Multi-Target Tracking with Sparse Group Features and Position using Discrete-Continuous Optimization

Billy Peralta (1) and Alvaro Soto (2)

(1)Universidad Católica de Temuco, (2)Pontificia Universidad Católica de Chile
bperalta@uct.cl asoto@ing.puc.cl

Abstract. Multi-target tracking of pedestrians is a challenging task due to uncertainty about targets, caused mainly by similarity between pedestrians, occlusion over a relatively long time and a cluttered background. A usual scheme for tackling multi-target tracking is to divide it into two sub-problems: data association and trajectory estimation. A reasonable approach is based on joint optimization of a discrete model for data association and a continuous model for trajectory estimation in a Markov Random Field framework. Nonetheless, usual solutions of the data association problem are based only on location information, while the visual information in the images is ignored. Visual features can be useful for associating detections with true targets more reliably, because the targets usually have discriminative features. In this work, we propose a combination of position and visual feature information in a discrete data association model. Moreover, we propose the use of group Lasso regularization in order to improve the identification of particular pedestrians, given that the discriminative regions are associated with particular visual blocks in the image. We find promising results for our approach in terms of precision and robustness when compared with a state-of-the-art method in standard datasets for multi-target pedestrian tracking.

1 Introduction

Automatic multi-target tracking is the computational task of detecting the trajectories of objects in a sequence of images. It has many and diverse applications in the real world, e.g. surveillance [1, 2], sports [3] and sensor networks [4]. In this work, we focus on multi-target tracking of pedestrians, where recent research has shown significant progress. Nonetheless, current techniques only offer good performance with easy conditions i.e. a static background and separated pedestrians. As the area of inspection becomes more crowded, the performance of these algorithms tends to decrease and they are clearly outperformed by humans.

Successful tracking methods are based on the premise of the presence of previous detections i.e. target people are detected by a generic pedestrian detector. In fact, this scheme can be combined with an online model to deal with appearance variation or scene lighting [6]. Some advantages of using pedestrian detections

(a) Results with model based on position [5]

(b) Results with our model based on sparse grouped features

Fig. 1. A comparison of our method with a state-of-the-art model [5]. In 1, the model based on location information tends to confuse as it only considers position, as shown by the yellow and light-blue tracklets. In contrast, our model 1(b) uses visual feature information with sparse selection of groups of features to select the more important visual components of the tracklets. In our case, all the tracklets are correctly identified.

as input as compared to a generic tracker are that this does not require initialization of the trackers and avoids model drifting. In single target tracking the task is usually accomplished by fitting a trajectory prediction function according to evidence given by detections; in multi-target tracking by contrast, we have multiple possible identities that complicate trajectory prediction by trackers. We have two specific problems: (i) the assignation of a unique identity to each detection, which is also called the data association problem, and (ii) trajectory estimation of each target.

Great emphasis is currently placed on the data association problem, since many discrete optimization techniques has been developed over several decades, even though this problem is NP-hard [7]. In the case of trajectory estimation, it is usually computed assuming known correct labelings. In order to solve the multi-target tracking problem, a key observation of Andriyenko et al. [5] is given by the complementarity between data association, which is a discrete optimization problem as the assignation of identities is nominal; and trajectory estimation, which is a continuous optimization problem as the position variable is numeric. In this case, the two problems are solved jointly using a discrete-continuous alternating optimization under a Markov Random Field (MRF) framework. Although this method is natural for this problem, it does not consider visual feature information and may potentially lose valuable information about the targets.

