Face Association across Unconstrained Video Frames Using Conditional Random Fields

Ming Du and Rama Chellappa

Center for Automation Research, University of Maryland,
College Park, MD 20742, USA
{mingdu,rama}@umiacs.umd.edu

Abstract. Automatic face association across unconstrained video frames has many practical applications. Recent advances in the area of object detection have made it possible to replace the traditional tracking-based association approaches with the more robust detection-based ones. However, it is still a very challenging task for real-world unconstrained videos, especially if the subjects are in a moving platform and at distances exceeding several tens of meters. In this paper, we present a novel solution based on a Conditional Random Field (CRF) framework. The CRF approach not only gives a probabilistic and systematic treatment of the problem, but also elegantly combines global and local features. When ambiguities in labels cannot be solved by using the face appearance alone, our method relies on multiple contextual features to provide further evidence for association. Our algorithm works in an on-line mode and is able to reliably handle real-world videos. Results of experiments using challenging video data and comparisons with other methods are provided to demonstrate the effectiveness of our method.

1 Introduction

We are interested in automatically assigning identity labels to a group of faces in each frame of real-world videos. An example of face association is demonstrated in Fig. 1. A successful solution to this problem has immediate applications: video-based face recognition[1,2], automatic video annotation, automatic collection of large-scale face dataset[3], just to name a few. We believe the following two aspects are especially important for solving this problem:

Bottom-up face association. Traditionally, automatic face association is an inherent outcome of multi-object tracking algorithms: Each target face initiates an independent tracker, which searches in successive frames for the best match over a neighborhood as determined by temporal coherence constraints. In addition to suffering from the same problems as the single-face tracker, such as drift errors, the trackers are frequently confused with each other due to subject's interactions and the similarity in face appearances. In recent years, the so-called "tracking-by-detection" approaches [4,5,6,7,8,9,10,11] have gained popularity as a result of the significant progress in object detection techniques. These methods first apply object detectors to every video frame and then connect the detection

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VII, LNCS 7578, pp. 167–180, 2012.

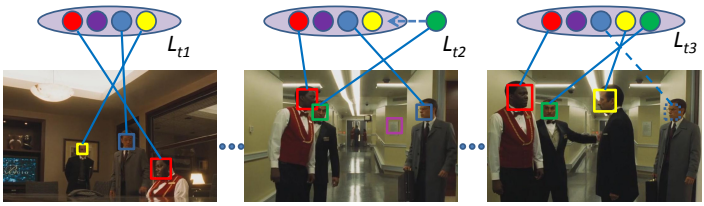

Fig. 1. Face association A face association algorithm solves the correspondence problem between face detections and the identity labels. At time $t1$, subject Purple (we use color of the corresponding label to refer to the person) does not appear in the scene. At time $t2$, A new subject, i.e. subject Green shows up and needs to be enrolled to the identity list. At the same time, there is a false detection (pink bounding box) which should not be assigned any label. At time $t3$, subject Yellow, who was absent at $t2$ returns to the scene. His face should be re-identified. The detector also misses subject Blue (the dotted bounding box), which should be retrieved by face association.

results into tracks[1]. In comparison with the tracking-based methods, the face association methods are free of drift errors, much more robust against shaky camera and occlusions and easier to recover from failure.

Context features. Recently, context-based vision has received increasing interest. Although there have been different notions of context in vision applications, it usually refers to information from regions of the image outside the region of interest and/or other sources such as maps, time stamp, etc. When speaking of face, the context can be the hair, clothing of a subject, or even other people in the image. The importance of context is even more obvious as we process real-world videos, in which low-resolution, blurred or arbitrarily-illuminated faces are more the norm than the exception. To effectively integrate evidence from both face appearance and context features, a systematic approach is needed.

Motivated by these considerations, we propose an automatic face association algorithm for real-world videos. To be specific, we solve the problem using a conditional random field (CRF), where each node represents a detected face (either true or false positives) in a video frame. Our contributions can be summarized as follows: First, we propose a general CRF-based framework to solve the face association problem for videos. Unlike many existing works in multi-object association [4,5,6], our method is an on-line procedure, which is crucial to real-world applications. Second, we leverage the abundant contextual features available in the video. The features enter the potential function through both unary and pairwise terms and significantly improve the performance of face association in terms of accuracy and robustness. Third, we introduce the concept of "null state" and apply logistic regression to handle false detections or novel faces. Therefore, the CRF in our case is dynamic in the sense that not only the features are characterized in a time-adaptive fashion, but also that the number

[1] We use the more particular term "face association" or "face labeling" to this bottom-up strategy, but avoid using the word "multi-face tracking" as used in some papers.

of nodes and states can vary at any time. All of the three aspects contribute to the proposed algorithm's capability for handling real-world videos: The data association strategy allows accurate localization, easy recovery from failures and quick scene adaption; The use of contextual features results in robust face labeling despite the presence of nuisance factors like significant camera motion, bad illumination or low-resolution; The time-varying structure handles situations when subjects enter or leave the scene. Moreover, our on-line method can process videos in a timely manner.

