Coupling Detection and Data Association for Multiple Object Tracking

Zheng Wu, Ashwin Thangali, Stan Sclaroff and Margrit Betke*
Departments of Computer Science, Boston University, Boston, MA 02215
{wuzheng,tvashwin,sclaroff,betke}@cs.bu.edu

Abstract

We present a novel framework for multiple object tracking in which the problems of object detection and data association are expressed by a single objective function. The framework follows the Lagrange dual decomposition strategy, taking advantage of the often complementary nature of the two subproblems. Our coupling formulation avoids the problem of error propagation from which traditional "detection-tracking approaches" to multiple object tracking suffer. We also eschew common heuristics such as "non-maximum suppression" of hypotheses by modeling the joint image likelihood as opposed to applying independent likelihood assumptions. Our coupling algorithm is guaranteed to converge and can handle partial or even complete occlusions. Furthermore, our method does not have any severe scalability issues but can process hundreds of frames at the same time. Our experiments involve challenging, notably distinct datasets and demonstrate that our method can achieve results comparable to those of state-of-art approaches, even without a heavily trained object detector.

1. Introduction

Although the problem of multiple object tracking has been studied for decades, a robust solution for analysis of visual data does not exist yet due to two major reasons: the lack of robust methods for object detection and poor scalability of data association methods to large numbers of objects. Most previous efforts have therefore followed two distinct directions of research: building stronger object detectors and designing better data association methods. As a result, almost all existing tracking systems use a "detection-tracking design" with two separate modules to address the detection and data association tasks.

We point out that the detection-tracking design has the inherent weakness that it requires the output of the detection module to be reliable in order for the data association module to work properly. Detection errors such as "false alarms" and "missed detections" otherwise propagate to the data association module and false matches need to be corrected later. In contrast, we show that error propagation from detection to data association can be avoided if both tasks, detection and data association, are combined into a single module and solved simultaneously by optimizing a single objective function. This coupling idea appears attractive but introduces new challenges as well: 1) What type of objective function should be used? Many existing detection methods have not even been formalized with an objective function. 2) How can the new objective function be solved? Many current data association methods are complicated and approximate solutions to intractable problems. A new objective function that couples detection and data association might be even more difficult to optimize. 3) How can scalability of the proposed method be ensured? Computer vision systems face demands for being able to track large numbers of objects in dense formations. Given such large input sizes, an efficient algorithm to optimize the new objective function must be found. In this paper, we address all the questions above with a formulation of a coupling function and a method to optimize it. Our method was tested both for monocular and multi-view video.

Inspired by the work of Alahi et al. [1], we propose a detection method with the classic sparse-signal recovery technique [8] for the dense-object tracking scenario, where the number of objects can be up to one hundred per frame and inter-object occlusions occur regularly. This method can be used to detect objects moving on the ground plane as well as objects moving in free 3D space. The sparsity constraint is important here because it can significantly reduce the number of false alarms and serves as a replacement of the heuristic technique of non-maximum suppression of hypotheses. We have to take care, however, that the approach does not lead to overly sparse results, that is, missed detections. We also impose a smoothness constraint for data association where we assume the state of each object follows a first-order Markov process and adopt the classical network flow formulation [9].

The overall objective function has a simple form and can be solved through Lagrange dual decomposition. The

*This material is based upon work partially supported by the National Science Foundation under IIS-0910908 and ONR 024-205-1927-5.

method distributes the coupling formulation to subproblems and coordinates their local solutions to achieve a globally optimal solution. For each subproblem, a very efficient off-the-shelf algorithm is available. The proposed paradigm also permits distributed computing.

Related Work. Our detection method is closely related to the method proposed by Alahi et al. [1], which addressed the pedestrian localization problem without tracking. We assume that foreground pixel estimation is possible, that is, binary foreground images are available as an input, for example, via a background subtraction method [20]. However, instead of using the L_2-norm to explain the image evidence [1], we chose the L_1-norm which strengthens the sparsity constraint without imposing any additional regularization terms. Furthermore, we extended the approach by Alahi et al. to handle the "ghost effect" caused by triangulation of objects in 3D space. We also enriched the formulation, which relied on a single object template, by supporting multiple templates for each object category.

