Another type of linkage $\{L_c\}$, is built by the relative distances among coexisting tracklets. First, we search all sets of coexisting tracklets that have overlapping frame indexes. Given their existence in the overlapping frame indexes, the tracklets should be mutually exclusive at any given time. No person can have more than one existence. Thus, the coexisting tracklets should not be linked. We use relative distance properties to build constrained linkages among all pairs of coexisting tracklets by the procedures described in Algorithm 1 . For each tracklet in $\{T_i^A\}$, we search the corresponding nearest neighbor tracklet in $\{T_j^B\}$ and build a linkage between them using the similarity measurement $M_{lc}(T_i^A, T_j^B)$, which takes into account for appearance affinity and relative distance constraints.

$$M_{lc}(T_i^A, T_j^B) = \Lambda^a(T_i^A, T_j^B)\Lambda^r(T_i^A)\Lambda^r(T_j^B), \quad (9)$$

where $\Lambda^a(T_i^A, T_j^B)$ is the same as Equation 6. $\Lambda^r(T_i^A)$ is used to impose relative distance constraints. Because coexisting tracklets should be exclusive, the connection between a tracklet in $\{T_i^A\}$ and a tracklet in $\{T_j^B\}$ should be one or none. We use this property to prevent false connections. When the relative distance between two tracklets (T_i^A, T_j^A) is smaller than ϑ or multiple tracklets in one set are connected to the same tracklet in the other set, the tracklets are very similar and hard to distinguish from each other. In this case, the linkages are disconnected. We define $\Lambda^r(T_i^A)$ as:

$$\Lambda^r(T_i^A) \begin{cases} 1, & \text{otherwise} \\ 0, & \text{if } D(T_i^A, T_j^A) < \vartheta, \forall T_j^A \neq T_i^A \in \{T_i^A\}. \end{cases}$$

This process is performed recursively until all pairs of sets of coexisting nodes have been evaluated.

After obtaining $\{L_l\}$ and $\{L_c\}$, all the linkages form a graph, G_T. The connected components in G_T are extracted and used to generate initial clusters.

Algorithm 1 Linking Coexisting Tracklets

Find all sets of coexisting nodes
for Every pair of sets of coexisting nodes: $\{T_i^A\}$, $\{T_j^B\}$ **do**
  Find maximum $M_{lc}(T_i^A, T_j^B)$ using Equation 9
  Built linkages between the pair of tracklets if $\max M_{lc}(T_i^A, T_j^B) > 0$
end for

3.3. Refinement Based on Gaussian Process (GP)

Empirical studies [25, 14] show CNN-based models are very sensitive to image blur and noise because the networks are generally trained on high quality images. Considering our recursive linking framework is initialized from CNN-based features to obtain better representations of the underlying face clusters, there would intrinsically exist some tracklets incorrectly linked to other tracklets. In order to find the incorrect association tracklets, we design a Gaussian process model to compensate for the deep feature limitations and to capture the richness of data. We apply the GP model to detect outliers, disconnect the linkages among outliers and other tracklets, and then reassign the outliers to refined clusters formed after the outliers are disconnected, thus yielding high-quality clusters.

3.3.1 Dimension Reduction Using GP

Gaussian process (GP) models, also known as kriging, are commonly used in many applications including machine learning and geostatistics [11]. Different from CNN-based approaches, GP models provide a flexible parametric approach to capture the nonlinearity and spatial-temporal correlation of the underlying system. Therefore, it is an attractive tool to be combined with the CNN-based approach to further reduce the dimension without losing complex, and important spatial-temporal information. Here, we illustrate the idea of reducing the dimension by fitting a GP model for each color channel with the spatial information. Three GP models are constructed obtained and the dimension is reduced to 18 parameters captured by the GP models. Note, the reduced dimension is not restricted to 18 and may be flexibly determined by changing the number of parameters in the GP models.

