Detection-Based Multi-Human Tracking Using a CRF Model

Alexandre Heili[1,2] Cheng Chen[1] Jean-Marc Odobez[1,2]
[1] Idiap Research Institute – CH-1920 Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne – CH-1015, Lausanne, Switzerland
{aheili,cchen,odobez}@idiap.ch

Abstract

In surveillance videos, the task of tracking multiple people is of primary importance and is often a preliminary step before applying higher-level algorithms, e.g. to analyze interactions or to recognize behaviors. In this paper, we take a tracking-by-detection approach and formulate multi-person tracking as a statistical data association problem which seeks for the optimal label field in which detections belonging to the same person have the same label. Specifically, unlike most previous works that rely on generative approaches, we use a Conditional Random Field (CRF) model, whose pairwise detection factors, defined for both distance and color features, are modeled using a two-hypothesis framework: a pair of detections corresponds either to the same person or not. Parameters of these two-hypothesis model factors are learned in a fully unsupervised way from data. Optimization is conducted using a deterministic sliding window method. Qualitative and quantitative results on several different surveillance datasets show that our method can generate robust and accurate tracks in spite of the noisy output of the human detector and of occlusions.

1. Introduction

In video surveillance context, multi-person tracking is a very important topic. Its solving can benefit many applications. For example, knowing the location of different people over time can greatly help the semantic analysis of video, such as group/interaction detection [17, 16], scene understanding [18] and so on. On the other hand, the output of a multi-person tracker can be fed to some higher level process such as behavior cue extraction for action/event recognition [8]. However, multi-human tracking remains a challenging task, especially in single camera tracking situations, or in multi-camera cases with small overlap or high crowding, notably due to low image quality, sensor noise, dimension loss due to projection of 3D objects in image planes, occlusions, clutter, unpredictable motions and appearance changes of people.

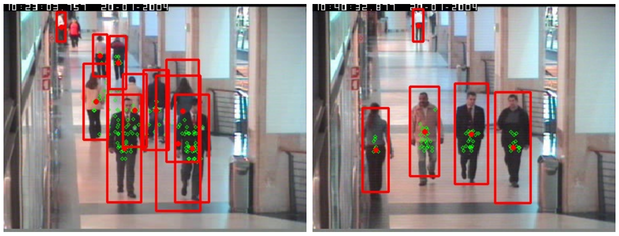

Figure 1. Examples of detector outputs showing (left) a missed detection and a false alarm; (right) detection accuracy issues, like legs cut or extended due to projected shadows on the floor.

As task-specific object detectors become more and more reliable, one approach for multi-person tracking is to rely solely on the output of human detectors, which is called "tracking-by-detection" or "detection-based tracking". In this paradigm, human detection is performed first on the images. Then, the tracking step attempts at associating the detections corresponding to the same person by assigning labels to the detection outputs. The main advantage is that discriminatively trained detectors are often more powerful at assessing the presence of humans in an image compared to standard generative models. Another advantage is that no manual (re-)initialization is needed since it is implicitly handled through the use of the detector output at every frame. However, to be successful, it is important to deal with human detector inherent flaws: missed detections and false alarms, but also unprecise localization and size due to the presence of projected shadows or partial occlusion for instance. Another more general challenge lies in the fact that people often have similar appearances. Some of these different challenges are illustrated in Figure 1.

In this paper we propose a new detection-based multi-person tracking method. The relationships between human detections are modeled using a *CRF model* whose factor terms encapsulate the likelihoods of detection pairs *within a short-time interval*. As a consequence, unlike almost all previous methods, the association of detection pairs is not only based on a similarity measure but on a *dissimilarity measure* as well. In addition, the parameters of the CRF factors are learned automatically in an unsupervised manner, allowing the model to adapt itself to different settings. The main contributions of this paper are the following:

- embedding of the multi-person tracking problem into a CRF framework with pairwise similarity and dissimilarity hypotheses;
- an unsupervised way to learn model parameters; and
- an efficient sliding window optimization algorithm to perform the labeling.

Qualitative and quantitative experiments on surveillance data validate our method.

The rest of the paper is organized as follows. Section 2 discusses related works. Section 3 elaborates on our CRF model and parameter learning. Section 4 explains our optimization procedures, while Section 5 presents our experimental results. Section 6 concludes the paper.

2. Related Work

Several approaches to multi-object tracking have been proposed in the literature. One class of methods relies on a Bayesian framework in which the states are recursively estimated using sequences of observations [12] [20]. In this section however, we will focus on summarizing the state-of-the-art detection-based multiple-object tracking papers which are directly related to our work.