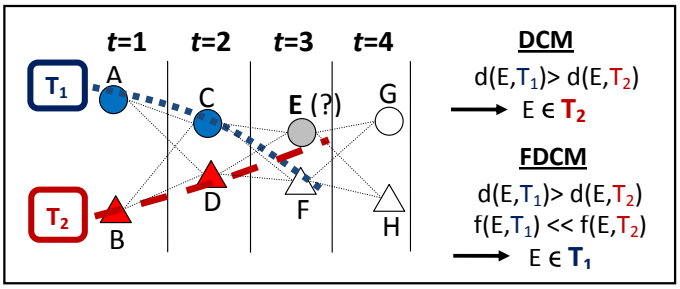


Fig. 2. The difference between our model (FDCM) and a state-of-the-art model (DCM) [5]. We have two objects that are detected in four moments ($t = 1$ to 4) originating 8 nodes in MRF (node A to node H). We assume two known trackers T_1 and T_2 and that the geometric and feature distances (d and f) are defined. With DCM, the node E has label T_2 because it is geometrically near to trajectory of tracker T_2: $d(E, T_1) > d(E, T_2)$. With FDCM, the node E has label T_1 as it considers feature distance between nodes and trackers: $d(E, T_1) + f(E, T_1) < d(E, T_2) + f(E, T_2)$. The feature distance between node E and T_1 is small because T_1 is formed by circles and node E is a circle; on contrast, T_2 is composed by triangles. In similar way, node F belongs to trackers T_2.

In multi-target tracking of pedestrians, we observe that if the targets are well separated, position information is usually enough to assign a correct label to each detection. In a more challenging setting, as in the case of crowded scenes, visual feature information can be very valuable for distinguishing between multiple individuals by using particular features such as the type or color of their clothes. In the present work, we embed a feature-dependent function inside the association potential component of a Markov Random Field. We use a multi-class logistic regression model to represent the relation between features of detections and the identity of trackers. Furthermore, we expect that in a frame sequence, a person will usually have a particular discriminative region of image; therefore, we also propose the use of group Lasso regularization. Figure 2 shows the intuition of our idea, if we use geometric distance node E has label T_2, but if we consider feature and geometric distance, node E has label T_1. Finally, the estimation of parameters of this function assumes a fixed labeling and adopts a Markov Random Field solution based on a pseudo-likelihood approximation.

2 Background

2.1 Related work

Multi-target tracking research has made significant progress in recent years. Kuo et al. [8] presented a discriminative appearance model for multiple targets using an online learning scheme based on the AdaBoost algorithm. Training samples are selected considering a sliding time window and spatial-temporal constraints. Although this model appears effective, it does not consider position information

which can be useful in case of occlusion or poor lighting conditions. Berclaz et al. [9] proposed a mathematical model based on constrained linear programming to address data association. This model is based on discretization given by an occupancy map where each cell can be filled with a target. A disadvantage of this method is that it can involve a very large number of variables i.e. if we assume a square occupancy map, the number of nodes is related quadratically to the length of the side of the square. Benfold and Reid [1] described a multi-target tracking system considering accurate estimations of head locations. They solve data association using a MCMC scheme combined with KLT tracking and HOG pedestrian detections in a multithreaded application. This system appears efficient because of its real-time performance. Nonetheless, a potential problem with this model is its complexity due to the combination of multiple algorithms.

Yang and Nevatia [10] proposed an online model based on Conditional Random Fields (CRF) to find and discriminate multiple targets. Although this model is efficient, the non-submodularity of energy function hinders the CRF inference provoking the use of heuristic solutions, and therefore, there is no guarantee that they will find an optimal solution. Butt and Collins [11] address the problem of multiple tracking by using detections in triplets, where a triplet is defined as three consecutive detections of a pedestrian. The use of triplets gives natural access to motion information i.e. with this information they estimate the speed of a particular candidate target in the triplet. Nevertheless, this model is supported by the reliability of the triplets, which can be uncertain as they may overlap other triplets. They use an heuristic process in order to avoid conflicts between nearby triplets, using the nearest detections. Hoffman et al. [12] proposed a hierarchical model for multiple tracking. They present a probabilistic model that finds the paths for each level of hierarchy. This path-finding problem is usually solved by the Hungarian algorithm. Nevertheless, due to the complexity of the hierarchical model, they require some heuristics for post-processing the results, where the parameters can be difficult to obtain.

