Long-Term Time-Sensitive Costs for CRF-Based Tracking by Detection

Nam Le, Alexander Heili, Jean-Marc Odobez

Idiap Research Institute, Martigny, Switzerland
École Polytechnique Fédéral de Lausanne, Switzerland

Abstract. We present a Conditional Random Field (CRF) approach to tracking-by-detection in which we model pairwise factors linking pairs of detections and their hidden labels, as well as higher order potentials defined in terms of label costs. Our method considers long-term connectivity between pairs of detections and models cue similarities as well as dissimilarities between them using time-interval sensitive models. In addition to position, color, and visual motion cues, we investigate in this paper the use of SURF cue as structure representations. We take advantage of the *MOTChallenge* 2016 to refine our tracking models, evaluate our system, and study the impact of different parameters of our tracking system on performance.

1 Introduction

Automated tracking of multiple people is a fundamental problem in video surveillance, social behavior analysis, or abnormality detection. Nonetheless, multi-person tracking remains a challenging task, especially in single camera settings, notably due to sensor noise, changing backgrounds, high crowding, occlusions, clutter and appearance similarity between individuals. Tracking-by-detection methods aim at automatically associating human detections across frames, such that each set of associated detections univocally belongs to one individual in the scene [1, 2]. Compared to background modeling-based approaches, tracking-by-detection is more robust to changing backgrounds and moving cameras.

In this paper, we present our tracking-by-detection approach [3] formulated as a labeling problem in a Conditional Random Field (CRF) framework, where we target the minimization of an energy function defined upon pairs of detections and labels. The specificities of our model is to rely on cue specific and reliability weighted long-term time-sensitive association costs between pairs of detections. This work was original proposed in [4, 3], and in this paper, we explored the use of additional cue (SURF) for similarity modeling, and the exploitation of training data to better filter detections or learn the cost models. In the following, we introduce the main modeling elements of the framework, then present the changes more specific to the *MOTChallenge* before presenting the results and analysis of our framework on the *MOTChallenge* data.

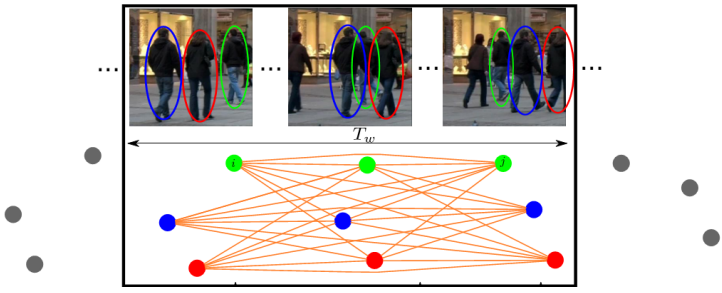

Fig. 1. Tracking as graph clustering task. The detections form the nodes, and a long-term connectivity is used, i.e. all links between pairs of nodes within a temporal window T_w are used to define the cost function. Long-term connectivity combined with time-interval sensitive discriminative pairwise models and visual motion enables dealing with missed detections, e.g. due to occlusion, as well as skipped frames.

2 CRF Tracking Framework

Our framework is illustrated in Figure 1. Multi-person tracking is formulated as a labelling problem within a Conditional Random Field (CRF) approach. Given the set of detections $Y = \{y_i\}_{i=1:N_y}$, where N_y is the total number of detections, we search for the set of corresponding labels $L = \{l_i\}_{i=1:N_y}$ such that detections belonging to the same identity are assigned the same label by optimizing the posterior probability $p(L|Y, \lambda)$, where λ denotes the set of model parameters. Alternatively, assuming pairwise factors, this is equivalent to minimizing the following energy potential [3]:

$$U(L) = \left(\sum_{(i,j)\in\mathcal{V}} \sum_{r=1}^{N_s} w_{ij}^r \, \beta_{ij}^r \, \delta(l_i - l_j) \right) + \Lambda(L), \tag{1}$$

with the Potts coefficients defined as

$$\beta_{ij}^r = \log \left[\frac{p(S_r(y_i, y_j)|H_0, \lambda^r_{\mathbf{\Delta}_{ij}})}{p(S_r(y_i, y_j)|H_1, \lambda^r_{\mathbf{\Delta}_{ij}})} \right], \tag{2}$$

and where $\Lambda(L)$ is a label cost preventing creation of termination or trajectories within the image (see [3] for details). The other terms are defined as follows.