Conventionally, the deterministic approach to multi-object tracking based on detections attributes a cost to each association between the detected objects in successive frames based on motion constraints [21] and object descriptors such as color. The problem is then formulated as a combinatorial optimization and can be solved for example with optimal assignment methods like the Hungarian algorithm or greedy search methods. However, this one-to-one correspondence scheme is very sensitive to local ambiguities and cannot handle directly occlusions, entries or exits.

Instead of solving a frame-to-frame correspondence problem, the association can also be performed on a multi-frame basis [10, 22, 3]. Dependencies are then often modeled using graphs, and the optimization problem then consists in finding the best paths between all the detections in separate frames. The process can be applied on potentially large time windows, so as to overcome the sparsity in the detection sets induced by missed detections and also to deal with false alarms, but the complexity of the optimization increases rapidly.

Alternatively, to reduce the computational cost and progressively increase the temporal range for correspondences, hierarchical approaches can be considered. For instance, in [11], the lower level associates pairs of detections based on their similarity in position, size and appearance. The resulting tracklets are fed into a Maximum A Posteriori (MAP) association problem which is solved by the Hungarian algorithm, and further refined at a higher level to model scene exits and occluders. As there are fewer tracklets than detections, the complexitiy of the optimization is reduced, but any wrong association made at the low-level is then propagated to the next level.

In terms of optimization, flow-based techniques have notably been used. In [23], the authors use the same MAP formulation as in [11] but embed it in a network framework where min-cost flow algorithm can be applied. The authors of [6] directly formulate the problem as finding the flow of humans on a discrete grid space that maximizes the cost of going through the detections. This formulation yields an objective function which is a linear expression of the estimated number of objects at each time and location of the grid. The main advantage is that by assuming a continuous version of the problem, Linear Programming techniques can be applied and the global optimum can be reached. Impressive results are obtained, but only results in overlapped multiview indoor room scenarios are shown, where relatively clean detections from background subtraction images are used. This use of multiple cameras to tackle the occlusion issue that typically arises as the video is more crowded is quite usual. However, in surveillance applications, this is often not the common case, and it poses the problem of data integration and synchronization between several sensors.

To the contrary of the methods described above, by formulating the tracking as a CRF problem, our approach does not only optimize the label field on a similarity hypothesis basis, but also relies on a dissimilarity information to assess the labeling. By contrasting the two hypotheses for each detection pair, the model it more robust to assess the appropriateness of a given association. This effect is reinforced by connecting detection pairs not only between adjacent frames, but between frames within a short time interval (from ±0.5s to ±2s). Our method takes inspiration from the framework of [13], which addresses the problem of tracking sound sources in a one dimensionnal space and showed that robust short-time clusters can be obtained.

Note that recently the authors of [4] also used a CRF in a tracking-by-detection context. However, their approach is different insofar as they model the affinities and dependencies between tracklets and do not work at the detection level. Moreover, the optimization they propose works offline, they do not model dissimilarities explicitly, and the model parameters are learned through supervised training rather than using an unsupervised data-driven approach as we do.

3. CRF Modeling with Two-Hypothesis Factors

In this Section we first introduce the model and then present its components.

3.1. Problem Formulation

Let us assume that the human detection step has been performed on each frame of a video sequence. The set of detection outputs $R = \{r_i\}_{i=1:N_r}$ is the input to our tracker, where N_r is the total number of detections. Each detection r_i consists of a set of observations, which include an occurrence time t_i (or frame number), as well as some features. In this paper, two features are used : X_i, the position of the detection expressed in the ground plane, and h_i, the color descriptor. X_i is calculated by projecting the bottom center of the detection bounding box into the ground plane, assuming the camera calibration or ground plane homography is available. As color descriptor h_i, we used the multi-resolution color histogram in the HSV color space. We use multi-resolution color histogram to reduce the quantization effects. Morover, to avoid taking many pixels from the background, the color histograms are computed within an elliptical region enclosed in the detection bounding boxes. Additional features like optical flow could be incorporated in future work. Thus, a detection is represented by $r_i = (t_i, X_i, h_i)$.

The task of multi-object tracking consists in linking those detections across frames, using some similarity measures. This task can be formulated as a labeling problem, where we want to assign labels to detections according to the identity of the object they represent. Let us define the label field $L = \{l_i\}_{i=1:N_r}$ for that purpose, where l_i denotes the label identity for detection r_i. Detections corresponding to the same object should possess the same label, meaning there would be ideally one label per track.