A Gaussian process model can be written as

$$y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \quad (10)$$

where $y \in \mathbb{R}$ is the intensity of a color and $\boldsymbol{x} \in \mathbb{R}^p$ is the input. In this research, $\boldsymbol{x}$ represents the spatial information, so $p = 2$. The mean function $\mu(\boldsymbol{x})$ is assumed to be a function of $\boldsymbol{x}$ with unknown parameters $\boldsymbol{\beta}$, say, $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. In addition, $Z(\boldsymbol{x})$ is a Gaussian process with mean 0 and $\text{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$, where $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$ is the correlation function and $\boldsymbol{\theta}$ is a vector of unknown correlation parameters. There are

various correlation functions discussed in the literature. Here we focus on a commonly used product form of power exponential functions:

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{k=1}^{p} R_k(|x_{ik} - x_{jk}|) \quad (11)$$

$$= \prod_{k=1}^{p} \exp(-\theta_k |x_{ik} - x_{jk}|^{\alpha}), \quad (12)$$

where $0 < \alpha \leq 2$ is a tuning parameter and $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with $\theta_k \geq 0$ for $i = 1, 2$. Because the correlation parameters, θ_k's, are not constrained to be equal, the model can handle different signals in each input dimension which makes Equation (11) particularly attractive to the analysis of complex underlying system.

Given n realizations of a particular color channel $\boldsymbol{y} = (y_1, ..., y_n)^\top$ and the corresponding spatial information $X = (\mathbf{x_1}^\top, ..., \mathbf{x_n}^\top)^\top$, the joint log-likelihood function for (10) can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\boldsymbol{\beta}) - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})| - \frac{n}{2}\log(\sigma^2),$$

where $\Sigma(\boldsymbol{\theta})$ is the $n \times n$ correlation matrix with the ij^{th} element equal to $\phi(\mathbf{x_i}, \mathbf{x_j}; \boldsymbol{\theta})$. The maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$ and σ can be obtained by

$$\widehat{\boldsymbol{\beta}} = (X^\top \Sigma^{-1}(\boldsymbol{\theta}) X)^{-1} X^\top \Sigma^{-1}(\boldsymbol{\theta}) \boldsymbol{y}, \quad (13)$$

$$\widehat{\sigma}^2 = (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})/n. \quad (14)$$

By maximizing the logarithm of the profile likelihood, the MLE of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{n \log(\widehat{\sigma}^2) + \log|\Sigma^{-1}(\boldsymbol{\theta})|\}. \quad (15)$$

For the estimation of correlation parameters $\boldsymbol{\theta}$, there are some likelihood-based alternatives. These alternatives include the restricted maximum likelihood (REML) and penalized likelihood approaches. In this paper, we focus on the study of MLEs, but the results can be further extended to the likelihood-based alternatives.

According to Equation (13, 14, 15), there are six parameters $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \widehat{\boldsymbol{\theta}})$ obtained for each color channel, therefore a total of 18 parameters are obtained to capture the underlying information of a given face image.

3.3.2 Outlier Detection and Reassignment by GP

We introduce the outlier detection and reassignment scheme in this section. Our idea is to measure how isolated a tracklet is when compared to the spatial surrounding neighborhood. More precisely, by comparing the local density of a tracklet to the local densities of its neighbors, we can identify tracklets that have a substantially lower density than their neighbors, as shown in Figure 2 (c2). These tracklets are considered outliers, and may belong to other clusters. We detect these outlier tracklets and reconnect them to one of the clusters by extracted GP features $r \in R^{18}$. For each cluster, we use a simple unsupervised outlier detection method, Local Outlier Factor (LOF) estimator [4], to compute the local density deviation of a given tracklet with respect to its neighbors. After detecting outliers, we refine the original clusters by disconnecting the linkages among outliers and other tracklets as there might be incorrect association among those linkages.

We further use extracted GP features to link all isolated tracklets to the refined clusters. We evaluate the appearance similarity between an isolated tracklet and every refined cluster. For any given isolated tracklet, we evaluate all pairwise distances between the isolated tracklet and every tracklet in one cluster and use the shortest distance as the similarity measure between the isolated tracklet and the cluster. We also enforce a temporal constraint to prevent multiple tracklets with overlapping frame indexes in the same cluster. Next, we determine the cluster that has the shortest distance to the given isolated tracklet and assign the tracklet into that cluster.