2 Related Work

A large body of literatures exists on object tracking. Interested readers may refer to [12] for a comprehensive review. General tracking algorithms can be extended to the multi-target case by initiating multiple independent trackers [13]. Though straightforward in idea, this approach has many limitations stated in Section 1.

Automatic face labeling from TV/film/news videos has attracted increasing interest in recent years. Most of the existing works treat face association as a constrained clustering problem. Therefore, researchers have focussed their attention in obtaining a good distance measure [14,15,16,3,2]. In most of these works, simple association strategies like the K-means or agglomerative clustering are applied, and are shown to produce promising results on videos with a small number of faces. But in many practical situations we need to process crowded scenes.

Association-based approaches have also been widely used to simultaneously localize multiple pedestrians in videos. According to the role a detector plays, they fall into three categories. In the first case, detection results are integrated into a tracking framework [7,8], functioning as part of the proposal distribution or the observation model. Alternatively, the associations can be performed directly on detected results in individual frames [4,5]. As the intermediate case, positive detections can activate trackers running for a short temporal window. The obtained "tracklets" are then connected to form global assignments[9,10,11,6]. Although substantial progress has been made, one or more of the following drawbacks are present in most of these works: 1)The method is only compatible with off-line processing. 2)Empirically determined parameters or heuristically defined function forms are used to combine multiple cues into the affinity model or the energy function. 3)Generative models and joint distributions, instead of discriminative models and conditional distributions, are employed for association. As a result, the dependence of observations on the interplay among local associations cannot be effectively captured by the model.

A CRF-based approach was suggested in [6] for multi-pedestrian association. Our approach differs from [6] in a fundamental aspect; In our work, the nodes in CRF are identity labels and face candidates detected in the current frame, while [6] defines each node to be a tracklet pair or a label pair, working only in an off-line mode. Hence, both the features and the learning process used in our work are different from theirs. Our work also differs from the recently published work in [17] in the sense that the energy function of the latter does not contain global

parameters and therefore its parameter learning is only local to each feature function. In contrast, by following the canonical CRF formulation, our model naturally inherits CRF's strength in modeling the interactions between features.

The effectiveness of contextual features has been studied for face labeling and recognition in consumer photos. Gallagher and Chen [18] learned group priors, i.e. the co-occurrence of people, and used it to resolve ambiguous label of faces in a graphical model framework. Clothing appearance has been combined with the face using an MRF model to improve the recognition accuracy in [19]. Relative poses among subjects in a family album can also be a useful cue[20].

3 Problem Formulation

Suppose there are N detected faces in the current frame F_t of the input video. Let $\mathbf{Y}_t = \{y_1, y_2, ...y_N\}$ denote the set of unknown labels we would like to associate with these faces. Let L be the number of all the subjects that have appeared in the scene up to frame F_{t-1}, then the state(label) space is $\mathfrak{L}_t = \{0, 1, 2, ..., L\}$. Here we introduce a "null" state with the label 0 to account for false detections and novel faces. Note that both the number of detected faces and the state space vary with time, and the mapping from $\mathbf{Y}_t$ to the state space is many-to-one.

We create a graph $G = (V, E)$ and let vertices $V = \{y_1, y_2, ...y_N, \mathbf{X}\}$, where $\mathbf{X}$ is a global observation node [2]. To make the maximum use of information encoded in the contexts, we let the label nodes to be fully connected to each other. CRFs model only the conditional probability $p(\mathbf{Y}|\mathbf{X})$ instead of the joint probability $p(\mathbf{Y}, \mathbf{X})$:

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \frac{1}{Z(\mathbf{X}, \theta)} \prod_{c \in C} \Psi_c(\mathbf{Y}_c|\mathbf{X}, \theta), \tag{1}$$

where C is the set of all cliques in the graph and Ψ_C is the potential function defined for clique c, θ is CRF parameter and Z is the normalization factor.