An interesting work is presented by Andriyenko et al. [5], in which data association and trajectory estimation problems are addressed respectively by a joint discrete and continuous optimization, on a Markov Random Field where the nodes set is given by the pedestrian detections and the edge set is given by the pairs of nearby detections. The discrete optimization assumes that the trajectory and label costs are known; then they estimates the unary and pair-wise terms of MRF; they subsequently apply a graph cut algorithm based on α-expansion as the energy function remains sub-modular. In this MRF, the unary term is proportional to the distance between detections and the spline associated with a specific label. In the case of the pair-wise term of two nearby nodes, it is greater than zero if the labels of the two nodes are different; otherwise it is zero. The continuous optimization of trajectories is hard given the presence of label cost. They assumes the labels as known, disregard the label cost, and fit a cubic spline model over the targets positions considering one label at a time in order to estimate the trajectories, where a change is accepted only if the global energy function is decreased. After the trajectory fitting, they calculate

the label cost. This discrete-continuous optimization process is done alternately until convergence of the MRF energy is reached.

The work of Andriyenko et al. [5] has the disadvantage that does not consider the visual feature information available, which may be helpful in differentiating between different trackings in complex environments. It is unlikely that two people will have the same appearance and clothing, and in this case an algorithm could use their local characteristics to facilitate the tracking process. Our proposal is based on a non-trivial augmentation of the association potential function by incorporating a term that indicates the consistency between the features of candidate detections and the average features of a candidate target. We will now describe the model of [5].

2.2 Discrete-Continuous model

Following the notation of [5], we assume an input set of M detections $D = (d_1, ..., d_M)$ and a set of labels represented by variable f with nominal values $f = (1, ..., N)$; this labeling variable assigns each detection to one of N trajectories $T = (T_1, ..., T_N)$ and identifies a false alarm using the outlier label $\oslash$. Each index detection d is associated solely with one ordered pair (i, d), where i is the detection index in a frame in relative time index t. Using this convention, each detection d is associated with a position p_i^t in the image. Using a MRF framework, the model graph is identified by Q where the nodes D are given by the set of detections inside a MRF framework and the pair of nearby nodes represents the edge set V. Specifically the edge set is defined by a temporal restriction between pairs of detections: $V(d_t^j, d_{t+1}^k) = 1$, if only $\left\| p_t^j - p_{t+1}^k \right\|^2 < \tau$. The energy of this MRF model is defined by the following equation:

$$E_d^T(f) = \sum_{d \in D} U_d(f_d, T) + \sum_{d,d' \in V} S_{d,d'}(f) + \sum_{i=1}^{N} \hat{E}_v^{te}(T_i) + \kappa h_f(T) + lnZ \quad (1)$$

In this case, the first term U_d represents the association potential function, the second term $S_{d,d'}$ represents the interaction potential function; and together they make up the discrete data association. The third term $\hat{E}_v^{te}(T_i)$ represents the continuous trajectory model. And finally, the fourth term $h_f(T)$ represents the label cost term, where this term penalizes complex configurations of MRF solutions. The term Z represents the partition function of the energy; in [5], the partition function is not stated as they do not need to estimate any parameter, however, in our model we need operate with it as we require to learn the weight of the features.

The association potential function component U is defined as:

$$U_{d_j^t}(l, T) = c_j^t \left\| p_j^t - T_l(t) \right\|^2 \quad (2)$$

If the detection is labeled as an outlier, it is penalized with a constant outlier cost $\oslash$, which is modulated by c_j^t. The interaction potential function component S is given by the following Potts model:

$$S_{d_t^j, d_{t+1}^k}(f) = S_{d_t^j, d_{t+1}^k}(f_{d_t^j}, f_{d_{t+1}^k}) = \eta\delta(f_{d_t^j} - f_{d_{t+1}^k}) \tag{3}$$

The continuous trajectory model evaluates the smoothing degree of trajectories. In this case, this model uses a cubic B-spline to fit the trajectories of targets. In our model, we do not alter this component as it is unrelated with visual feature information.

The label cost function penalizes excessive complexity of trajectories. This function is the sum of five terms: The dynamic cost h_i^{dyn}, which penalizes complex splines by adding the cubic coefficients of splines, $C_i(r, 3)$, i.e. it prefers simpler curves. The persistence cost h_i^{per}, which penalizes unreliable trajectories by adding the distance to border of image ($\bar{b}$) and the inverse size of the same trajectory. The high-order data fidelity cost h_i^{fid}, which punishes trajectories that are far away from detections over a long period by adding the square of the cardinal of the subsequences of such trajectories G_k; we use a quadratic penalty instead of cubic as in [5] as we obtained better experimental results. The mutual exclusion cost h_i^{col}, which penalizes collisions between trajectories by considering physical constraints i.e. two objects cannot be in the same position. Finally, the regularization cost, which penalizes the cardinality of the set of trajectories. For more details, please see [5].