We want to find the label field which maximizes the posterior probability $p(L|R)$. In a traditional generative model, we can use a Maximum A Posteriori (MAP) formulation, and can equivalently maximize $p(R|L)p(L)$. Typically, $p(L)$ defines a prior over the label field and is often modeled as Markov Random Fields (MRF) decomposed as a product of potential functions over the maximal cliques, and assuming conditional independences of the detections, $p(R|L) = \prod_i p(r_i|l_i)$ denotes the data likelihood. Note that such an approach is not appropriate for association, since we do not know in advance the number of classes and the term $p(r|l)$ only involves one detection and cannot be defined in advance. Rather, in this paper, we adopt a Conditional Random Field (CRF) formulation [15], and model directly the conditional probability as follows:

$$p(L|R) = \frac{1}{Z(R)} \left(\prod_{(i,j)} \prod_{k=1}^{N_{f_2}} \Phi_k(l_i, l_j, r_i, r_j) \right) \cdot \left(\prod_i \prod_{l=1}^{N_1} \Psi_l(l_i, r_i) \right) \cdot \Omega(L) \tag{1}$$

where the Φ_k denote the N_{f_2} pairwise factors, the Ψ_l denote the N_1 unitary factors, and $Z(R)$ is a normalization term which does not depend on L, and $\Omega(L)$ is a factor on the whole label field. Figure 2 illustrates the factor graph of our model. Unlike the generative approach which represents the joint probability distribution, CRFs present the advantage that no assumptions on the dependencies among the observed variables r_i need to be specified, as we directly model the label field conditional distribution.

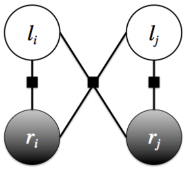

Figure 2. Factors of our Graphical Model. Shaded nodes denote observed variables, whereas unshaded node represent hidden ones.

Given this global model, the main technical points to address are the definition of the Φ_k functions, the learning of their parameters, and the optimization procedure. They are described next.

3.2. Two-Hypothesis Factors Φ_k

In this paper, we only consider the pairwise factors and omit the other terms. The definition of the pairwise factors follows a two hypotheses short-term framework. More precisely, each pair of detections in the factor is supposed to follow either one of two hypotheses, namely, either the pair corresponds to a same object (hypothesis H_1) or not (hypothesis H_0). And we only consider a predefined short term horizon T_{short} and impose $\Phi_k(l_i, l_j, r_i, r_j) = 1$ if $|t_i - t_j| > T_{short}$. In other words, in the graph there is no link between two detection nodes if they are more than T_{short} frames apart. Following this approach, the factors are defined as:

$$\Phi_k(l_i, l_j, r_i, r_j) = p_k(f_k(r_i, r_j)|H(l_i, l_j)) \quad \text{if} \quad 1 \leq |t_i - t_j| \leq T_{short} \tag{2}$$

where $H(l_i, l_j) = H_0$ if $l_i \neq l_j$, $H(l_i, l_j) = H_1$ if $l_i = l_j$, and $f_k(r_i, r_j)$ denotes a similarity measure between detections for the k^{th} factor.

We use $N_{f_2} = 2$ similarity functions, one for each of our features. For the position, the similarity is simply the Euclidean distance in the ground plane. For the color, the similarity is evaluated as the Bhattacharyya distance D_h between the multi-resolution histograms of the two detections:

$$\begin{cases} f_1(r_i, r_j) = X_i - X_j \\ f_2(r_i, r_j) = D_h(h_i, h_j) \end{cases} \tag{3}$$

Finally, the probability distributions for each feature type are defined as follows. For the position feature, we assume

Algorithm 1 Position model parameter learning

```
for $T = 1$ to $T_{short}$ do
    Initialize empty set $S_1^T$
    for $i = 1$ to $N_r$ do
        $j = \arg\min_{k \text{ } s.t. \text{ } |t_k - t_i| = T} |X_i - X_k|$
        $m = \arg\min_{k \text{ } s.t. \text{ } t_k = t_j \text{ } and \text{ } k \neq j} |X_i - X_k|$
        Add $f_1(r_i, r_j)$ and $f_1(r_i, r_m)$ to the set $S_1^T$
    end for
    Learn parameters $\Sigma_T^{\text{diff}}$, $\Sigma_T^{\text{same}}$ through EM from $S_1^T$
end for
```

that it follows a Gaussian distribution whose Covariance depends on the two label hypotheses H_0 or H_1, and also *on the time gap* $|t_i - t_j|$:

$$\begin{cases} p(f_1(r_i, r_j)|H_0) = \mathcal{N}(X_i - X_j \; ; \; 0, \Sigma_{|t_i - t_j|}^{\text{diff}}) \\ p(f_1(r_i, r_j)|H_1) = \mathcal{N}(X_i - X_j \; ; \; 0, \Sigma_{|t_i - t_j|}^{\text{same}}) \end{cases} \tag{4}$$