A CRF does not model the data distribution $p(\mathbf{X})$, which is what we have observed. Therefore, a CRF is capable of incorporating non-local features, and the edge potentials between the label nodes can be either dependent or independent of the observation nodes. The property makes it especially useful for modeling contextual features. We will assume the potential function to possess the log-linear form: $\ln p(\mathbf{Y}|\mathbf{X}, \theta) = \sum_{c \in C} \sum_k \theta_k f_k(\mathbf{Y}_c, \mathbf{X}) - \ln Z(\mathbf{X}, \theta)$, where $f_k(\mathbf{Y}_c, \mathbf{X})$ are feature functions. The log-linear form not only imposes positivity, but also has a close relationship to the Maximum Entropy models. At the tth frame, we are trying to solve for the optimal label configuration $\mathbf{Y}_t^*$ that maximizes the conditional probability. It is then used to renew the models used in feature functions and update the state space to $\mathfrak{L}_{t+1}$.

[2] We omit the time index to keep the notations simple as long as it does not cause any confusion.

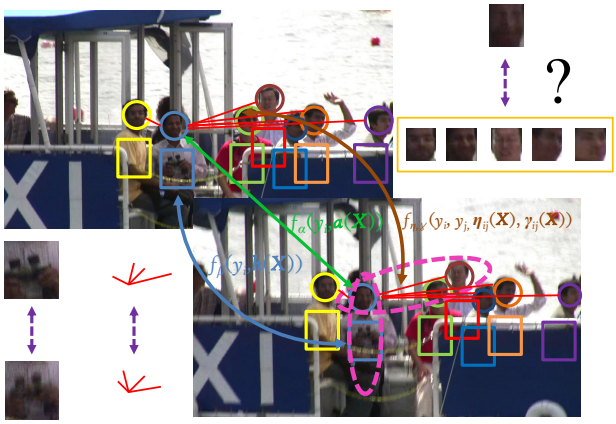

Fig. 2. Context-aided face matching The face appearance alone usually is not sufficient as a strong feature to perform association. Contextual information, such as clothing appearance and relative poses, can be incorporated to make a more confident decision. Two frames from a video of faces in a moving boat collected at distances of several tens of meters are shown in this figure.

4 Context-Aided Association

In this work, we incorporate face appearance with four kinds of contextual features: clothing appearance, relative scale, relative position and uniqueness of identity. The conditional probability is thus defined as:

$$\log p(\mathbf{Y}|\mathbf{X},\theta) = -\log Z(\mathbf{X},\theta) + \sum_{i\in V}\theta_\alpha f_\alpha(y_i,\mathbf{a}_i(\mathbf{X})) + \sum_{i\in V}\theta_\beta f_\beta(y_i,\mathbf{h}_i(\mathbf{X})) + \sum_{(i,j)\in E}\theta_\gamma f_\gamma(y_i,y_j,\gamma_{ij}(\mathbf{X})) + \sum_{(i,j)\in E}\theta_\eta^T \mathbf{f}_\eta(y_i,y_j,\eta_{ij}(\mathbf{X})) + \sum_{(i,j)\in E}\theta_\lambda f_\lambda(y_i,y_j), \tag{2}$$

where $f_\alpha, f_\beta, f_\gamma, \mathbf{f}_\eta$ and f_λ are the feature functions for the four aforementioned features, respectively. We demonstrate our use of contextual features in Fig. 2. In the following, we define each feature function individually.

4.1 Potentials and Feature Functions

Face Appearance. Face appearance provides the most direct evidence about a subject's ID, though for our case its power has been impaired by nuisance factors. We maintain an Online Appearance Model (OAM)[21] for each existing face track, motivated by the algorithm's success in modeling appearances with

strong temporal coherence. In an OAM, object appearance is represented by a mixture of three components, namely the stable, wander and lost components. The stable component models steady and long-term appearance; The wander component is responsible for modeling the short-term change in appearance; The lost component accounts for outliers. Considering that alignment errors in cropping exist due to the nature of the sliding-window detection strategy, in the OAM we use Gabor features, which can tolerate small translation and scale variations, in lieu of raw intensity values.

The model parameters are updated by an Online-EM procedure. Denote the set of Gabor coefficients of a detected face at time t as $\mathbf{a}_t = \{a_{n,t}\}, n = 1, 2, ..., N$, and the set of existing, recently updated OAMs as $\mathcal{A}_{t-1} = \{\mathbf{A}_{l,t-1}\}, l = 1, 2, ..., L$. In the E step, we calculate the ownership probabilities of the face with respect to the lth OAM:

$$\mathbf{o}_{l,q}(\mathbf{a}_t) = \frac{m_{q,t} p_q(\mathbf{a}_t|\mathbf{A}_{l,t-1})}{\sum_q m_{q,t} p_q(\mathbf{a}_t|\mathbf{A}_{l,t-1})} \ , \ q = \{wander, stable, lost\}. \tag{3}$$

p_{wander} and p_{stable} are the two normal distributions whose parameters are updated every frame for each OAM. p_{lost} is a uniform distribution over the domain of observation feature values. The feature function, which evaluates how likely a node y is in state l, is defined as:

$$f_\alpha(y = l, \mathbf{a}(\mathbf{X}_t)) = \sum_n \log \sum_q o_{l,q}(a_{n,t}) p_q(a_{n,t}|\mathcal{A}_{l,t-1}) \ , \ q = \{wander, stable\}. \tag{4}$$

The M step happens after the label of the CRF has been determined through inference. We use the appearance of the node that has been labeled as subject l to update the parameters of the lth OAM. The set of updating equations can be found in [21]. We illustrate an example of OAM in Fig. 3 (a)-(d).