3 Our Approach

In our work, we augment the association potential term by including visual feature information with an embedded feature selection based on image regions. We consider an input set of M detections $D = (d_1, ..., d_M)$ with a corresponding set of features $X = (x_1, ..., x_M)$ where each component $x_i \in \Re^K$ is given by $x_i = (x_i^1, ..., x_i^K)$. As in the case of the detection variable, each feature vector x_d is associated with a unique ordered pair (t, j). By considering N possible trackers, we add an auxiliary class variable depending on a particular label variable $l \in (1, ..., N)$, $Y = (y_1^l, ..., y_M^l)$ where each component $y_i^l \in \{-1, +1\}$ depends on the assumed detection label f_i and is given by the indicatrix formula $y_i^l = 2 * I(l = f_i) - 1$. Moreover, we use a logistic regression model for connecting class and feature variables, Y and X, given by $Y = \sigma(W^T X)$, with $W \in \Re^{KxN}$ and σ as sigmoid function. Our task is to find the best set of weights W that explains the relation between the set of features X and set of class variables Y in order to maximize the likelihood of the MRF.

3.1 MRF model

The association potential function is enhanced with the addition of a term related to the classification of detections according to tracker label information. As the

association potential function U depends on label l, the parameter of the logistic regressor is a vector $W_l \in \Re^K$. We regularize this vector considering sparsity and natural groups with group Lasso penalization [13, 14]. The number of non-overlapping groups of visual features is G. By using the bijective correspondence between an index of detections d and an ordered pair (t, j), the association potential function is reformulated by considering the regularization terms in $R(w_l)$ as:

$$U_{d_t^j}(l, T, x_t^j) = c_j^t(\left\| p_t^j - T_l(t) \right\|^2 + \alpha \, log(1 + exp(-y_l^d(w_l^T x_t^j)))) + R(w_l) \quad (4)$$

$$R(w_l) = \lambda_S \left\| w_l \right\|^1 + \lambda_G \sum_{g=1}^{G} \left\| w_l^g \right\|^2 \quad (5)$$

MRF training is hard to solve as it needs to calculate the value of the partition function, which demands an exponential number of configurations. As we require to find an optimal set of weights W, we cannot avoid this problem as in [5]. To do this, we use the pseudolikelihood approximation[15], where the energy function is approximated using a local partition function instead of the global partition function:

$$E_{PL:d}^T(f) = \sum_{d \in D} U_d(f_d, T) + \sum_{d, d' \in V} S_{d,d'}(f) + \sum_{i=1}^{N} \hat{E}_v^{te}(T_i) + \kappa h_f(T) + ln Z_{loc} \quad (6)$$

The majority of terms are defined in Equation 1 (Subsection 2.2). The local partition function (Z_{loc}) is given for each node and represents the sum of the energies under all the possible labels [15] which has the advantage of being much more maneuverable than the global partition function. We also apply the same strategy as [5] and solve the model by alternating between optimization of the trajectory set, T, and the labelings set, f. By fixing T, we can ignore this term and have the pseudolikelihood energy expressed as a function of w given by:

$$E_{PL:d}^T(w_l) = \sum_{d \in D} \left\{ \left[c_j^t(\left\| p_t^j - T_l(t) \right\|^2 + \alpha \, log(1 + exp(-y_l^d(w_l^T x_t^j))) + \lambda_S \left\| w_l \right\|^1 \right. \right.$$
$$\left. \left. + \lambda_G \sum_{g=1}^{G} \left\| w_l^g \right\|^2) + \sum_{d' \in V_d} S_{d,d'}(f_d, f_{d'}) \right] - ln(Z_{loc}^d) \right\} \quad (7)$$

where the local function partition $Z_{loc}^d(w_l)$, considering $Q_t^j = \left\| p_t^j - T_m(t) \right\|^2$, is given by:

$$Z_{loc}^d(w_l) = \sum_m exp \left\{ c_j^t(Q_t^j + \alpha \, log(1 + exp(-y_m^d(w_l^T x_t^j))) + \sum_{d' \in V_d} S_{d,d'}(f, f_{d'}) \right\} \quad (8)$$