After all tracklets have been connected into one of the clusters, we obtain final clusters. Finally, we assign a specific identity to each cluster and generate final tracking.

4. Experiments

We empirically demonstrate the effectiveness of our proposed method on two distinct types of challenging unconstrained video datasets and compare with state-of-the-art methods, especially with variants in [51].

4.1. Details

Dataset: Experiments are conducted on two datasets: (1) Edited High-quality Music Video Dataset. The dataset made available by [51] contains 8 edited music videos. The videos contain dramatic facial appearance changes, frequent camera view and shot changes, and rapid camera motion. (2) Unedited Body-worn Camera Video Dataset. We introduce a new highly-challenging dataset of 4 realistic and unedited body-worn camera videos. All videos were captured by police officers in different incidents, and thus have limited quality control. The videos in this dataset have very severe camera movement and heavy occlusion. There is a large number of dark scenes and many tracks with non-frontal faces.

Experiment settings: All parameters have the same settings and remain unchanged for all videos: high detection threshold for creating new track is 0.8, low detection threshold for tracking is 0.1. φ is 0.5; ζ is 0.9;

ϑ is 0.7; α is 2. Note that, in contrast to [51], our method does not apply shot change detection and does not assume that the total cluster number is known a priori.

Evaluation metrics: (1) Clustering. We use Weighted Clustering Purity (WCP) [51] to evaluate the extent to which faces can be clustered automatically according to their identities. WCP is given as $WCP = \frac{1}{N}\sum_{c\in C} n_c \cdot p_c$ where N is the total number of faces in the video, n_c is the number of faces in the cluster $c \in C$, and its purity, p_c, is measured as the fraction of the largest number of faces from the same person to n_c, and C is the total number of clusters. (2) Tracking. We report tracking results based on the most widely accepted evaluation metrics, the CLEAR MOT [35], including Recall, Precision, F1, FAF, MT, IDS, Frag, MOTA, and MOTP.

4.2. Experiments on Edited High-quality Music Video Dataset

Clustering. We report the clustering accuracy of our method and other competitors, HOG[12], AlexNet[27], VGG-face[37], and variants in [51] in Table 1. Table 1 shows that our method achieves substantial improvement compared to the best competitor (e.g., from 0.56 to 0.86 in Westlife), demonstrating the superiority of our method. Furthermore, we analyze the effectiveness of our outlier reassignment scheme. The third to last row reports the performance of our method without outlier reassignment, which achieves performance comparable to the state-of-the-art methods. But, as showed in the second to last row, our integrated framework can compensate the deep feature insufficiency and bring the full potential of the proposed method. The last row presents the number of clusters estimated automatically by our method versus the ground-truth number of clusters. It shows that our method does not suffer from the basis ambiguity difficulty, but is able to generate number of clusters automatically and reliably.

Face Tracking. We report the face tracking results in Table 2. Our method is compared with ADMM[2], IHTLS[13] and variants in [51]. Table 2 shows the proposed method improves tracking performance against the existing methods for most metrics. Overall, we achieve better performance in terms of Recall, Precision, F1, MOTA and MOTP. Specifically, our method noticeably increases most tracked (MT), and reduces the number of identity switching (IDS) and track fragments (Frag). This implies our co-occurrence tracker can robustly construct longer trajectories, and face IDs are correctly maintained by our recursive linking framework.

Qualitative Results. Figure 3 shows sample tracking results of our algorithm. In some frames, we can see that different persons have very similar face appearance, multiple main singers appear in a cluttered background filled with audiences, or some faces have heavy occlusions by other cast members. As shown, the proposed algorithm is capable of generating invariant face identities and tracking them reliably across different shots in the entire unconstrained video.