Probabilistic Occupancy Maps were proposed by Fleuret et al. [11] so that, during the detection step, "hard decisions" about the presence of objects do not have to be made. The maps enable the system to make "soft decisions" about the presence of objects on a grid in a probabilistic way. However, unless the estimated density map is very selective around the true locations of objects, non-maximum suppression of hypotheses is needed in order to avoid producing multiple tracks that only slightly differ. Non-maximum suppression is a greedy local operation and does not consider the overall effect for the entire image. In past work, it has been applied during the detection or data association stages. For many approaches to data association, for example that use the classical network flow formulation [3, 9, 19], suppressing false alarms is still a critical issue. Our proposed coupling of detection and data association provides an effective solution for suppressing false alarms. Furthermore, because of the modeling of the joint image likelihood, which is a global operation, our system does not need to perform any type of non-maximum suppression in the detection or data association stages. We also want to emphasize that our coupling framework is general in the sense that it does not require a particular approach to data association. The network flow approach used here may be replaced by other types of data association methods.

We note that the idea to couple detection and data association in a single objective function was proposed by Leibe et al. [14], who coupled the two through a quadratic Boolean function and optimized it according to the Minimum Length Description criterion. We instead base our method on the foundations of Bayesian estimation theory. Our objective function is linear and straightforward to extend to higher-order cases. We stress the problem of scalability, which was not discussed by Leibe et al. Once the number of object grows, the explosion of track hypotheses is always the most difficult challenge, no matter what the exact form of the objective function is. Another important difference is that the trained detector proposed by Leibe et al. is fixed and all detection hypotheses are accumulated in a pool waiting for further selection. In contrast, the behavior of our detector changes under the influence of data association.

An important motivation for designing the coupling framework was to perform occlusion reasoning. By introducing temporal information, we wanted to improve the detection rate of our tracking system, especially when objects are partially or completely occluded. A part-based detector may be able to handle partially occluded objects with sufficient resolution [16], but it fails when objects are completely occluded or the resolution of an object is too small. Research efforts for multiple object tracking typically treat occluded objects as missed-detection events and iteratively grow or stitch tracklets together before and after occlusions [2, 17, 19]. These approaches follow the "detection-tracking strategy" and rely on good detectors for initialization. The output of our approach can be used as an improved starting point for their methods, because our formulation is less sensitive to initialization and requires very few parameters to be set. Our contributions are:

1. A novel framework for coupling the subproblems of detection and data association of multiple-object tracking in a single objective function.
2. A system that optimizes the objective function using a sparsity-driven detection method and a network flow data association method and achieves a high detection rate and tracking accuracy, even when objects are temporarily occluded.
3. A mechanism to suppress false alarms without having to perform non-maximum suppression.

Our framework is flexible – the methods for solving the tracking subproblems are not unique and may be substituted. Because there are opportunities for other instantiations of the coupling framework, we hope that our work provides a new direction for multiple-object tracking research.

2. Coupling Detection and Data Association

We formulate the multiple object tracking problem as a maximum-a-posteriori estimation problem. Given a collection $\mathbf{Y}$ of binary image evidence (foreground pixels), we estimate the state of all objects $\mathbf{X}$ in the scene as follows:

$$\begin{aligned} & \max_{\mathbf{X}} p(\mathbf{X}|\mathbf{Y}) \\ \propto \; & \max_{\mathbf{X}} p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) \\ = \; & \max_{\mathbf{X}} \prod_t p(\mathbf{Y}_t|\mathbf{X}_t)p(\mathbf{X}_1) \prod_t p(\mathbf{X}_t|\mathbf{X}_{t-1}) \\ = \; & \max_{\mathbf{X}} \prod_t p(\mathbf{Y}_t|\mathbf{X}_t) \prod_i p(x_{i,1}) \prod_t p(x_{i,t}|x_{i,t-1}) \quad (1) \end{aligned}$$

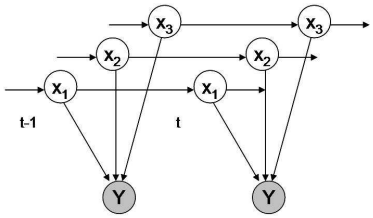

Figure 1. The graphical model for the multiple object tracking problem. The image observation $\mathbf{Y}$ is jointly generated by all objects in the scene, here x_1, x_2, and x_3.

Here, $p(\mathbf{Y}_t|\mathbf{X}_t)$ is the image likelihood conditioned on **all** objects. The joint state of all objects is governed by a Markov process and objects are independent from each other, so $p(\mathbf{X})$ can be factorized with respect to each individual object. We do not further factorize the image likelihood because all objects jointly generate the image. This enables us to handle occlusions. The graphical model for our generative process is depicted in Fig. 1.