By fixing the terms independent of w_l with an auxiliary variable K_d and the regularization terms as $R(w_l)$, the optimization equation is re-expressed as a function of w_l^* by:

$$w_l^* = \arg\min_{w_l} \sum_{d \in D} \left\{ K_d + c_j^t \alpha \, log(1 + exp(-y_l^d (w_l^T x_t^j))) - ln(Z_{loc}^d) \right\} + R(w_l) \quad (9)$$

3.2 Estimation of MRF parameters

Equation 9 is hard to solve due to the presence of the local partition function and the regularization term. To solve Equation 9, we heuristically follow a strategy similar to Lee et al. [16], who first solve an unregularized version of the cost function and then add the sparsity constraints. In our case, we use the additivity property of gradient and follow a three-step strategy: **i)**. First, we solve an unregularized version of the energy function given by Equation 9 where the component referring to local partition function can be separated. **ii)**. As we have a logistic regression term that approximates to the unregularized version of the energy function, we add the Lasso and Group-Lasso regularization terms in order to have a standard logistic regression model with Group-Lasso regularization [13, 14]. In this step, we estimate the gradient of this model by using the direction of the solution given by the code based on the SLEP package [17]. **iii)**. Finally, we approximate the total gradient of Equation 9 by summing the results of the gradients given in steps **i)** and **ii)**, and substracting the gradient of the version of Equation 9 without the local partition function and the regularization term, which is simple to solve.

In the first step, we solve an unregularized versions of energy ($E_{PL:d}^{T(u)}$) using an optimization based on Newton method. The gradient is given by:

$$\nabla_{w_l} E_{PL:d}^{T(u)}(w_l) = -\alpha \sum_{d \in D} c_j^t \left[p_d(\bar{y}_l/d; w_l) y_l^d x_t^j - \left\langle p_d(\bar{y}_m/d; w_l) y_m^d x_t^j \right\rangle_{p_m(y_m/d; w_l)} \right] \quad (10)$$

In the previous equation, we use the notation $p_d(\bar{y}_m/d; w_l) = 1 - p_d(y_m/d; w_l)$. We accelerate the calculations by using the Hessian of unregularized energy function that is given by:

$$H_{w_l}(E_{PL:d}^{T(u)}(w_l)) = -\alpha \sum_{d \in D} \left\{ c_j^t x_t^j x_t^{j^T} \; [p_d(y_l/d; w_l)(1 - p_d(y_l/d; w_l)) \right.$$
$$\left. - \langle (1 - p_d(y_l/d; w_l)) p_d(y_m/d; w_l) \rangle_{p_m(y_m/d; w_l)} \right] \Big\} \quad (11)$$

The equation requires a known labeling. We use an iterative scheme based on hard partitioning according to labels [15]: first we fix the labeling and calculate the weight. Once the weights are fixed, we calculate the labeling using the same scheme as Andriyenko using the max-flow solution. In this case, after labeling we

also optimize the continuous trajectory model and the label cost. We continue this process until we reach a convergence criterion. We also use the random change heuristic in the configuration of splines and accept them if only they decrease the energy function [5].

4 Experiments

4.1 Datasets

Our method is evaluated on four public datasets of pedestrian video sequences and two own datasets. The first three datasets come from TUD dataset, they are: Campus(TUD-CAMP), Crossing (TUD-CROS) and Stadtmitte (TUD-STAD) subsets [18]. These videos have 91, 201 and 179 frames respectively, and show pedestrians in street scenes. The low viewpoint means that the targets often occlude one another. The fourth dataset is PEDS-2009-S2L1 which has 795 frames and considers a high viewpoint. The detections are performed using a SVM classifier based on Histogram of Gradients [19]. The fifth and sixth datasets (U-HALL-1 and U-HALL-2) were obtained from an indoor environment in a university campus; the first is a easy case of multi-tracking where the people are usually separated. In order to have a variety of experimental settings, we simulate a perfect pedestrian classifier using human labeling. The feature-based methods of bibliographic revision have not available their code, for such reason, we only compare with [5]. In this work, our main goal is to improve the pure MRF framework for multi-tracking tasks considering feature information. Table 1 shows the details of each dataset.