Table 1: Clustering purity comparisons with the state-of-the-art methods on 8 music videos. The best results are highlighted with the bold.

MUSIC VIDEO DATASET								
Method	T-ara	Pussycat Dolls	Bruno Mars	Hello Bubble	Darling	Apink	Westlife	Girls Aloud
HOG[12]	0.22	0.28	0.36	0.35	0.19	0.20	0.27	0.29
AlexNet[27]	0.25	0.31	0.36	0.31	0.18	0.22	0.37	0.30
VGG-face[37]	0.23	0.46	0.44	0.29	0.20	0.24	0.27	0.29
Pre-trained[51]	0.31	0.31	0.50	0.34	0.24	0.29	0.37	0.33
Siamese[51]	0.69	0.77	0.88	0.54	0.46	0.48	0.54	0.67
Triplet[51]	0.68	0.77	0.83	0.60	0.49	0.60	0.52	0.67
SymTriplet[51]	0.69	0.78	**0.90**	0.64	0.70	0.72	0.56	0.69
W/o GP outlier reassign.	0.87	0.77	0.78	0.63	0.68	0.64	0.70	0.61
The proposed framework	**0.89**	**0.79**	0.85	**0.70**	**0.73**	**0.92**	**0.86**	**0.92**
Estimated / GT cluster no.	6/6	6/6	11/11	4/4	7/8	6/6	4/4	5/5

Table 2: Quantitative comparisons with the state-of-the-art tracking methods on music video dataset.

MUSIC VIDEO DATASET									
Method	Recall↑	Precision↑	F1↑	FAF↓	MT↑	IDS↓	Frag↓	MOTA↑	MOTP↑
ADMM[2]	75.5	61.8	68.0	0.50	23	2382	2959	51.7	63.7
IHTLS[13]	75.5	68.0	71.6	0.41	23	2013	2880	56.2	63.7
Pre-Trained[51]	60.1	88.8	71.7	**0.17**	5	931	2140	51.5	79.5
mTLD[51]	69.1	88.1	77.4	0.21	14	1914	2786	57.7	80.1
Siamese[51]	71.5	89.4	79.5	0.19	18	986	2512	62.3	64.0
Triplet[51]	71.8	88.8	79.4	0.20	19	902	2546	61.8	64.2
SymTriplet[51]	71.8	89.7	79.8	0.19	19	699	2563	62.8	64.3
Ours	**81.7**	**90.2**	**85.3**	0.27	**32**	**624**	**1645**	**69.2**	**86.0**

Speed. We have measured execution speed of the proposed method on music videos that typically have several faces to be tracked in each frame. In one 5-minutes music video, there are 21,747 face observations over a sequence of 5,000 frames, our implementation takes about 25 minutes after feeding the detection results. The running time is implemented with unoptimized C++ and Matlab code, single thread execution on a Mac with Intel 2.5 GHz i7 CPU and 16 GB memory.

4.3. Experiments on Unedited Realistic Body-worn Camera Dataset

To further test the capability of our method, we conduct experiments on unedited realistic body-worn camera dataset and compare the results with variants in [51].

Clustering. We compare the clustering results with HOG[12], AlexNet[27], VGG-face[37], pre-trained, Siamese and SymTriplet in [51]. Table 3 shows our method outperforms other methods with noticeable margin on all videos in the body-worn camera dataset. This problem is particularly challenging. For example, in Foot Chase video, our method achieves weighted purity of 0.73. But even for the best-performing feature in [51], SymTriplet, it only achieves purity of 0.45. HOG, AlexNet, VGG-face also perform poorly. The possible reason is that 3 body-worn camera videos (Foot Chase, TS1 and TS3) only have 640x480 resolution, and these

Figure 3: Sample tracking results of the proposed algorithm. The first two rows are Westlife and Hello Bubble from music video dataset. The bottom row is Foot Chase from body-worn camera dataset. The ID number and color of face bounding box for each person are kept. (Refer to the supplementary material for more results.)

Table 3: Clustering purity comparisons on 4 body-worn camera videos.