Without modeling the likelihood for the entire image but instead making certain independence assumptions, one can further factorize the first term of Eq. 1, a technique used by most earlier tracking approaches. A side effect of the independence assumption is that it yields ad-hoc choices (e.g., non-maximum suppression) because the number of objects is also a hidden variable to be inferred. In contrast, if the likelihood for the entire image is modeled, context and the relationship between objects are naturally brought into consideration. This observation has been recognized widely for the topic of scene recognition [10]. Directly estimating the joint hidden states is difficult here because we do not even know the dimension of the joint state! We propose a decomposition technique to tackle the MAP estimation problem. After taking the negative logarithm of Eq. 1, we rewrite the optimization problem as follows:

$$\begin{aligned} &\min_{\mathbf{X}_1,\mathbf{X}_2} g(\mathbf{X}_1,\mathbf{Y}) + h(\mathbf{X}_2) \\ s.t. \quad &\mathbf{X}_1 = q(\mathbf{X}_2), \end{aligned} \tag{2}$$

where g is the function that models the detection problem, h the function that models the data association problem and q the function that enforces the agreement between the solutions $\mathbf{X}_1$ and $\mathbf{X}_2$ of the two subproblems. More specifically, $g(\mathbf{X}_1,\mathbf{Y})$ is minimized to estimate the states $\mathbf{X}_1$ of objects from image evidence $\mathbf{Y}$ and $h(\mathbf{X}_2)$ is minimized to infer the states $\mathbf{X}_2$ of objects from motion or other types of prior knowledge. Both coupling variables $\mathbf{X}_1$ and $\mathbf{X}_2$ could be discrete or continuous. If a filtering technique that works in the continuous domain is used to solve the data association subproblem, q here could be a quantization mapping. Eq. 2 is a classic setup in operation research: a minimization problem with a coupling constraint. This type of formulation has been applied to the labeling problem, e.g., MRF-based image segmentation [13]. In the remainder of our paper, we will show that the coupling formulation is also useful for solving the tracking problem. We first define functions g and h in Sections 2.1 and 2.2 respectively, by giving specific examples of detection and data association methods.

2.1. Multiple Object Detection

Inspired by the sparsity-driven people localization method proposed Alahi et al. [1], we propose the following L_1-norm minimization formulation as our object detector. First we discretize the space in which objects move. If our target is a rigid object, then for each possible location in 3D, we can reproject the object to the image plane. The reprojected foreground image can be seen as a template or a "codeword." The codeword can be just an image in the single-view case, or a concatenation of images in the multiple-view case. By collecting all codewords in discretized 3D space, we build the dictionary $\mathbf{D}$ for a particular category of objects, see Figs. 2 and 3. The length of each codeword is the size of the observed image(s), while the number of entries in the dictionary is determined by the discretization. Usually, the step of creating the codeword dictionary can be performed offline. But for tracking small objects in a 3D volume, as in Fig. 3, the discretization of the entire volume is infeasible. In this case, we only consider valid triangulations formed from 2D detections using epipolar geometry and build the dictionary on the fly. Here a triangulation is valid if the reconstruction error is within a certain tolerance.

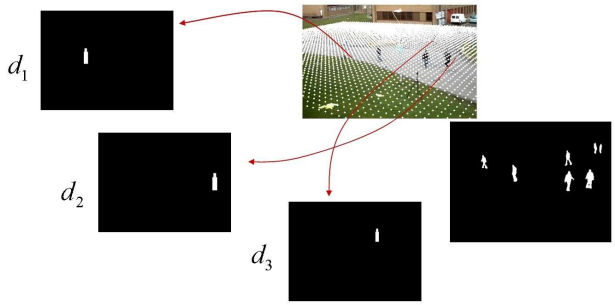

Figure 2. For objects that move on the ground plane, our method discretizes the plane into a grid, where the binary image of the instantiation of an object at each grid point is a codeword (e.g., d_1, d_2, and d_3).

Given the binary foreground image $\mathbf{Y}$ after background subtraction, we want to to find the best way to instantiate the codewords from the dictionary such that the generated image is as close to observation $\mathbf{Y}$ as possible. Mathematically, we want to minimize the following L_0-norm, defined as the Hamming distance from zero, where $\mathbf{X}$ is an binary vector to indicate which codeword to select from the dictio-

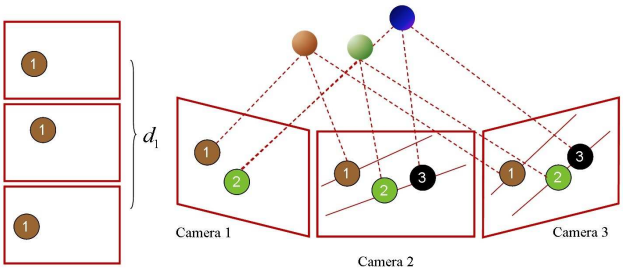

Figure 3. For objects that move in a 3D volume, our method constructs the pool of candidate locations in 3D by triangulation, keeping the reconstruction error below a threshold. The images of the re-projection of each candidate object is one codeword.