4.2 Implementation details

We compare our feature-based discrete-continuous method (FDCM) with the discrete-continuous model (DCM) of [5], which is a state-of-art method of multi-target tracking. For FCDM, we use two combined features based on texture and color information: (i) dense HOG over a 6x3 grid with 576 resulting features and (ii) 3-level pyramidal RGB features producing 315 features. As our model uses a group sparse regularization [13], we consider the same arbitrary groups of features for all datasets. They correspond to a 2x2 non-overlapping spatial grid considering texture and color features as separated, with a total of eight groups. In relation to free cost parameters of a particular dataset, we use the best results in another similar own dataset and apply over such particular dataset. The sparsity regularizacion free parameters (λ_S, λ_G) are obtained by the best result in an independent dataset where these parameters result to be $\lambda_S = 0.1$ and $\lambda_G = 0.2$ and are applied to all datasets. In relation to performance metrics, we need to reduce the effects of random initial solutions. Therefore, we run the optimization procedure with 20 randoms seeds, and pick the three results with lowest energy and average the metrics. On the other hand, we do not initialize the algorithms with online individual trackers because this input can mask the real performance of each algorithm.

Table 1. Datasets details.

Dataset name	# frames	# persons	Scene
TUD-CAMP	91	10	Street
TUD-CROS	201	13	Street
TUD-STAD	179	8	Street
PETS	795	19	Outdoor
U-HALL-1	200	3	Hall
U-HALL-2	200	9	Hall

4.3 Metrics

For quantitative analysis, we use CLEAR MOT metrics [20]. We use Multi-Object Tracking Accuracy (MOTA), which combines all errors (number of missing targets, false positives and identity switches) into a normalized score between zero and one; and Multi-Object Tracking Precision (MOTP), which measures the bounding box overlap between tracked targets and ground truth detections with a normalized score between zero and one. Additionally, we use a variant of MOTA that penalizes the logarithm of the number of identity switches (MOTAL). In order to consider the performance on rough detections, we also measure the classical precision and recall. Considering [5], we also show the metric identity switches (ID-SW), which accumulates the identity changes of people; and the rate of false detections per frame (FAR).

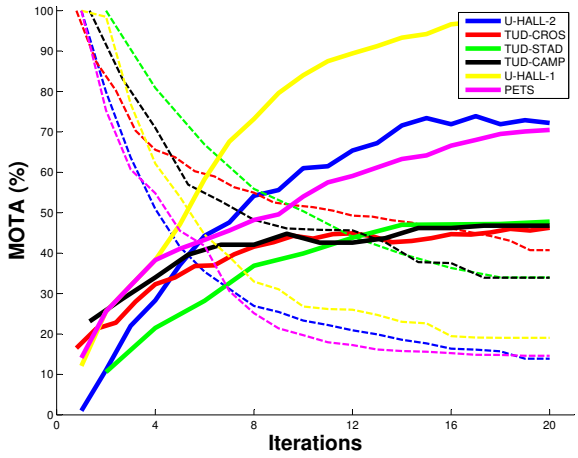


Fig. 3. Convergence of the optimization. The energy keeps decreasing for all iterations (dashed lines, rescaled for visualization). The decreasing of energies is generally reflected in the improving of tracking accuracies (solid lines).

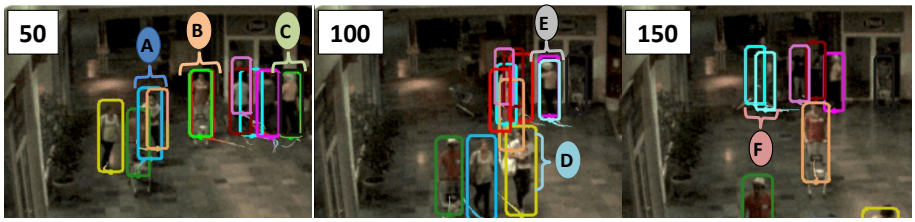
4.4 Results

First, we analyze the convergence behaviour of FDCM by examining the relation between energy and multi-tracking accuracy, MOTA. Figure 3 shows that the most significant performance increase generally appears within the first few iterations, however, the optimization procedure is still able to find better configurations in posterior steps. The energy behaviour is variable and depends on each dataset, for example, while U-HALL and TUD-CAMP tend to converge in terms of accuracy, the energy function appears first stabilized in TUD-CAMP dataset. Now, we are going to analyze and compare multi-tracking performance for the tested algorithms.