Clothing Appearance. As a contextual feature, clothing appearance assists the goal of face association, especially for real-world videos because: 1) It occupies a larger area than face and hence is easier to extract from a distance. 2) The between-class variation for clothing appearance is usually more distinguishable than face appearance. Given F_c, the center of a face, we locate the torso by using a probabilistic mask $p(I \in torso|F_c)$ (See Fig. 3(e)), which is learned from the H3D (Human in 3D) dataset[22]. If the clothing histogram feature for a detection is denoted as $\mathbf{h}$, the color feature function for the tth frame is defined as:

$$f_\beta(y = l, \mathbf{h}) = \log(1 - d(\mathbf{h}, \mathbf{h}_{l,t-1}), \tag{5}$$

where d is the chi-square distance between two histograms. The histogram model $\mathbf{h}_l$ is also updated at every frame with a forgetting parameter. We assume mutual occlusion to be happening when two torsos are in significant overlap. In this situation, we associate the intersection region to the torso with more consistent color distributions.

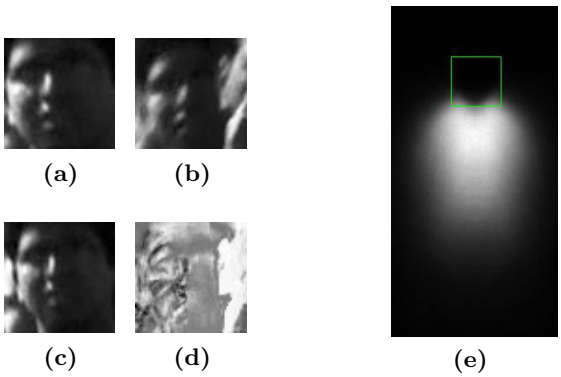

Fig. 3. The OAM and the probabilistic mask of torso From frame $t-20$ to $t-2$ there was partial occlusion, which is still present in the mean of the S(stable) component of the recently updated OAM $\mathcal{A}_{t-1}$(b). The occlusion disappeared at frame $t-2$. So in the current frame t we get a clean face $\mathbf{a}_t$(a). (c) is the mean of the W(wander) component of $\mathcal{A}_{t-1}$, which captures this recent appearance change. (d) is produced by subtracting the posterior mixture probability of S component from that of the W component. We can see that the previously occluded region is much better accounted for by the W component than by the S component. (e) The learned probabilistic mask of torso. The green square marks the position of the reference face.

Relative Pose. Shaky cameras are common for real-world videos. Unfortunately video stabilization algorithms often fail when complicated or textureless background(water surface, wall etc.) are present. However, the relative scale and distance features do not suffer as much, and they maintain a temporal coherence at the same time. Note that these features cannot be defined in a MRF framework as MRF's edge potentials cannot condition on non-local observations. We approximate the camera with a para-perspective model. This is a reasonable model since for the camera whose field of view can hold a group of people, the depth variations of the scene points that we are interested in are usually small in comparison to Z_0, the distance between the frontal plane and the image plane. Another assumption implied by the model is that the movement of a face along the camera axis is also insignificant compared to Z_0. Let the scale-normalized distance(SND) between the images of two rigid objects A and B be $\mathbf{\Delta}_{AB} = [(\mu_B - \mu_A)/\omega_A, (\nu_B - \nu_A)/\omega_A]$, where (μ_A, ν_A) is the image coordinate of A's center, ω_A is the size of A's image, and so on. It is easy to show that, when the focal length and the principal point of the para-perspective camera changes, the difference of $\mathbf{\Delta}_{AB}$ between two consecutive frames satisfies: $\mathbf{\Delta}_{AB,t} - \mathbf{\Delta}_{AB,t-1} = \delta_{AB}/\tau$, where δ_{AB} is the displacement of A w.r.t. B in the world coordinate system (we disregard the camera-axis direction for the aforementioned reason) during the same time interval and τ is a constant factor. That is, the SND's change is only dependent on the object's motions and is independent of the camera's zoom or translation.