For the color feature, we use a non-parametric model. More precisely, we discretize the Bhattacharrya measure, and its probability using a multinomial over these indices:

$$\begin{cases} p(f_2(r_i, r_j)|H_0) = \mathbf{m}_{|t_i - t_j|}^{\text{diff}}(D_h(h_i, h_j)) \\ p(f_2(r_i, r_j)|H_1) = \mathbf{m}_{|t_i - t_j|}^{\text{same}}(D_h(h_i, h_j)) \end{cases} \tag{5}$$

where $\mathbf{m}(D)$ denotes the probability of the multinomial for the index D. In the next paragraph, we explain how we learn these model parameters from training data in an unsupervised way. In summary, following all our assumptions, the posterior probability we want to maximize is defined as:

$$p(L|R) = \frac{1}{Z(R)} \prod_{\substack{(i,j) \\ |t_i - t_j| \leq T_{short}}} \prod_{k=1}^{N_{f_2}} p(f_k(r_i, r_j)|H(l_i, l_j)) \tag{6}$$

Since we maximize over the label field L, the normalization term $Z(R)$ can be omitted during the optimization.

3.3. Unsupervised Model Training

The goal of the training phase is to learn model parameters automatically from the data to avoid manual setting. In practice, for a given time interval T, we collect for all detections the ground-plane distances to their closest and second closest detections separated by T time steps, from which we learn the parameters of the assumed two-component Gaussian Mixture Model through Expectation-Maximization (EM). The means are constrained to zero and only the covariances ($\Sigma_T^{\text{diff}}, \Sigma_T^{\text{same}}$) are trained. Algorithm 1 sums up the automatic learning procedure for this model.

Figure 3 shows the learning result for $T = 3$. We can see that the Gaussian distribution representing hypothesis H_1 is more peaky, which makes sense because under a short time interval, detections representing the same object are more likely to be close than to be far apart. In addition (not shown here), we notice that the spread of the covariances increases w.r.t. the interval T for hypothesis H_1 but not for hypothesis H_0, as one would expect.

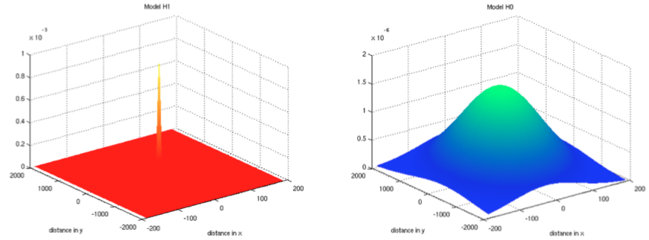

Figure 3. Learned position model for T=3: H_1 (left) and H_0 (right)

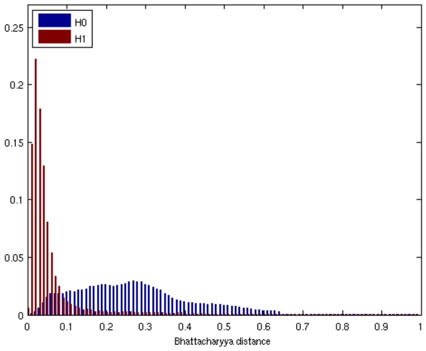

Figure 4. Learned histogram of Bhattacharyya distances for T=3: H_1 (red) and H_0 (blue)

A similar approach is used to learn the multinomial parameters (equivalent to an histogram of Bhattacharyya distances) of the color model. However, given the non-parametric nature of the model, collecting all the measures and assuming they follow a two-component model is not possible for training. We thus adopt a more straightforward method. For each time interval T, and for each detection, we assume that its closest detection (if it exists) in the ground plane from all detections T frames apart in the past and the future corresponds to the same object. The Bhattacharyya distance between their respective color histograms is thus added (after discretization) to the corresponding multinomial histogram under hypothesis H_1. Similarly, the second closest detections are used to estimate the color model parameter under hypothesis H_0. The resulting histograms are then smoothed by applying a moving average algorithm to account for the limited amount of training data. Figure 4 illustrates the result for $T = 3$. We see that under hypothesis H_1, pairs are more likely to be very similar in terms of color.