nary and N the number of codewords:

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_0, \text{ where } \mathbf{X} \in \{0,1\}^N. \tag{3}$$

Because of the way we construct the dictionary, the selection variable $\mathbf{X}$ also encodes the positions of objects in 3D. The L_0-norm can be seen as our approximation to the negative logarithm of image likelihood $p(\mathbf{Y}|\mathbf{X})$ defined in Eq. 1. It is in general difficult to optimize, so we take the $L1$-norm instead. According to the well-studied sparse signal recovery theory [8], the recovery of $\mathbf{X}$ using the $L1$-norm is "almost" accurate if $\mathbf{X}$ is sparse (only has a few of non-zero entries). Because of occlusion, the real imaging process we model here should actually be a linear combination of codewords followed by a quantization step, i.e., $Q(\mathbf{D}\mathbf{X})$. A common way to handle quantization is to treat its effect as noise. When the observation $\mathbf{Y}$ is considered to be noisy, the sparse signal recovery theory still applies.

We use the primal-dual interior-point algorithm to minimize the $L1$-norm in the following linear programming problem, which is equivalent to the minimization problem in Eq. (3):

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{U}} \ \mathbf{1}^T\mathbf{U} \\ s.t. \quad &-\mathbf{D}\mathbf{X} - \mathbf{U} + \mathbf{Y} \leq 0, \\ &\mathbf{D}\mathbf{X} - \mathbf{U} - \mathbf{Y} \leq 0, \\ &\mathbf{0} \leq \mathbf{X} \leq \mathbf{1}, \end{aligned} \tag{4}$$

where $\mathbf{U}$ is an auxiliary variable. At each iteration, the primal-dual algorithm evaluates a "duality gap" that indicates how accurate the current solution is. This property is useful because the algorithm can come to an early stop when sufficiently accurate results have been obtained.

The above formulation with the L_1-norm is a relaxation version of the original problem ($\mathbf{X}$ is continuous in Eq. 4). We found that, after rounding the continuous solution, the resulting discrete solution can be further improved by a greedy local search. Previous work has tried the L_2-norm [1], which leads to a quadratic programming problem. But so far, throughout our experiments, the L_1-norm produced much better results than the L_2-norm, even though theoretically they are related.

Figure 4. The shape of pedestrians viewed from the front or side can be approximated by two binary templates.

The original L_0-norm minimization can also be directly solved through a sampling-based technique [12] that samples from a rich set of templates. Our approach instead enforces sparsity and uses a minimal number of templates.

In case we need to consider shape variations of the objects, we just enrich our dictionary by providing multiple templates that model these variations. We then impose a uniqueness constraint on our selection variable $\mathbf{X}$, i.e, the system can only choose one of the multiple templates to explain our image evidence as a valid solution. The following modified minimization formulation supports the two versions a and b of a pedestrian template shown in Fig. 4 and is used in our experiments:

$$\begin{aligned} &\min_{\mathbf{X}} \|\mathbf{Y} - (\mathbf{D}_a\mathbf{X}_a + \mathbf{D}_b\mathbf{X}_b)\|_0, \\ s.t. \quad &\mathbf{X}_a + \mathbf{X}_b \leq \mathbf{1}, \mathbf{X}_a, \mathbf{X}_b \in \{0,1\}^N. \end{aligned} \tag{5}$$

Because of occlusion or inaccurate foreground estimation, our detection algorithm cannot always produce the desired solution. We therefore bring in the idea of improving detection results by "temporally smoothing detections" through data association – in the sense that we use our data association results to recover missed detections and suppress false alarms within the detection step. The basic idea is to introduce a temporal prior on our selection variable $\mathbf{X}$ to reflect different preferences. Before we describe this approach, we first review the classic data association formulation that it uses.

2.2. Network Flow Data Association

Classical data association methods represent every detection in every frame as a node in a network and every potential match between detections across time as an arc with an associated cost (Fig. 5). The goal is to find paths through the network that correspond to the trajectories of objects, i.e., sequences of associated detections so that the sum of costs along the paths is minimized. The network flow data association method [9] represents the number of objects in the scene as the amount of flow through the network. As the number of objects present is unknown a priori, the method searches for the amount of flow that produces the minimum cost. Several techniques, involving augmentation of the network and constraints on the flow capacity of arcs, ensure that paths are produced that are mutually exclusive and can represent true object trajectories. We here selected the network flow formulation as our data association method because several efficient algorithms exist [6].