We define the relative distance feature function as:

$$f_{\eta,\mu}(y_i = l_1, y_j = l_2, \eta_{ij}(\mathbf{X}_t)) = \log \mathcal{L}(\Delta\mu_{i,j,t} - \Delta\mu_{l_1,l_2,t-1}|m_\mu, b_\mu)$$
$$f_{\eta,\nu}(y_i = l_1, y_j = l_2, \eta_{ij}(\mathbf{X}_t)) = \log \mathcal{L}(\Delta\nu_{i,j,t} - \Delta\nu_{l_1,l_2,t-1}|m_\nu, b_\nu), \quad (6)$$

where $\mathcal{L}$ is the Laplace distribution: $\mathcal{L}(x|m, b) = \frac{1}{2b}\exp(-\frac{|x-m|}{b})$. The choice of Laplace distribution over the more frequently used Gaussian distribution is justified by two considerations: First, the Laplace distribution has longer tails, therefore it is more robust against outliers. Second, in our experiments, the Laplace distribution can approximate the empirical distribution of the features more accurately (See Fig .4). Parameters of the Laplace distributions are estimated from the training data using the maximum likelihood method.

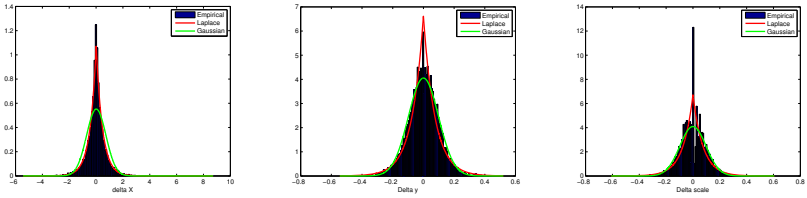

Fig. 4. Distributions of relative positions The empirical distributions are visualized using histograms. The fitted Laplace distribution is plotted in red, and the Gaussian distribution is plotted in green. Parameters are set as the maximum likelihood estimates. The distribution of the x-direction distance variation is much larger than that of the y-direction, which makes sense since the human moves horizontally much more often than vertically.

In a similar fashion, we define the feature function for relative size as:

$$f_\gamma(y_i = l_1, y_j = l_2, \gamma_{ij}(\mathbf{X}_t)) = \log \mathcal{L}(\frac{\omega_{j,t}}{\omega_{i,t}} - \frac{\omega_{l_2,t-1}}{\omega_{l_1,t-1}}|m_\gamma, b_\gamma) \quad (7)$$

Uniqueness. The uniqueness constraint follows intuitively from a simple fact: No person can appear more than once in the same frame. However, the constraint does not apply to the null state that will be discussed in Section 4.2, as multiple new faces and false detections can be present at the same frame. This feature function has the following form:

$$f_\lambda(y_i, y_j) = \begin{cases} -\inf & \text{if } y_i = y_j \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

As we can see, this hard constraint dominates other feature functions and enforces a zero probability when it is violated, but has no influence when it is satisfied. Thus, θ_λ is fixed as 1.

4.2 The Null State

So far, the definitions of the feature functions have not considered the issue of the null state. In other words, we have not defined $f(y_i, \mathbf{X})$ or $f(y_i, y_j, \mathbf{X})$ when $y_i = 0$ or $y_j = 0$. It is very challenging to explicitly model the null state, which is an open-universe set. The problem becomes even more complicated when we attempt to guarantee the function value for the null state to be in an appropriate numerical scale comparable to that of other states.

If we denote the domain of a unary feature function f as $\mathfrak{X}$, then f define a map $Z : \mathfrak{X} \to R^L$:

$$Z_p(\mathbf{X}) = \mathbf{f}_p = [f(y_i = 1, \mathbf{X}), f(y_i = 2, \mathbf{X}), ..., f(y_i = L, \mathbf{X})]^T, \tag{9}$$

where L is the number of the non-null states. We now construct a second map: $Z'_p : R^L \to I^{L+1}$ as follows:

$$Z'_p(\mathbf{f}) = \mathbf{f}'_p = [f'_0, f'_1, ..., f'_L] \ , \ f'_l = \frac{e^{\mathbf{w}_l^T \phi(\mathbf{f})}}{\sum_{l'=0}^{L} e^{\mathbf{w}_{l'}^T \phi(\mathbf{f})}}, \tag{10}$$

where I is the closed interval $[0, 1]$, and ϕ is a set of nonlinear basis functions. The pairwise case is a little more complicated. We define a "null state set" $\mathcal{N} = \{(l_1, l_2) | l_1 = 0 \vee l_2 = 0\}$ for the edges of the CRF, which contains $L + 1$ elements. We can similarly learn a map for a pairwise feature function $Z'_q : R^{(L^2)} \to I^{(L+1)^2}$, but with an additional constraint: $\forall (l_1, l_2) \in \mathcal{N} : f'_{(l_1,l_2)} = \rho$. This is intuitive as there is no reason to favor one null state in $\mathcal{N}$ over another in the eye of an pairwise feature function. Although the logistic regression is a classification algorithm, its output is continuous and so can be interpreted as class-conditional probabilities. So the models define a map for feature functions, with the desired property that their outputs with respect to different states have comparable magnitudes. Note that we need to learn a model for each different case of state numbers. We trained models for $L = 2, ..., 20$.