First, the energy involves N_s feature functions $S_r(y_i, y_j)$ measuring the similarity between detection pairs as well as confidence weights w_{ij}^r for each detection pair, which mainly depends on overlaps between detection (see [3] for details). Importantly, note that a *long-term connectivity* is exploited, in which the set of valid pairs $\mathcal{V}$ contains *all pairs* whose temporal distance $\mathbf{\Delta}_{ij} = |t_j - t_i|$ is lower than T_w, where T_w is usually between 1 and 2 seconds. This contrasts with most frame-to-frame tracking or path optimization approaches.

Secondly, the Potts coefficients themselves are defined as the likelihood ratio of the probability of feature distances under two hypotheses: H_0 if $l_i \neq l_j$ (i.e.

Fig. 2. Position models. The different iso-contours of value 0 of the Potts costs for different values of $\mathbf{\Delta}$ (i.e. location of detections occurring after $\mathbf{\Delta}$ frames around each shown detection and for which $\beta = 0$), learned from sequences MOT16-01 and MOT16-03. In the region delimited by a curve, association will be favored, whereas outside it will be disfavored. Curves show different moving directions and amplitudes in each sequence.

detections do not belong to the same face), or H_1 when labels are the same. In practice, this allows to *incorporate discrimination*, by quantifying how much features are similar and dissimilar under the two hypotheses, and not only on how much they are similar for the same identity as done in traditional path optimization of many graph-based tracking methods. Furthermore, note that as these costs depend on the set of parameters $\lambda^r_{\mathbf{\Delta}_{ij}}$, they are *time-interval sensitive*, in that they depend on the time difference $\mathbf{\Delta}_{ij}$ between the detections. This allows a fine modelling of the problem and will be illustrated below.

Finally, in Eq. 1, $\delta(.)$ denotes the Kronecker function ($\delta(a) = 1$ if $a = 0$, $\delta(a) = 0$ otherwise). Therefore, coefficients β^r_{ij} are only counted when the labels are the same. They can thus be considered as costs for associating or not a detection pair in the same track. When $\beta^r_{ij} < 0$, the pair of detections should be associated so as to minimize the energy 1, whereas when $\beta^r_{ij} > 0$, it should not.

2.1 Features and association cost definition

Our approach relies on the unsupervised learning of time sensitive association costs for $N_s = 8$ different features. Below, we briefly motivate and introduce the chosen features and their corresponding distributions. We illustrate them by showing the Potts curves (for their learning see next section), emphasizing the effect of time-interval sensitivity and their easy adaptation to different datasets.

Position. The similarity is the Euclidean distance $S_1(y_i, y_j) = \mathbf{x}_i - \mathbf{x}_j$, with $\mathbf{x}_i$ the image location of the i^{th} detection y_i. The distributions of this feature are modelled as zero mean Gaussians whose covariance $\Sigma^H_{\mathbf{\Delta}}$ depends on the hypothesis (H_0 or H_1) and the time gap $\mathbf{\Delta}$ between two detections. Fig. 2 illustrates the learned models by plotting the zero iso-curves of the resulting β functions. We can notice the non-linearity with respect to increasing time gaps $\mathbf{\Delta}$ (especially for small $\mathbf{\Delta}$ increases), and the difference between sequences in viewpoints, moving directions, and amplitudes is captured by the models.

Motion cues. Motion similarity between detection pairs is assessed by comparing their relative displacement and their visual motion. The similarity is

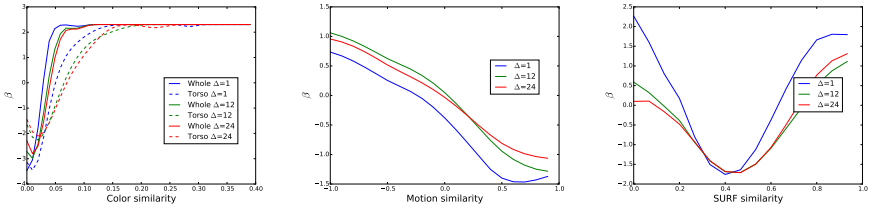

Fig. 3. Automatically learned Potts functions β for the different similarity functions and some values of $\mathbf{\Delta}$ in sequence MOT16-01. Left: color distances. Middle: motion cue distance. Right: SURF cue distance.

computed as the cosine of the angle between these two vectors. Intuitively, if a person moves in a given direction, the displacement between its detections and their visual motion will be aligned, leading to a motion similarity close to 1. The resulting β curves in the middle plot of Fig. 3 confirm the above intuition, as the β decreases at the cosine value increases.