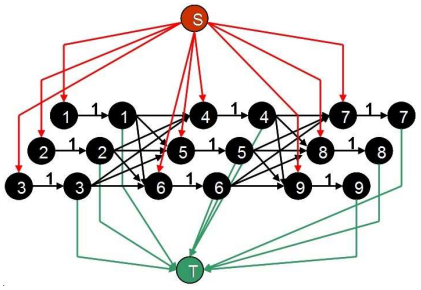

Figure 5. Data association as a minimum-cost network-flow problem. A flow of amount 1 along a path from the source S (track initiation) to the sink T (track termination) represents a single object. Here, three detections, (1,2,3), (4,5,6) and (7,8,9), were made in each of three frames. Duplicate nodes with capacity-one arcs ensure mutually disjoint paths are computed. Here, 3–9 paths can be represented. Minimizing the costs for flow=4, for example, may yield paths S115577T, S2266T, S3344T, and S88T.

2.3. The Coupling Framework

To couple our detection and data association methods, we propose the following objective function, where $\sum_t \|\mathbf{Y}_t - \mathbf{D}\mathbf{X}_t\|_1$ approximates the negative logarithm of the image likelihood $p(\mathbf{Y}|\mathbf{X})$ and the sum of flow costs $\sum_t \sum_i \sum_j c_{i,j}^{(t)} f_{i,j}^{(t)}$ approximates the negative logarithm of the Markov motion prior $p(\mathbf{X})$ described in Eq. 1:

$$\min_{\mathbf{X},\mathbf{f}} \quad \sum_t \|\mathbf{Y}_t - \mathbf{D}\mathbf{X}_t\|_1 + \sum_t \sum_i \sum_j c_{i,j}^{(t)} f_{i,j}^{(t)} \tag{6}$$

$$\text{s. t.} \quad \sum_i f_{i,n}^{(t)} = \sum_j f_{n,j}^{(t)}, \quad \forall \, \textit{frames}\ t, \forall \, \textit{codewords}\ n \tag{7}$$

$$x_{t,n} = \sum_j f_{n,j}^{(t)}, \quad \forall t, \forall n \tag{8}$$

$$f_{i,j} \geq 0 \text{ and } \mathbf{X}_t \in \{0,1\}^N.$$

Selection variable $\mathbf{X}$ indicates the presence of an object at a particular location in discretized space. Flow variable $\mathbf{f}$ is used in the min-cost flow problem, where $f_{i,j} = 1$ means there is a match between detections i and j, which belong to the same track. The cost function (6) is the summation of two *local* terms to minimize; the first term represents the costs of sparsity-driven object detection (Sec. 2.1) and the second term measures the costs of temporal data association in the min-cost flow formulation (Sec. 2.2). The first set of constraints (7) ensures a balance of flow. The second set of constraints (8) ensures *consistency* between the two local variables $\mathbf{X}$ and $\mathbf{f}$. Since this is a linear/integer programming problem, we can apply a general large-scale LP solver to find the optimal solution after linear relaxation. However, because of the special structure of the objective function, we can decompose the problem into two kinds of subproblems, each of which can be solved with an efficient algorithm, and ensure to coordinate the separate minimizers until an agreement is achieved. This approach can be pursued by formulating the Lagrangian dual problem (9) to the minimization problem (6):

$$\begin{aligned} L(\boldsymbol{\lambda}) \quad = \quad & \min_{\mathbf{X},\mathbf{f}} \sum_t \|\mathbf{Y}_t - \mathbf{D}\mathbf{X}_t\|_1 + \boldsymbol{\lambda}_t^T \mathbf{X}_t \\ & + \sum_t \sum_i \sum_j (c_{i,j}^{(t)} - \lambda_{t,i}) f_{i,j}^{(t)} \\ \text{s. t.} \quad & \sum_i f_{i,n}^{(t)} = \sum_j f_{n,j}^{(t)}, \quad \forall t, \forall n \\ & f_{i,j} \geq 0 \text{ and } \mathbf{X}_t \in \{0,1\}^N. \end{aligned} \tag{9}$$

It can be separated into $(T+1)$ independent subproblems, where T is the number of frames:

$$\begin{aligned} g_t(\boldsymbol{\lambda}) \quad = \quad & \min_{\mathbf{X}_t \in \{0,1\}^N} \|\mathbf{Y}_t - \mathbf{D}\mathbf{X}_t\|_1 + \boldsymbol{\lambda}_t^T \mathbf{X}_t \\ h(\boldsymbol{\lambda}) \quad = \quad & \min_{\mathbf{f}>0} \sum_t \sum_i \sum_j (c_{i,j}^{(t)} - \lambda_{t,i}) f_{i,j}^{(t)} \\ \text{s. t.} \quad & \sum_i f_{i,n}^{(t)} = \sum_j f_{n,j}^{(t)}, \quad \forall t, \forall n. \end{aligned} \tag{10}$$