BODY-WORN CAMERA DATASET				
Method	Foot Chase	TS1	TS3	DVHD2
HOG[12]	0.40	0.52	0.58	0.50
AlexNet[27]	0.40	0.54	0.58	0.59
VGG-face[37]	0.43	0.46	0.72	0.72
Pre-trained[51]	0.42	0.54	0.61	0.74
Siamese[51]	0.41	0.54	0.68	0.56
SymTriplet[51]	0.45	0.55	0.69	0.77
W/o GP outlier reassign.	0.64	0.74	0.77	0.75
The proposed framework	**0.73**	**0.80**	**0.80**	**0.81**
Estimated / GT cluster no.	4/5	3/3	2/2	3/3

Table 4: Quantitative comparisons with the state-of-the-art tracking method [51] on body-worn camera dataset.

BODY-WORN CAMERA DATASET									
Method	Recall↑	Precision↑	F1↑	FAF↓	MT↑	IDS↓	Frag↓	MOTA↑	MOTP↑
mTLD[51]	75.1	79.2	75.8	0.14	7	70	400	52.7	93.5
Pre-Trained[51]	75.1	79.2	75.8	0.14	7	61	404	52.9	93.5
Siamese[51]	75.1	79.2	75.9	0.14	7	55	404	52.8	93.5
SymTriplet[51]	75.1	79.8	75.9	0.13	7	52	390	53.9	93.5
Ours	**78.6**	**93.8**	**85.4**	**0.07**	**11**	**39**	**188**	**69.8**	**93.6**

methods cannot cope with such low resolution. In addition, SymTriplet [51] requires sufficient negative pairs generated from tracklets that co-occur in the same shot. But in body-worn camera videos, many shots contain only a single person. Consequently, they are unable to train their network and fine-tune features well. However, these problems are addressed by our proposed method. We believe the significant performance difference lies in our GP model is designed to compensate the insufficiency of the CNN-based initialized linking framework and capture the false positive tracklet associations. Again, the last row shows the proposed method is able to generate the number of clusters automatically and reliably.

Face Tracking. We report the face tracking results on body-worn camera videos in Table 4. Our method is compared with 4 variants in [51]. The body-worn camera videos are captured with limited quality control; thus, they usually contain undesirable motion blur caused by camera shake. Video quality degradation yields more false positive detections, which increases the tracking difficulty. Table 4 shows the proposed method produces overall superior performance for all metrics. This implies the proposed method can overcome the difficulty and handle lower quality videos better.

Qualitative Results. Figure 3 shows that the proposed algorithm is able to robustly track multiple faces with their correct identities in the shaking and low resolution unconstrained videos. More qualitative results are available in the supplementary material.

5. Conclusions

We have introduced a prior-less algorithm for reliably tracking multiple faces in unconstrained videos, where extensive motion and variations exist and affect the way by which many heretofore existing methods perform. Experiments on two distinct video datasets demonstrated the superiority of the proposed method when compared to the state-of-the-art methods that require intensive training to fine-tune the networks or manual video analysis to obtain the number of clusters. In the future, we intend to explore modeling the similarity of other body parts to extend our framework's capability.

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1926–1933. IEEE, 2012.

[2] M. Ayazoglu, M. Sznaier, and O. I. Camps. Fast algorithms for structured robust principal component analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1704–1711. IEEE, 2012.

[3] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225. IEEE, 2014.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[5] L. M. Brown and Q. Fan. Enhanced face detection using body part detections for wearable cameras. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 715–720. IEEE, 2016.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[7] J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, and J. Young Choi. Visual tracking using attention-modulated disintegration and integration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4321–4330, 2016.

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1559–1566. IEEE, 2011.

[9] F. Comaschi. *Robust online face detection and tracking*. PhD thesis, Technische Universiteit Eindhoven, 2016.

[10] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal. Online multi-face detection and tracking using detector confidence and structured svms. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015.

[11] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[13] C. Dicle, O. I. Camps, and M. Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2311, 2013.