Appearance (color). Detections are represented by multi-level color histograms in 4 different regions: the whole body and its subparts, the head, torso, and leg regions. The similarity between histograms of the same region of the detections is measured using the Bhattacharyya distance D_h, and the distributions of this distance is modelled using a non-parametric method. Example of Potts curve β are shown in Fig. 3, Left. We can notice here that the statistics associated to each region are relatively different, and although we would not expect so, also varies with the time gap $\mathbf{\Delta}$ between detections.

Appearance (SURF). Color is sometimes not sufficient to discriminate between people. We thus propose to exploit more structured appearance measures. More precisely, we rely on SURF [5] descriptors computed at interest points detected within the detection bounding box, although better re-identification oriented descriptors could be used. They are invariant to scale, rotation, and illumination changes and are thus suitable for person representation under different lighting conditions or viewpoint changes. As similarity measure, we use the average Euclidean distances between pairs of nearest keypoint descriptors from the two detections. We model the distributions of the similarity measures with a non-parametric approach. As can be seen in the right plot of Fig. 3, the Potts coefficient β is negative for a SURF similarity around 0.4, thus encouraging association for such values. On the other hand, positive coefficients for larger distances - around 0.7 - discourage the association. The β values are surprisingly positive for smaller values, but this can be explained by the fact that small values are very rarely observed, and due to some smoothing applied to probability estimates, β values are either saturated or close to neutral when the distance is small (see [3]).

2.2 MOT-Challenge - Parameter learning, optimization

Here we comment on changes and modifications made for the MOT16 benchmark (in addition to evaluating the benefit of SURF features). They relate to detection preprocessing, parameter learning, and optimisation.

Detection filtering. The quality of the detections have a direct impact on the performance of the system. In our work, we rely on the Deformable Part-based Model (DPM) detector [6] [1]. In [3], a simple scheme based on size was used to eliminate obvious false detections when calibration was available. Here, we take advantage of the training data to learn simple rules and parameters to increase the precision of the detector according to the following factors.

- Detection size: Because in *MOTChallenge* 2016, training and test sequences are paired with roughly the same viewpoint, the groundtruth (GT) bounding boxes from the training video can be used to filter detections in test sequences. Assuming that the height of one detection linearly relates to its horizontal coordinate, one can estimate the most likely range of height for one detection. Detections that fall out of the range are omitted to remove obvious false alarms and big detections that cover multiple people.
  Concretely, let $[x, y, h]$ be the coordinates and height of on GT bounding boxes. At training time, for each x, one can find h_{min}, h_{max} to be the minimum and maximum height of all boxes with the same horizontal coordinate. The relationship between h_{max}, h_{min} and x and be estimated through linear regression: $h_{min} = a_m \times x + b_m$ and $h_{max} = a_M \times x + b_M$.
  At test time, for one detection $[x_{test}, y_{test}, h_{test}]$, one can find a predictive range $[\bar{h}_{min}, \bar{h}_{max}]$ to accept detections that fall within that range. This constraint helps removing obvious big false alarms or detections covering multiple people. From table Tab. 1 the filter gives a boost in precision with a small decrease in recall and all tracking metrics are improved.
- Detection score: we can vary the threshold T_{dpm} of the DPM detector to find an appropriate threshold that provides a good compromise between recall and precision.

Parameter learning. Given our non-parametric and time interval sensitive cost model, the number of parameters to learn in λ is quite large. In [3], a two step *unsupervised approach* was used to train the model directly from data. Broadly speaking, a first version of the model is learned for small time interval assuming that closest detections of a given detection in the next frames correspond to the same person. These modes were used to run the tracker a first time. The resulting tracks (usually with high purity) were then used to lean the full model.