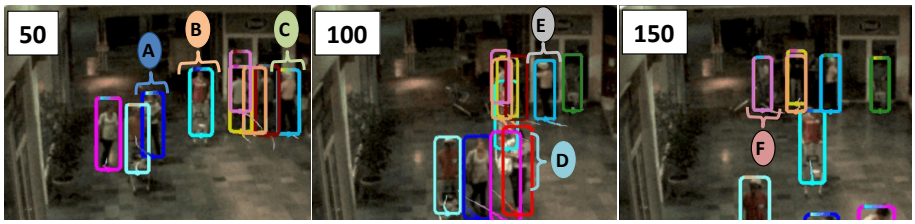
Table 2. The results of normalized scores with a larger margin over 1% are shown in bold font; those with a smaller margin are shown in italic font in order to differentiate the level of the margins. The best results of unnormalized scores (FAR and ID-SW) are shown in bold font. In MOTAL, MOTA, Precision, FAR and ID-SW, FDCM shows better results than DCM. In MOTP and Recall, FDCM is slightly better than DCM.

	Metric						
Dataset	***MOTAL***	***MOTA***	***MOTP***	***PREC***	***REC***	***FAR***	***ID-SW***
TUD-CAMP (DCM)	47.9	43.7	68.9	77.4	66.7	1.01	16
TUD-CAMP (*FDCM*)	**49.6**	**45.9**	*69.2*	**80.1**	**68.2**	**0.85**	**15**
TUD-CROS (DCM)	47.2	42.8	74.1	73.5	73.6	1.47	63
TUD-CROS (*FDCM*)	**53.0**	**47.4**	*74.4*	**78.3**	*74.1*	**1.12**	**50**
TUD-STAD (DCM)	48.2	46.3	75.3	78.1	71.7	1.51	43
TUD-STAD (*FDCM*)	**54.4**	**51.1**	*75.4*	**81.2**	*71.9*	**1.05**	**41**
PETS (DCM)	84.4	83.7	73.5	98.5	*85.7*	0.07	34
PETS (FDCM)	**87.2**	**86.9**	*73.8*	*98.8*	85.6	**0.05**	**29**
U-HALL-1 (DCM)	94.6	94.5	77.2	95.2	99.8	0.11	1
U-HALL-1 (*FDCM*)	**99.1**	**99.1**	77.2	**99.3**	99.8	**0.01**	**0**
U-HALL-2 (DCM)	80.4	75.3	76.2	87.5	94.0	1.12	87
U-HALL-2 (*FDCM*)	**82.1**	*76.1*	**77.2**	*88.4*	*94.7*	**1.03**	**81**

Table 2 shows the results for each dataset, comparing the metrics. As measured by MOTAL and MOTA, our technique is able to clearly outperform the discrete-continuous method by an average of 4 and 3%, respectively. As measured by MOTP, the two techniques show similar results with a slight advantage of 0.4 %. FDCM outperforms DCM in terms of Precision and Recall by an average of 3% and 0.7 %, respectively. By analyzing more qualitative metrics, FDCM slightly beats DCM in the rate of false detections by an average of 0.26. In terms of change of identities, FDCM is superior by 6 less switches of identity on average. We observe that the two techniques achieve a similar results in terms of MOTP and Recall, it is natural as both metrics are related; this contrasts strongly with the MOTA and Precision results where the advantage of FDCM is notorious. These results are explained because FDCM is dependent on visual feature information, which ensures greater precision in tracking due to the use