[14] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pages 1–6. IEEE, 2016.

[15] S. Duffner and J.-M. Odobez. Track creation and deletion framework for long-term online multiface tracking. *IEEE Transactions on image processing*, 22(1):272–285, 2013.

[16] A. Ellis and J. Ferryman. Pets2010 and pets2009 evaluation of results using individual ground truthed single views. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 135–142. IEEE, 2010.

[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[18] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[19] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[20] J. Hong Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon. Online multi-object tracking via structural constraint event aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1392–1400, 2016.

[21] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *European Conference on Computer Vision*, pages 788–801. Springer, 2008.

[22] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, B. Schiele, and S. I. Campus. Arttrack: Articulated multi-person tracking in the wild. In *Proc. of CVPR*, 2017.

[23] U. Iqbal, A. Milan, and J. Gall. Pose-track: Joint multi-person pose estimation and tracking. *arXiv preprint arXiv:1611.07727*, 2016.

[24] S. Jin, H. Su, C. Stauffer, and E. Learned-Miller. End-to-end face detection and cast grouping in movies using erdos-renyi clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5276–5285, 2017.

[25] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel. How image degradations affect deep cnn-based face recognition? In *Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the*, pages 1–5. IEEE, 2016.

[26] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele. A multi-cut formulation for joint segmentation and tracking of multiple objects. *arXiv preprint arXiv:1607.06317*, 2016.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1217–1224. IEEE, 2011.

[29] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[30] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth. Tracking the trackers: An analysis of the

state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017.

[31] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[32] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

[33] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2953–2960. IEEE, 2009.

[34] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[35] A. Milan, K. Schindler, and S. Roth. Challenges of ground truth evaluation of multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 735–742, 2013.

[36] P. Motlicek, S. Duffner, D. Korchagin, H. Bourlard, C. Scheffler, J.-M. Odobez, G. D. Galdo, M. Kallinger, and O. Thiergart. Real-time audio-visual analysis for multiperson videoconferencing. *Advances in Multimedia*, 2013:4, 2013.

[37] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[38] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE, 2011.

[39] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3178–3185. IEEE, 2012.

[40] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[41] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1815–1821. IEEE, 2012.

[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[43] M. Tapaswi, O. M. Parkhi, E. Rahtu, E. Sommerlade, R. Stiefelhagen, and A. Zisserman. Total cluster: A person agnostic clustering method for broadcast videos. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 7. ACM, 2014.

[44] C. Tomasi and J. Shi. Good features to track. *CVPR94*, 600:593–593, 1994.

[45] B. Wang, G. Wang, K. Luk Chan, and L. Wang. Tracklet association with online target-specific metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1241, 2014.

[46] B. Wu, S. Lyu, B. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos, ieee intl. In *Conf on Computer Vision (ICCV)*, 2013.

[47] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4705–4713, 2015.

[48] S. Xiao, M. Tan, and D. Xu. Weighted block-sparse low rank representation for face clustering in videos. In *European Conference on Computer Vision*, pages 123–138. Springer, 2014.

[49] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1918–1925. IEEE, 2012.

[50] H. Yu, Y. Zhou, J. Simmons, C. P. Przybyla, Y. Lin, X. Fan, Y. Mi, and S. Wang. Groupwise tracking of crowded similar-appearance targets from low-continuity image sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 952–960, 2016.

[51] S. Zhang, Y. Gong, J.-B. Huang, J. Lim, J. Wang, N. Ahuja, and M.-H. Yang. Tracking persons-of-interest via adaptive discriminative features. In *European Conference on Computer Vision*, pages 415–433. Springer, 2016.

[52] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *arXiv preprint arXiv:1706.00384*, 2017.

[53] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Joint face representation adaptation and clustering in videos. In *European Conference on Computer Vision*, pages 236–251. Springer, 2016.

[54] X. Zhao, D. Gong, and G. Medioni. Tracking using motion patterns for very crowded scenes. In *Computer Vision–ECCV 2012*, pages 315–328. Springer, 2012.