In the context of the MOT challenge, we took advantage of the availability of training data to learn the models from the ground truth (GT), and applied these models to the test data. We also considered relearning the parameters from the

[1] Although the detector is the same that produced the public detections, we used our own output to exploit the detected parts for motion estimation.

obtained tracking results before reapplying the model to evaluate the impact of taking into account the noise inherently present in the data.

Optimization. We mainly followed the approach of [3]. For computational efficiency, we used a sliding window algorithm that labels the detections in the current frame as the continuation of a previous track or the creation of a new one, using an optimal Hungarian association algorithm relying on all the pairwise links to the already labelled detections in the past T_w instants. A second step (Block ICM) is then conducted, which accounts for the cost labels and allows the swaping of track fragments at each time instant.

3 Experiments

In [3], the original model was evaluated on the CAVIAR, TUD sequences, PETS-S2L1, TownCenter, and ParkingLot sequences and was providing top results. The new MOT16 benchmark contains 14 sequences with more crowded scenarios, more scene obstacles, different viewpoints and camera motions and weather conditions, making it quite challenging for the method which did not incorporate specific treatments to handle some of these elements (camera motion, scene occluders). The MOT16 challenge thus allows to better evaluate the model under these circumstances.

3.1 Parameter setting

For each test sequence, there is a training sequence in similar conditions. As explained earlier, we have used the training sequences to learn Potts models, and used them on the test data. Other parameters (e.g. for reliability factors) were set according to [3] and early results on the training data. Unless stated otherwise, the default parameters (used as well on test data) are: $T_w = 24$, $\mathbf{\Delta}_{sk} = 3$ (i.e. only frame 1, 4, 7, ... are processed), $d_{\min} = 12$ (short tracks with length below $d_{\min}$ were removed), $T_{dpm} = -0.4$, and linear interpolation between detections were produced to report results.

3.2 Tracking evaluation

We use the metrics (and evaluation tool) of the MOT challenge. Please refer to [9] for details. In general, except the detection filtering, results (MOTA) were not affected much by parameters changes.

Detection filtering. Tab. 1 reports the metrics at detection level and tracking level when applying the linear height filtering and with different detection threshold T_{dpm}. The filter gives a boost in precision with a small decrease in recall and all tracking metrics are improved thanks to fewer false alarms. We can also observe that threshold $T_{dpm} = -0.4$ provides an appropriate trade-off between precision and recall and good tracking performance.

Tracking window T_w and step size $\mathbf{\Delta}_{sk}$. Different configurations are reported in Tab. 2. One can observe that with longer tracking context T_w (default $T_w = 24$

	Raw detection	Filtered detection		
	$T_{dpm} = -0.5$	$T_{dpm} = -0.5$	$T_{dpm} = -0.4$	$T_{dpm} = -0.3$
Detection Recall	35.4	35.1	34	32.4
Detection Precision	78.3	86.1	89.9	92.4
MOTA	25.2	29.1	29.8	29.3
MOTP	74.1	74.2	74.3	74.6

Table 1. Detection filtering. Detection precision-recall and tracking performance (nota: tracks are not interpolated in results).

Parameters	Rec.	Pre.	FAR	MT	PT	ML	IDS	FM	MOTA	MOTP
Default	38.7	85.9	1.32	49	180	288	211	634	32.1	74.7
T_w =12	35.9	90.5	0.78	39	181	297	275	636	31.9	75.1
$\mathbf{\Delta}_{sk}$ =1	40.5	82.8	1.74	52	188	277	273	1199	31.8	73.7
$\mathbf{\Delta}_{sk}$ =3	38.7	85.9	1.32	49	180	288	211	634	32.1	74.7
$\mathbf{\Delta}_{sk}$ =6	35.5	88.9	0.92	33	177	307	217	459	30.8	75.1
Unsup. models	38.0	86.6	1.22	43	183	291	237	692	31.9	74.7
W/o match. sim.	36.6	89.5	0.89	48	157	312	210	555	32.2	75
With match. sim.	37.2	88.8	0.98	49	161	307	203	638	32.3	74.8

Table 2. Evaluation of our tracking framework with various configurations. Results with the default parameters are shown first, and then performance obtained when varying one of the parameters (provided in first column) are provided.

vs shorter $T_w = 12$), tracks are more likely to recover from temporary occlusions or missed detections, resulting in higher MT, ML. When detector is applied scarcely (e.g. $\mathbf{\Delta}_{sk} = 3$ or 6), we observe a performance decrease (e.g. decrease of MT, increase of ML). Nevertheless, applying the detection every $\mathbf{\Delta}_{sk} = 3$ frames reduces the false alarms and improves IDS and FM metrics. Since detection is one of the computation bottlenecks, this provides a good trade-off between performance and speed. When $\mathbf{\Delta}_{sk} = 3$, the overall tracking speed also is increased by up to 6 times.