4.3 Parameter Estimation and Inference

The parameters of the CRF are estimated using a regularized maximum likelihood procedure: Given M labeled data pair $\{\mathbf{X}^{(m)}, \mathbf{Y}^{(m)}\}_{m=1,...,M}$, we maximize:

$$E = L + \lambda \|\theta\|^2 = \sum_{m=1}^{M} \log p(\mathbf{Y}^{(m)} | \mathbf{X}^{(m)}, \theta) + \lambda \|\theta\|^2. \tag{11}$$

For this purpose we need to compute

$$\frac{\partial L}{\partial \theta_p} = \sum_{m=1}^{M} [\sum_{i \in V} f_p(y_i^{(m)}, \mathbf{X}^{(m)}) - \sum_{\mathbf{Y} \in \mathfrak{Y}} p(\mathbf{Y} | \mathbf{X}^{(m)}, \theta) \sum_{i \in V} f_p(y_i, \mathbf{X}^{(m)})] \tag{12}$$

$$\frac{\partial L}{\partial \theta_q} = \sum_{m=1}^{M} [\sum_{(i,j) \in E} f_q(y_i^{(m)}, y_j^{(m)}, \mathbf{X}^{(m)}) - \sum_{\mathbf{Y} \in \mathfrak{Y}} p(\mathbf{Y} | \mathbf{X}^{(m)}, \theta) \sum_{(i,j) \in E} f_q(y_i, y_j, \mathbf{X}^{(m)})], \tag{13}$$

where f_p and f_q are unary and pairwise functions. Note this requires an enumeration of all the possible configurations in the label space $\mathfrak{y}$, which is generally infeasible. Alternatively, we use the Gibbs sampler to generate samples from the label space and infer $p(\mathbf{Y}|\mathbf{X}^{(m)}, \theta)$. At the test time, the MAP solution: $\mathbf{Y}^* = \underset{\mathbf{Y}}{\text{argmax}}\, p(\mathbf{Y}|\mathbf{X}, \theta)$ is solved by the Loopy Belief Propagation (LBP) algorithm. Occasionally the LBP may fail to converge, in which case we switch to the variational Mean Field algorithm for a max marginal solution.

4.4 Removal of False Detections and Recovery of Missed Detections

Depending on the detectors applied, some faces may be missed and some detections may be false positives. Detections assigned with a null label become candidates for false positive or novel faces. They are examined for a number of consecutive frames. Those with low re-appearing rate are considered as false positives and discarded, and the others are kept as novel faces. For the faces which are previously detected but are missing from the current MAP solution, we propose hypothesis according to their most recent SND features w.r.t. other subjects. We then generate samples of bounding boxes $\{s_i = (\mu_i, \nu_i, \omega_i)\}_{i=1,\ldots,N}$ over the neighborhood of the hypothesis and evaluate for them the unary feature functions: $f_\alpha(y_{s_i} = l_m, \mathbf{a}(s_i))$, $f_\beta(y_{s_i} = l_m, \mathbf{h}(s_i))$, where l_m is the label of the missed face. Only if one or more samples yield function outputs significantly higher than a conservatively-set threshold will a recovery of the missing face be enforced. Otherwise the hypothesis is simply rejected.

5 Experiments

Dataset. We collected a real-world video dataset in outdoor environments consisting of 67 video sequences at 10 different scenes. The number of subjects showing up in each frame ranges from 2 to 14. The database is challenging in the following aspects: 1) Videos were acquired at a distance ranging from 50 to 200 meters. Such a capture distance results in low-resolution face images. 2) We intentionally introduce perturbations due to zoom-in and zoom-out operations. The level of camera shaking can be roughly measured by the displacement-scale ratio (DSR), i.e. the displacement of a face's image center due to camera motion/the width(height) of the face, both in pixels. It is not rare to see DSRs as large as 1.5 or even 2.5 in our databases, i.e. within two consecutive frames, the center of a 50×50 bounding box shifts 125 pixels due to camera motion. 3) Face appearances are subject to blur, pose variations and occlusions caused by sunglasses, cell phones, scene objects or other people. 4) The illumination condition is uncontrolled and could often be extreme in its variation. We use an independent set of 42 short video sequences for parameter estimation. This 10 GB dataset together with the source code will soon be released on our website.