Now the dual problem is to maximize $\sum_t g_t(\boldsymbol{\lambda}) + h(\boldsymbol{\lambda})$ with variable $\boldsymbol{\lambda}$. Here we use a subgradient method to solve the "master problem," the primal-dual interior point algorithm to solve the first T subproblems with parallel computing, and the push-relabel algorithm to solve the min-cost flow subproblem. The dual decomposition technique [7] then yields the following Coupling Algorithm:

COUPLING ALGORITHM FOR TRACKING

For $k = 1, 2, ..., K$ (max iterations), **do**

- Solve T sparsity-constrained detection problems with the interior point algorithm:
  $\mathbf{X}_t \leftarrow \arg\min g_t(\mathbf{X}_t, \boldsymbol{\lambda})$.
- Solve the min-cost flow data-association problem with the push-relabel algorithm:
  $\mathbf{f} \leftarrow \arg\min h(\mathbf{f}, \boldsymbol{\lambda})$.
- **If** $x_{t,n} = \sum_j f_{n,j}^{(t)}$ for all t, **Then Return** $\mathbf{X}_t, \mathbf{f}$
- Update dual variables $\lambda_{t,n} = \lambda_{t,n} + \alpha_k (x_{t,n} - \sum_j f_{n,j}^{(t)})$, $\alpha_k = \frac{1}{k}$ (step size).

Return $\mathbf{X}_t, \mathbf{f}$

The Coupling Algorithm performs as desired in our tracking context: The Lagrange multiplier λ serves as a weighting parameter. For the detection subproblem, a higher value of λ implies a lower preference for detection at a particular location. For the data association subproblem, a higher value of λ leads to a lower edge cost, so it attracts flows passing through that edge. When agreement is

achieved, the optimal global solution is obtained for the primal objective function. The detection output is guaranteed to be smooth because of the influence of data association. The flow computation produces tracks as the final output. By changing the value of λ dynamically, false alarms can be suppressed and detections missed due to occlusions can be recovered.

3. Experiments

3.1. Datasets and Evaluation Metrics

We applied our approach to solve two notably different problems: tracking of pedestrians walking on the ground plane and tracking of wild animals flying in 3D space. We used five video sequences that contained between 23 and 127 objects to be tracked (#O in Table 1). For pedestrian tracking, we used two sequences from the PETS2009 benchmark: the 1st view of sequence S2L1 (795 frames) and of sequence S1L1-2 (241 frames). The ground truth was provided by Andriyenko[1]. For flying animal tracking, we used the infrared videos and ground truth provided by Wu et al. [18]. The data consists of three sequences (100,200,100 frames respectively) with increasing densities of up to one hundred objects per frame. For each sequence, we used the available three views to perform 3D reconstruction.

We used the two metrics Mostly Tracked (MT), the number of objects for which $\geq 80\%$ of the trajectory is tracked, and Mostly Lost (ML), the number of objects for which $\leq 20\%$ of the trajectory is tracked [16]. We also adopted the commonly used tracking metrics Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) [5]. MOTA takes into account false positives, missed targets, and identity mismatches. Its ideal value is 1. In order to compute MOTA, we choose 1 m as the miss/hit threshold for the PETS data and 0.3 m for the infrared data. MOTP measures the average distance between ground-truth trajectories and system-generated trajectories. We computed Euclidean distances for the 3D case (infrared data) and overlap ratio between ground-truth and system-generated bounding boxes for the 2D case (PETS data).

3.2. Implementation Details

We used an improved version of a mixture model for background subtraction [20] to estimate the foreground. For the PETS data, we uniformly discretized the ground plane into square elements of 30×30 cm^2 (Fig. 2) and, to enable comparisons with previously published results, restricted our evaluation to objects moving in the shaded area defined by Andriyenko et al. [2] and shown in Fig. 7. For template comparisons, we assume a pedestrian's height to be 180 cm (Fig. 4) and that a flying animal can be represented by a sphere of 15-cm radius (Fig. 3).