Supervised vs unsupervised models. The "Unsup. model's' line in Tab. 2 provides the results when using association models trained from the raw detection *in an unsupervised fashion as in [3]*, which can be compared against of the default ones obtained using tracking models trained from the labeled GT boxes provided in *MOTChallenge* 2016. Interestingly, although the *unsupervised* approach suffer from missing detections and unstable bounding boxes, it performs very close to the supervised models in most tracking metrics.

Matching similarity. Because of the complexity, we used $T_w = 15$ for sequence MOT16-04, the rest use the default parameters. Although SURF matching can be discriminative for objects, it is less effective in human tracking because of clothing similarity, and data resolution where most features are found on human boundaries rather than within. This is reflected in Tab. 2, where only minor

	FAR	MT	ML	IDS	FM	MOTA	MOTP
LTTSC_CRF	2.0	9.6 %	55.2 %	481	1012	37.6	75.9

Table 3. Results on the MOT 2016 test data.

improvement in IDS, ML, MT, and PT are observed. In future work, better tracking oriented cues could be used, such as those developed for re-identification.

3.3 Evaluation on test sequences

The results of the method configured with detection filtering and the default parameters for the tracker are reported in Tab. 3. Overall, the performance are better, showing that the method generalizes well (with its limitation) and qualitative results are aligned with those of the training sequences. The comparison with other trackers can be found in the MOT website[2]. Overall, our tracker achieved fair ranking in comparison to other methods. Considering methods based on the public detections, our tracker exhibit a good precision (rank $5^{th}/20$ on the IDS metric and $8^{th}/20$ on Frag metric) but is penalized by a low recall, resulting on a ranking of $11^{th}/20$ for MOTA. It is important to note that our modeling framework was taken as is from previous paper, and not adapted or over-tuned to the MOT challenge (e.g. for camera motion or other things). In addition, as our framework can leverage any cue in a time-sensitive fashion, other state-of-the-art features like those based on supervised re-identification learning can be exploited and would positively impact performance.

4 Conclusion

We presented a CRF model for detection-based multi-person tracking. Contrarily to other methods, it exploits longer-term connectivities between pairs of detections. Moreover, it relies on pairwise similarity and dissimilarity factors defined at the detection level, based on position, color and also visual motion cues, along with a feature-specific factor weighting scheme that accounts for feature reliability. Experiments on MOTChallenge 2016 validated the different modeling steps, such as the use of a long time horizon T_w with a higher density of connections that better constrains the models and provides more pairwise comparisons to assess the labeling, or an unsupervised learning scheme of time-interval sensitive model parameters. The results also give us hint at future directions such as occlusion and perspective reasoning, handling the high-level of miss-detections, or adapting our framework better to moving platform scenario.

Acknowledgement This research was supported by the European Union project EUMSSI (FP7-611057).

[2] https://motchallenge.net/results/MOT16/

References

1. Berclaz, J., Fleuret, F., Fua, P.: Multiple object tracking using flow linear programming. In: Winter-PETS. (2009) 1–8
2. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. In: CVPR. (2012) 2034–2041
3. Heili, A., Lopez-Mendez, A., Odobez, J.M.: Exploiting Long-Term Connectivity and Visual Motion in CRF-based Multi-Person Tracking. IEEE Trans. on Image Processing (2014)
4. Heili, A., Chen, C., Odobez, J.M.: Detection-Based Multi-Human Tracking Using a CRF Model. In: IEEE ICCV-VS, Int. Workshop on Visual Surveillance, Barcelona. (2011)
5. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: ECCV. Springer (2006) 404–417
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9) (2010) 1627–1645
7. Dubout, C., Fleuret, F.: Exact acceleration of linear object detectors. In: ECCV. Springer (2012) 301–311
8. Dubout, C., Fleuret, F.: Deformable part models with individual part scaling. In: BMVC. (2013)
9. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)