Face Detection. We apply the cascaded Harr-feature face detector [23] to each frame, followed by a skin detector using HSV color space thresholding. We mark

those face candidates with a unreasonable portion of skin pixels (We empirically determine the upper and lower threshold to be 0.85 and 0.2) as tentative false positives and the remaining ones as tentative true positives. The CRF inference is applied to the tentative true positives. However, the tentative false positives are not simply discarded. When their locations are coincided with a sample in the missing face recovery procedure described in Section. 4.4, we will assign bonus score to that sample.

Evaluation Metrics. To qualitatively evaluate the performance of our algorithm, we borrow the set of metrics commonly used in multi-pedestrian tracking works, including:

- **GT**: the number of ground-truth face tracks.
- Recall: correctly labeled faces / total ground-truth detections.
- Precision: correctly labeled faces / total labelings made.
- **MA**: the percentage of mostly associated face tracks, which are correctly labeled for more than 80%.
- **MW**: the percentage of mostly wrongly associated trajectories, which are correctly labeled for less than 20%.
- **PA**: the percentage of partially successfully associated face tracks.
- **Frag**: fragments, the number of times that a ground-truth face track is interrupted.
- **IDS**: ID switch, the total number of times that a ground-truth face track changes its associated label.

Results. Sample face association results on the video database are presented in Fig. 5. The white bounding boxes mark the detections which are labeled by our algorithm as false positives. The black ones mark the faces that are recovered by our scheme as presented in Section 4.4. As can be seen from the result, despite large scene variation, our method is able to reliably associate faces: Faces can be re-identified after occlusion moves away (e.g. subject 5 in the 4th sequence); The false detections can be removed accurately; Even those "lost" faces can be retrieved. Some errors do exist in our results. The most common failure mode of our method happens when a previously occluded subject re-appear with a new face pose. As a result, both face appearance and relative distance/scale features may fail to associate the re-appearing face with existing tracks. An example of this can be found in the 1st sample sequence of Fig. 5, where subject 4 was labeled as a new subject when being detected again. (Note, however, that the subject marked with a null label in the last image of the 3rd sequence is not an error. It is a novel face and is in its examination phase before finally accepted into the ID list, as discussed in Section 4.4.) Another major source of error is the uncertainty of localization existing in the output of face detector. Although the Gabor filter help alleviate the problem to some extent, sometimes the bounding boxes of the same face in consecutive frames can exhibit high variations in scale and position. This kind of error affects all the feature functions. An efficient face registration module can improve the performance of the current implementation and therefore is one of our future directions.

We also compare our result with two alternative approaches. The first one is similar to the approaches used in [16,2,1]. In this method, we first connect frame-based detection results with a short-duration face tracker. Whenever the tracking result has significant overlap with a detection result both in spatial location and in appearance, it will be reset by the latter. The tracker will stop when it fails to find such an overlap detection for five consecutive frames. The intermediate gap between detections shorter than five frames will be bridged by the tracker. The obtained tracklets are then grouped by applying constrained agglomerative clustering. The "cannot link" constraint is the same as the uniqueness constraint mentioned in 4.1. The face tracker is based on a particle filter and the OAM observation model. We set the number of particles to be 200. The tracker worked on videos stabilized according to SIFT feature correspondences. Note that our CRF method is applied to the original, un-stabilized videos. The other compared approach is the min-cost flow algorithm as proposed in [5]. It is based on a generative Hidden Markov Model, which does not incorporate observations into the transition probability. Comparisons in terms of the criteria defined above are summarized in Table 1. As shown, our method clearly outperforms the Tracking-Clustering method and has better

Fig. 5. Sample face association results Images of each row are from the same video sequence. The number of frames in-between is 100. The subjects in the 1st and the 4th rows are captured by remote cameras located at more than 50 meters away. The subjects in the 2nd row are on the shore and were imaged by cameras on a boat moving towards them. The subjects in the 3rd row are in a moving boat. The figure is best viewed in color and with pdf magnification.

overall performance than the Min-Cost method. Besides, both of these alternative approaches work in an off-line mode, but ours operates in an on-line fashion, which is more difficult and more advantageous for real-world applications. Also shown in this table are the results obtained by removing the face/clothing/relative pose feature function each at a time. We keep the uniqueness feature all the time, otherwise we would get a very noisy result. We see that face appearance is still the most important evidence, as the performance drops most drastically after we remove it from the feature set. Without this feature, the ID switch error will be frequent because of the confusion caused by similar clothing (For example, the second sample sequence of Fig. 5). Relative distance and scale feature rank the second place in terms of importance, especially for videos which are captured by a shaky camera or on a moving platform. Disregarding this feature leads to high occurrences of fragment error. This mainly happens when blur or occlusion causes appearance features to be unreliable and spatial correlation is the only possible cue to maintain correspondences. Performance degradation after dropping the clothing feature can be mostly accounted by cases in which blurred and low-resolution faces are present. In general, every feature play an important role in this framework, and we achieve the best performance by combining all of them using the CRF framework.