4. Optimization

4.1. Energy Minimization Formulation

At testing time, the goal is to find the optimal label field by maximizing our objective function given in Equation 6. It can be shown that it is equivalent to minimizing the energy function [9]:

$$U(L) = \sum_{\substack{(i,j)\\ |t_i - t_j| \leq T_{short}}} \beta_{ij}^{\text{potts}}.\delta(l_i - l_j) \qquad (7)$$

where $\delta(.)$ denotes the Kronecker function ($\delta(a) = 1$ if $a = 0, 0$ otherwise), and the potentials between each pair (also called Potts coefficient) are defined by:

$$\beta_{ij}^{\text{potts}} = \log \left[\frac{\prod_{k=1}^{N_{f_2}} p(f_k(r_i, r_j)|H_0)}{\prod_{k=1}^{N_{f_2}} p(f_k(r_i, r_j)|H_1)} \right] \qquad (8)$$

The interpretation is the following. If $\beta_{ij}^{\text{potts}}$ is negative, the minimization encourages to have $\delta(l_i - l_j) = 1$, which means that the pair (r_i, r_j) is more likely to correspond to the same object (hypothesis H_1 prevails) to an extent related to the amplitude of β_{ij}. On the contrary, when $\beta_{ij}^{\text{potts}}$ is positive, hypothesis H_0 prevails.

4.2. Sliding Window Solution

There exist several methods to optimize the energy function in Equation 7. In this paper, we propose to use an online Sliding Window (SW) algorithm, which performs the optimization in an iterative manner. We have also tested a global stochastic method, namely Simulated Annealing (SA), but in addition to being offline and computational intensive, it did not produce better results.

The SW algorithm works as follows. Let F_t denote all the detections in frame t, and P_t denote all past detections from frame $t - T_{short}$ to frame $t - 1$. That is, F_t contains the detections we want to assign labels to in the current step, and P_t contains all the detections which already have a label and which have a link to detections in F_t under the T_{short} horizon constraint. Note that for each detection r_i in F_t, the potential label $pot(r_i)$ can either be one of the labels in P_t (indicating an existing tracklet), or a new label (indicating the emergence of a new person, or noise): $pot(r_i) = \{unique(l_{P_t})\} \cup \{l_{new}\}$, where l_{P_t} is the set of labels in P_t.

The method works in two steps. First, we evaluate all possible labeling combinations for the detections in F_t. In practice, this can be achieved quickly by building a second graph between the detections in F_t and the potential labels, as illustrated in Figure 5. Algorithm 2 shows how to compute the log-likelihood terms involved in each link of this graph. Then, in a second step, we apply a standard ICM optimization step (SA algorithm with temperature 0) to all detections in $F_t \cup P_t$ to allow taking into account more recent detections in the labeling of past detections.

Algorithm 2 Algorithm to build new graph

```
for $r_i \in F_t$ do
    for $l \in pot(r_i)$ do
        $LL_i(l) = \sum_{r_j \in P_t} \log(\prod_k p(f_k(r_i, r_j)|H(l, l_j)))$
    end for
end for
```

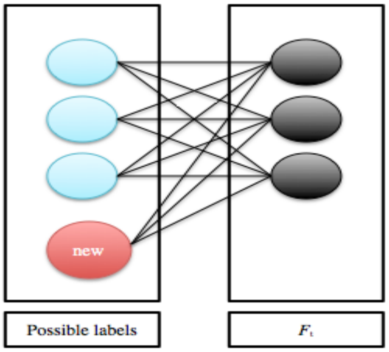

Figure 5. Graph containing the log-likelihood of each possible label for each detection in F_t

5. Experimental Results

5.1. Datasets

Surveillance data, even though recorded daily at a massive scale all around the world, are seldom made available for public research, especially with annotations. Their usability poses the ethical problem of privacy and personality rights. Therefore, a few freely available datasets like CAVIAR [1] have been extensively used for various computer vision tasks and present the advantage to provide ground truth for several interesting features. In this dataset, challenges arise from occlusions and also specular reflection. We used the shopping mall corridor view of this dataset, which comprises 26 videos. We also used videos from Torino metro stations. The method presented in [19] has been used to perform human detection on video sequences from CAVIAR and Torino.

We also tested our algorithms on the laboratory sequence from EPFL [2]. In this case, however, we used the output from a multi-view detector [5], which overall provides cleaner detection results.

5.2. Performance Measures

It is often difficult to compare results to other approaches because no common evaluation established benchmark has been adopted by the research community. Although attempts have been made to define evaluation measures for multiple object tracking [14] [7] there are no unique performance measures. In this paper, we used tracker purity TP and object purity OP as performance measures.