[1] http://www.gris.informatik.tu-darmstadt.de/ aandriye

Our network flow approach computes detection scores on the nodes and transitional costs on the arcs of the network as follows. The score on a node is $-\ln \frac{\rho}{1-\rho}$, where ρ is the ratio between the number of foreground pixels that can be explained by a codeword and the length of that codeword. To reduce the size of our dictionary, we removed codewords whose support was less than 0.2 [1]. For the PETS data, we used normalized correlation to compute the similarity between the two subimages in the bounding boxes. Because of the viewing angle of the cameras, the head of a pedestrian is not very likely occluded. We therefore decided to compute transitional costs only for the upper one-fourth of the bounding box. (This is the only step where we made use of an appearance feature, which is notably simple. More advanced feature representations and comparisons might be helpful but were not our concern here.) To reduce the number of arcs, we did not allow transitions that would model a pedestrian's unrealistic jump of more than 2 m. For the infrared data, the Euclidean distance between reconstructed points in 3D space was taken as the transitional cost. Because of our strategy to couple detection and data association, we only needed to initialize the costs on the arcs, not the nodes. In the first iteration of the Coupling Algorithm, without costs on the nodes, the network-flow minimizer simply chooses a zero flow as the best output. A drastic cost update (by subtracting $\lambda_{i,j}$) then occurs in subsequent iterations of the Coupling Algorithm.

3.3. Quantitative Results

Our quantitative evaluation provides the tracking results of two versions of our system, the single-template coupling tracker (CP1) and the two-template coupling tracker (CP2), on five datasets and compares them to the results of four related approaches, see Table. 1. We found that our trackers are more reliable than competing methods based on the MOTA, MT, and ML scores and comparably accurate based on the MOTP scores.

Our MOTA scores are slightly better and our MOTP scores are slightly lower than those of the occlusion-modeling OM method [2], which achieved the state-of-art performance on the PETS dataset by combining explicit occlusion reasoning, a full-body SVM classifier, tracklet stitching, and initialization via Extended Kalman Filtering. Our bounding box alignment cannot be expected to outperform a classifier-based detector with respect to the MOTP score. By simply enriching the templates that capture the variability of object shape, however, we can already improve performance: The Coupling Algorithm that uses two pedestrian templates (CP2) indeed produced better results than the single-template tracker (CP1) or the Integer Linear Programming (ILP) trackers [3, 4]. These ILP trackers run the detection and network-flow data association modules sequentially and do not take advantage of the complementary

Data	Method	#O	MT	ML	MOTA	MOTP
PETS	OM [2]	23	20	1	0.88	0.76
S2L1	ILP [4]	23	n/a	n/a	≤ 0.6	≤ 0.5
(795	ILP [3]	23	1	8	0.26	0.67
frames)	our CP1	23	20	0	0.94	0.70
	our CP2	23	22	0	0.94	0.70
PETS	OM [2]	36	20	7	0.64	0.67
S1L1-2	our CP1	36	18	1	0.80	0.50
(241)	our CP2	36	24	2	0.89	0.61
Infrared	RT [18]	19	19	0	0.80	9.0 cm
S1	S-RT	19	19	0	0.90	9.5 cm
(100)	our CP1	19	19	0	0.90	9.5 cm
Infrared	RT [18]	75	68	0	0.51	9.9 cm
S2	S-RT	75	62	3	0.81	9.9 cm
(200)	our CP1	75	71	1	0.92	9.7 cm
Infrared	RT [18]	127	60	8	-0.34	11.6 cm
S3	S-RT	127	72	10	0.43	11.6 cm
(100)	our CP1	127	95	5	0.87	11.4 cm

Table 1. Quantitative Results. The OM, ILP, RT, and S-RT trackers sequentially apply the detection and data association modules. Our CP1 and CP2 couple the modules. MOTA is ideally 1, MOTP also 1 or 0 cm. Results are either generated by the authors or copied from published papers. The scores for ILP [4] was read from a chart and were based on a different source of ground truth.

nature of the two subproblems.

Our experiments with the infrared data show that our sparsity-constrained detection method successfully suppressed ghost reconstructions in 3D space. The Reconstruction-Tracking (RT) method [18] was able to track true objects very well but also tracked a lot of ghost objects. By applying our sparsity constraint (S-RT), we could reduce the false alarm rate significantly. Moreover, the performance improvement between S-RT and CP1 shows the significant impact of our coupling idea.

The variables in the Coupling Algorithm can be optimized separately. This property enables us to process a long sequence in a batch mode. Throughout our experiments, we took the whole test sequence as our input. Each detection subproblem can be solved independently through parallel computing. The data association subproblem can also be solved efficiently even for a large network with one million nodes. This is because the complexity of the min-cost flow algorithm is typically governed by the number of edges and a simple "gating" technique allows us to remove most unnecessary edges during the construction of the network. The bottleneck is the L_1 minimization (5–10 s/frame in our experiments). Faster suboptimal algorithms exist, e.g. greedy matching pursuit. Because we were optimizing over the entire discretized space (the size of X in Eq. 3), the overall computational complexity is mainly determined by the quantization N, the number of grid points (2D case) or candidate triangulations (3D case), not the number of objects.