Table 1. Face association results on the real-world database

Method	GT	Recall	Precision	MA	MW	PA	Frag	IDS
Tracking-Clustering	371	70.2	74.4	63.6	11.1	25.3	133	167
Min-Cost Flow	371	74.1	77.9	72.8	10.0	17.2	120	124
CRF (no face feature)	371	58.6	60.5	57.7	20.5	21.8	163	199
CRF (no clothing feature)	371	71.7	74.2	67.5	12.4	20.1	135	146
CRF (no relative distance/scale feature)	371	66.5	69.3	60.1	14.0	25.9	172	151
CRF	371	81.6	83.5	78.2	8.6	13.2	121	101

6 Conclusions

In this paper, we presented an on-line face association algorithm. The method is based on CRFs, whose property allows us to conveniently incorporate multiple contextual features. The algorithm is able to deal with blurred or low-resolution faces, shaky/zooming camera, occlusion and subject entry/exit, and hence is suitable for real-world video processing. Our algorithm has achieved promising experimental results on a challenging face association database.

Acknowledgments. This work was partially supported by a MURI on Remote Biometrics in the Maritime Domain from the Office of Naval Research under the Grant N00014-08-1-0638.

References

1. Sivic, J., Everingham, M., Zisserman, A.:"Who are you?" – learning person specific classifiers from video. In: CVPR, pp. 1145–1152 (2009)
2. Everingham, M., Sivic, J., Zisserman, A.: "Hello! my name is.. buffy" – automatic naming of characters in tv video. In: BMVC, vol. 3, pp. 899–908 (2006)

3. Ramanan, D., Baker, S., Kakade, S.: Leveraging archival video for building face datasets. In: ICCV, pp. 1–8 (2007)
4. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
5. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR, pp. 1201–1208 (2011)
6. Yang, B., Huang, C., Nevatia, R.: Learning affinities and dependencies for multi-target tracking using a crf model. In: CVPR, pp. 1233–1240 (2011)
7. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.J.V.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009)
8. Cai, Y., de Freitas, N., Little, J.J.: Robust Visual Tracking for Multiple Targets. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 107–118. Springer, Heidelberg (2006)
9. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR, pp. 1–8 (2008)
10. Huang, C., Wu, B., Nevatia, R.: Robust Object Tracking by Hierarchical Association of Detection Responses. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008)
11. Song, B., Jeng, T.-Y., Staudt, E., Roy-Chowdhury, A.K.: A Stochastic Graph Evolution Framework for Robust Multi-target Tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 605–619. Springer, Heidelberg (2010)
12. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. ACM Comput. Surv. 38 (2006)
13. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. PAMI 26, 1208–1221 (2004)
14. Fitzgibbon, A.W., Zisserman, A.: On Affine Invariant Clustering and Automatic Cast Listing in Movies. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 304–320. Springer, Heidelberg (2002)
15. Berg, T.L., Berg, A.C., Edwards, J., Maire, M., White, R., Teh, Y.W., Learned-Miller, E., Forsyth, D.A.: Names and faces in the news. In: CVPR, vol. 2, pp. 848–854 (2004)
16. Sivic, J., Everingham, M., Zisserman, A.: Person Spotting: Video Shot Retrieval for Face Sets. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 226–236. Springer, Heidelberg (2005)
17. Yang, B., Navatia, R.: An online learned crf model for multi-target tracking. In: CVPR (2012)
18. Gallagher, A.C., Chen, T.: Using group prior to identify people in consumer images. In: CVPR, pp. 1–8 (2007)
19. Anguelov, D., Lee, K.C., Gokturk, S.B., Sumengen, B.: Contextual identity recognition in personal photo albums. In: CVPR, pp. 1–7 (2007)
20. Gallagher, A.C., Chen, T.: Using context to recognize people in consumer images. IPSJ Transactions on Computer Vision and Applications 1, 115–126 (2009)
21. Jepson, A.D., Fleet, D.J., El-Maraghi, T.: Robust online appearance model for visual tracking. In: CVPR, vol. 1, pp. 415–422 (2001)
22. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3d human pose annotations. In: International Conference on Computer Vision (2009)
23. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV 57, 137–154 (2004)