(a) Tracking results according to DCM



(b) Tracking results according to FDCM

Fig. 4. Multi-target tracking results for U-HALL dataset in frames 50, 100 and 150. In frame 50, the persons A, B and C have one correct detection with FDCM; whereas they have two detections with DCM. In frame 100, the person E has a correct detection with FDCM; whereas it has two detections with DCM. In frame 150, person F has a correct detection with FDCM; whereas it is detected twice with DCM. Nonetheless, there are some mistakes in FDCM, for example, person D is detected twice with FDCM; whereas it is correctly detected with DCM.

of more information; however, this process is fed by a list of detections given mainly by the MRF model, where spatial distance is decisive: two detections with similar visual features have not opportunity to be labelled with the same tracker if they are geometrically far. Moreover, the results in the number of identity switches and the rate of false detections confirm the advantages of our proposal in the multi-target tracking problem.

Finally, Figure 4 shows some examples of the results of our technique, FCDM, in comparison to continuous-discrete model, DCM, for the U-HALL dataset. It can be seen that our technique achieves greater precision in various detections. Visual feature information helps FCDM to identify the target correctly. Nonetheless, our technique has some failures as in frame 100, where the person D has two detections with FDCM and one correct detection with DCM. A possible explanation is uncertainty associated with the visual feature information in each detection. Nonetheless, by jointly considering quantitative and qualitative analysis, our method stresses the usefulness of visual feature information.

5 Conclusions

In this work, we present a method of introducing visual feature information inside a Markov Random Field model. Our model optimizes data association, appearance discrimination and trajectory estimation through an alternating optimization procedure in the discrete and continuous components of MRF energy. In particular, in the discrete component we obtain a sparse set of weights for weighting the features; this considers the natural group of features in order to optimize the global energy function, using an approximation based on the pseudo-likelihood function. The data association is solved using a graph cut algorithm based on an α-expansion. On the other hand, trajectory estimation is solved using analytic fitting of splines to assigned detections. We show that the proposed technique outperforms the base technique based on discrete-continuous optimization. In future work, we plan to use max-margin optimization in order to obtain better multi-target tracking results. Another possible avenue is the use of more detailed features which can be related to semantic parts of pedestrians.

Acknowledgement. This work was funded by FONDEF grant D10I1054.

References

1. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2011) 3457–3464
2. Bar-Shalom, Y., Fortmann, T.: Tracking and data association. Academic Press Boston (1988)
3. Liu, J., Carr, P., Collins, R., Liu, Y.: Tracking sports players with context-conditioned motion models. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 1–8
4. Liu, J., Chu, M., Liu, J., Reich, J., F.Zhao: Distributed state representation for tracking problems in sensor networks. In: Third International Symposium on Information Processing in Sensor Networks. (2004) 234–242
5. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: Conference on Computer Vision and Pattern Recognition. (2012) 1926–1933
6. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. (In: European Conference on Computer Vision)
7. Collins, J., Uhlmann, J.: Efficient gating in data association with multivariate gaussian distributed states. IEEE Transactions on Aerospace and Electronic Systems **28** (1992) 909–916
8. Kuo, C.H., Huang, C., Nevatia, R.: Multi-target tracking by on-line learned discriminative appearance models. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2010) 685–692
9. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33** (2011) 1806–1819
10. Yang, B., Nevatia, R.: An online learned crf model for multi-target tracking. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 2034–2041

11. Butt, A., Collins, R.: Multiple target tracking using frame triplets. In: Asian Conference on Computer Vision (ACCV). (2012) 163–176
12. Hofmann, M., Haag, M., Rigoll, G.: Unified hierarchical multi-object tracking using global data association. In: International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). (2013) 16–18
13. Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008) 53–71
14. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. Journal of Computational and Graphical Statistics **22** (2013) 231–245
15. Li, S.: Markov random field modeling in image analysis. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2001)
16. Lee, S.I., Lee, H., Abbeel, P., Ng, A.: Efficient l1 regularized logistic regression. In: National Conference on Artificial Intelligence. (2006) 1–9
17. Liu, J., Ji, S., Ye, J.: SLEP: Sparse learning with efficient projections. Volume 1813. Arizona State University (2009)
18. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. International Journal of Computer Vision **99** (2012) 259–280
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition. (2005) 886–893
20. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The clear 2006 evaluation. In: CLEAR. (2006) 1–44