Introduced in [14], they can be interpreted as precision and recall measures. To obtain TP, we first identify for each estimated tracklet ϵ_i the ground truth track $GT_{\hat{j}_i}$ it spends the most time with, and measure its purity TP_i as the percentage of time ϵ_i spends with $GT_{\hat{j}_i}$. More precisely, given an estimate ϵ_i, we compute at each time instant t its overlap $F^t_{i,j}$ with each GT_j and compare it to a coverage threshold t_C. $GT_{\hat{j}_i}$ is then chosen as $\hat{j}_i = \arg\max_j \sum_t \mathbf{1}(F^t_{i,j} > t_C)$ where $\mathbf{1}$ is the indicator function. Denoting by n_i the total number of frames ϵ_i exists, we have:

$$TP_i = \frac{\sum_t \mathbf{1}(F^t_{i,\hat{j}_i} > t_C)}{n_i} \tag{9}$$

and the overall tracker purity is obtained by averaging over the number N_ϵ of estimates:

$$TP = \frac{1}{N_\epsilon} \sum_{i=1}^{N_\epsilon} TP_i \tag{10}$$

Reversely, object purity can be computed by looking for each GT_j to the track $\epsilon_{\hat{i}_j}$ it spends the most time with.

These measures give an insight into how much the estimates are associated to a single ground truth track. In the absence of identity switches, the tracker purity is 1. On the other hand, the object purity drops with the increase of mis-detections and if their largest associated estimates are short-lived.

5.3. Learning Procedure and Optimization

For CAVIAR, a two-fold approach was used, in which unsupervised training of model parameters was done on half of the videos and then used on the remaining videos. For Torino data, we performed training on 2 clips for a total of 2385 frames. In the case of the EPFL data, the unsupervised training was conducted on the given video. Since the ground-plane coordinates of detection outputs from [5] are quantized on a grid, we added noise to the ground-plane positions while performing training in order to take the uncertainty on the real position into account. We chose a short-term horizon $T_{short} = 10$ frames for all the datasets, and applied our Sliding Window algorithm for optimization.

5.4. Results and Discussion

Table 1 gives the average measures and their standard deviations over all the testing videos of CAVIAR. We observe a high tracker purity with a low variability across the tested videos. The variations of the object purity around the mean are quite large and depend mainly on the complexity of the sequences.

Table 1. Performance evaluation on CAVIAR

average tracker purity TP	0.97 (0.04)
average object purity OP	0.51 (0.13)

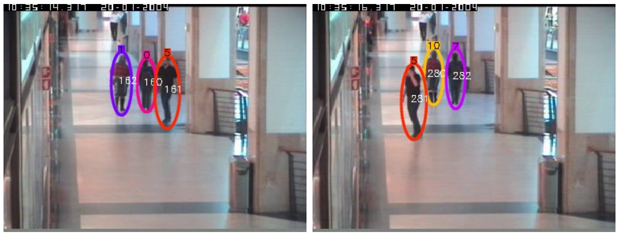

Figure 6. Example of long-term occlusion effect: before occlusion (left), after occlusion (right).

The results show that reliable tracklets with an average tracker purity close to 1 can be built on CAVIAR data using our approach. However, long occlusions (i.e. absence of detections) beyond T_{short} result in a significant object purity drop. Figure 6 shows how occlusions can affect the performance. The two persons in the back are being occluded by another one passing in the foreground. The person in the foreground correctly keeps his label, but as the occlusion is longer than T_{short} frames, the occluded people are given new labels when they reappear (i.e. they are detected again).

Figure 7 shows the output of our tracker on a more crowded sequence from CAVIAR. It shows that people with labels 1 and 9 are correctly tracked with a single label from the moment they are detected. The target labeled 2 is correctly tracked as well and then exits the scene to enter the shop on the left. For some other targets, the tracks are more fragmented, for example for label 4 which becomes 24 later.

Figure 8 shows the output of our tracker on the Torino data. As with CAVIAR, without heavy occlusions, the tracker is able to keep track of people correctly. One image also shows that a false alarm is treated correctly by the tracker which assigns a new label to it, that does not live long. This illustrates that false alarms are not a problem for the tracker, as we can observe from all used sequences. We also observe some missed detections on the top-right. Again, if some targets are missed by the detector on more than T_{short} successive frames, they will be assigned new labels when they reappear.

Training the model with a longer T_{short} has shown that it can resolve longer short-time occlusions during testing, but at the cost of a higher complexity of the graph. In order to get a higher object purity in monocular scenes, it will be necessary to deal with long-term occlusions. Long-term association can be considered by merging tracklets further apart in time, provided they present motion and appearance similarities. We are currently investigating this issue.