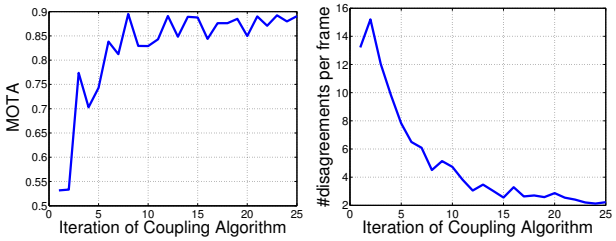

Figure 6. Within the first few iterations of our Coupling Algorithm (here for PETS-S1L1), the MOTA score quickly increases (left) and the number of disagreements per frame between the L_1 solutions and the network flow solutions quickly decreases (right).

In our experiments, we found that the Coupling Algorithm does not need to run many iterations before it reaches a good solution (Fig. 6). The same behavior was also observed in optimization work [15]. In practice, an early stop (25 iterations in our experiments) is sufficient for producing a good suboptimal solution. Other heuristic stopping criteria could also be used. Trackers, such as ILP, RT, and S-RT, that sequentially apply the detection and data association modules could be considered to perform the first iteration of our coupling algorithm. The results in Fig. 6 seem to indicate that the performance of these trackers may increase significantly with additional iterations, if they were placed within our coupling framework.

4. Conclusions

In this paper, we presented a novel multiple-object tracking framework that couples object detection and data association. The objective function is derived from Bayesian estimation theory (1). Its form is general and flexible (2). Our Coupling Algorithm combines a sparsity-driven detection method and a network-flow data-association method within this framework (9). Our approach enabled us to model the likelihood of the entire image so we could avoid non-maximum suppression. Through dual decomposition (10), a coupled objective function is optimized iteratively with off-the-shelf efficient algorithms for each subproblem. In future work, we plan to incorporate different detection and data association methods and seek the best combination. The experiments with both monocular and multi-view datasets show that coupling detection and data association can significantly improve tracking performance compared to the results of sequentially applying each module. Because of our evidence that this performance boost can generalize to existing tracking methods, we hope the proposed coupling concept inspires a new direction for multi-object tracking research.

Acknowledgements

We would like to thank Nathan Fuller and Prof. Thomas Kunz for help with data collection and Prof. David Castañón for feedback on an earlier draft of this paper.

Figure 7. Tracking results. Associated objects are shown with the same colored bounding cylinder/box. The PETS images show sequential frames; the infrared images are three simultaneously-recorded views of sequence S3. The 3D visualization shows 67 flight paths.

References

[1] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst. Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2009. 8 pp.

[2] A. Andriyenko, S. Roth, and K. Schindler. An analytical formulation of global occlusion reasoning for multi-target tracking. In *11th IEEE Intl. Workshop on Visual Surveillance*, 2011. 8 pp.

[3] A. Andriyenko and K. Schindler. Globally optimal multi-target tracking on a hexagonal lattice. In *11th European Conf. on Computer Vision*, pages 466–479, 2010.

[4] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2009. 8 pp.

[5] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008. Article ID 246309, 10 pp.

[6] D. P. Bertsekas. *Linear Network Optimization: Algorithms and Codes*. MIT Press, 1991.

[7] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[8] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51(12):4203–4215, 2005.

[9] D. A. Castañón. Efficient algorithms for finding the k best paths through a trellis. *IEEE Trans. on Aerospace and Electronic Systems*, 26(2):405–410, 1990.

[10] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 8 pp.

[11] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans PAMI*, 30(2):267–282, 2008.

[12] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *The 11th European Conference on Computer Vision (ECCV)*, pages 324–337, 2010.

[13] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2007. 8 pp.

[14] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE PAMI*, 30(10):1683–1698, 2008.

[15] B. Savchynskyy, J., S. Schmidt, and C. Schnörr. A study of Nesterov's scheme for Lagrangian decomposition and map labeling. In *IEEE Intl. Conf. CVPR*, 2011. 8 pp.

[16] B. Wu. *Part based Object Detection, Segmentation, and Tracking by Boosting Simple Feature based Weak Classifiers*. PhD thesis, University of South California, USA, 2008.

[17] Z. Wu, M. Betke, and T. H. Kunz. Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *IEEE Conf. CVPR*, 2011. 8 pp.

[18] Z. Wu, N. I. Hristov, T. H. Kunz, and M. Betke. Tracking-reconstruction or reconstruction-tracking? Comparison of two multiple hypothesis tracking approaches to interpret 3D object motion from several camera views. In *IEEE Wkshp Motion and Video Computing (WMVC)*, 2009. 8 pp.

[19] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, 2008. 8 pp.

[20] Z. Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. In *17th Intl. Conf. Pattern Recognition (ICPR), Vol. 2*, pages 28–31, 2004.