However, when given multiple cameras, we can use a multi-view detector that resolves the uncertainty due to oc-

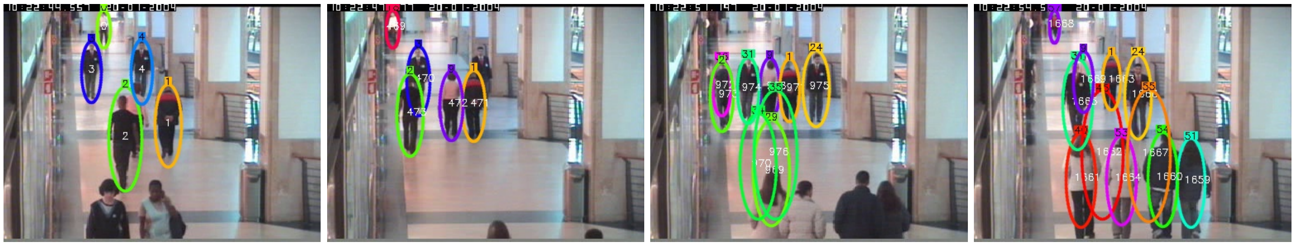

Figure 7. Example of tracker output on a more complex sequence from CAVIAR at $t = 509, 592, 675, 758$.

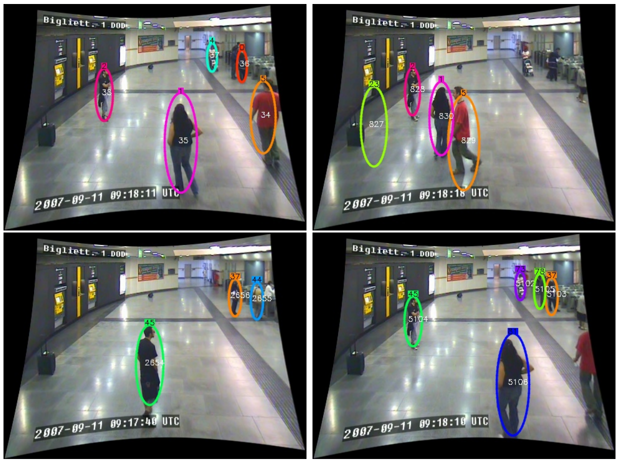

Figure 8. Examples of tracking output on typical surveillance data from a Torino metro station (each row is taken from a different sequence)

clusion by combining the views. In this context, we can expect to observe longer accurate tracks, which means a higher object purity. This is what we observe in the EPFL data, in which we apply our algorithm using the position feature alone. Figure 9 provides the output that we obtain in this case. The 6 targets are indeed correctly followed most of the time. For example, the persons with labels 6 and 30 are correctly tracked over the whole sequence, in spite of the numerous occlusions. Still, as the model is based on a short-term association, if some targets exit the field of view or are not detected for more than T_{short} frames, they are assigned a new label when they reenter the scene or are detected again. Because of exits of the scene and re-entries, 35 becomes 38, 43 becomes 54 and 58 becomes 64. Because of missed detection for more than T_{short} successive frames, 50 becomes 51 and 59 becomes 62. Some tracks are also fragmented for other reasons. For instance, 14 becomes 42 at one point where there is an abrupt change in the bounding box location. We also observe one instance of identity switch, as label 38 switches with label 51. However, tracking was performed using position alone. Other cues could hopefully help overcome this wrong association.

Computational cost. Given the detections, the time to extract the color features, build the graph and to perform the optimization takes around 200 milliseconds on the CAVIAR data.

6. Conclusion

We have formulated the multi-person tracking task as an association problem between detections. The association was expressed as a labeling process using a Conditional Random Field framework. The CRF encapsulates short-term dependencies between pairs of detections in the factor terms of the graph and are defined as probabilities of similarity measurements between detection pairs under two distinct hypotheses that they correspond to the same object (H_1) or not (H_0). Despite the use of simple features (location difference, Bhattacharyya color distance) and noisy detection, very good real-time performance has been achieved. Dynamic trajectory information could be incorporated in the framework, though it would require doing filtering and therefore increase the computational complexity. One limitation of our method is that it does not cope with long term occlusions. Current and future work will address the short-term limitation of our method, by merging tracklets using longer term motion and appearance models. Another improvement could consist in interpolating tracks to correct missed detections. We also plan to conduct more extensive evaluations.

Acknowledgments

This work has been funded by the Integrated Project VANAHEIM (248907) supported by the European Union under the 7^{th} framework program.

References

[1] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/.

[2] http://cvlab.epfl.ch/data/pom/.

[3] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *CVPR*, 2011.

[4] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *CVPR*, 2011.

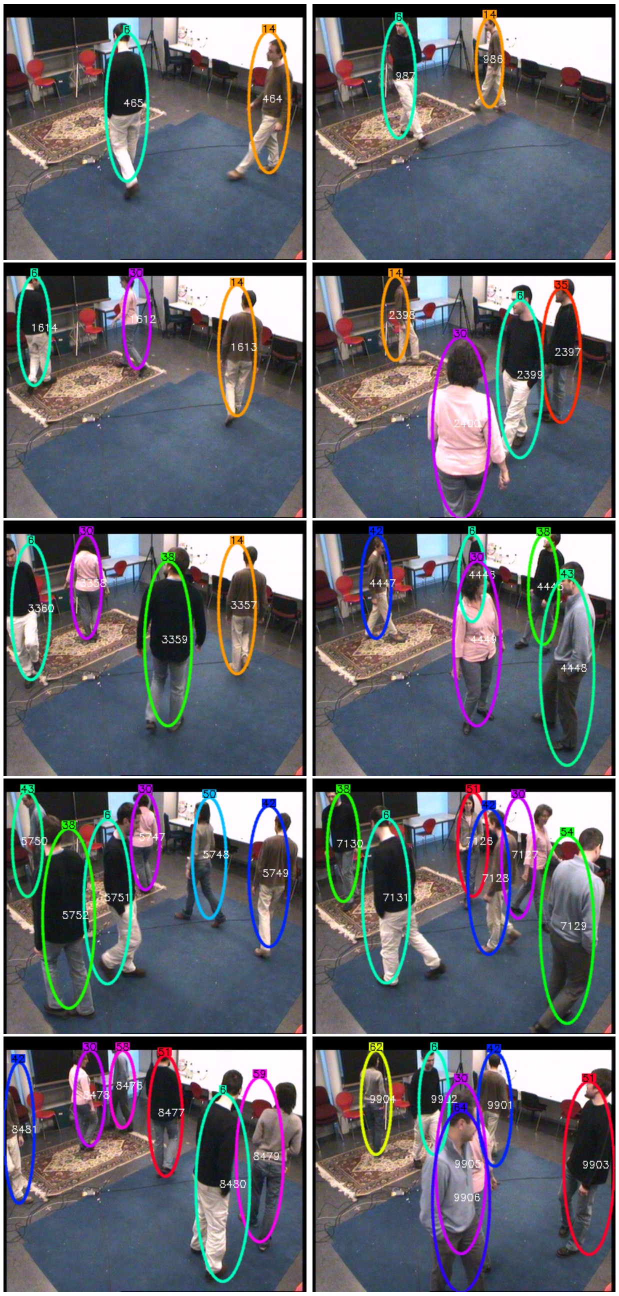

Figure 9. Examples of tracking output on detections from [5] (results displayed every 10 seconds).

[5] J. Berclaz, F. Fleuret, and P. Fua. Principled detection-by-classification from multiple views. In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, volume 2, pages 375–382, January 2008.

[6] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (Winter-PETS 2009)*, Snowbird, Utah, December 2009.

[7] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *J. Image Video Process.*, 2008:1:1–1:10, January 2008.

[8] C. Chen, A. Heili, and J. Odobez. Combined estimation of location and body pose in surveillance video. In *AVSS*, 2011.

[9] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[10] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.

[11] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proc. of ECCV*, 2008.

[12] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. Pattern Anal. Machine Intell.*, 27:1805–1819, 2005.

[13] G. Lathoud and J.-M. Odobez. Short-term spatio-temporal clustering applied to multiple moving speakers. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(5):1696–1710, July 2007.

[14] K. Smith, D. Gatica-Perez, S. Ba, and J. Odobez. Evaluating multi-object tracking. In *CVPR Workshop on Empirical Evaluation Methods in Computer Vision*, San Diego, June 2005.

[15] C. Sutton and A. Mccallum. An Introduction to Conditional Random Fields for Relational Learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[16] T. Yu, S. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *CVPR*, 2009.

[17] W. Ge, R. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *IEEE Workshop on Application of Computer Vision (WACV)*, 2009.

[18] X. Song, X. Shao, H. Zhao, J. Cui, R. Shibasaki, and H. Zha. An online approach: learning-semantic-scene-by-tracking and tracking-by-learning-semantic-scene. In *CVPR*, 2010.

[19] J. Yao and J.-M. Odobez. Fast human detection from videos using covariance features. In *8th European Conference on Computer Vision Visual Surveillance workshop (ECCV-VS)*, 2008.

[20] J. Yao and J.-M. Odobez. Multi-camera multi-person 3d space tracking with mcmc in surveillance scenarios. In *European Conference on Computer Vision, workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications (ECCV-M2SFA2)*, Marseille, Oct. 2008.

[21] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13+, Dec. 2006.

[22] Z. Wu, T. Kunz, and M. Betke. Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *CVPR*, 2011.

[23] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.