Computer Vision and Image Understanding 136 (2015) 103–116

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Identifying multiple objects from their appearance in inaccurate detections ☆

Julian F.P. Kooij, Gwenn Englebienne, Dariu M. Gavrila *

Intelligent Autonomous Systems Group, Informatics Institute, University of Amsterdam, Sciencepark 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 19 April 2014
Accepted 20 March 2015

Keywords:
Object recognition
Segmentation
Generative model
Unsupervised learning
Latent Dirichlet Allocation
Video surveillance

ABSTRACT

We propose a novel method for keeping track of multiple objects in provided regions of interest, i.e. object detections, specifically in cases where a single object results in multiple co-occurring detections (e.g. when objects exhibit unusual size or pose) or a single detection spans multiple objects (e.g. during occlusion). Our method identifies a minimal set of objects to explain the observed features, which are extracted from the regions of interest in a set of frames. Focusing on appearance rather than temporal cues, we treat video as an unordered collection of frames, and "unmix" object appearances from inaccurate detections within a Latent Dirichlet Allocation (LDA) framework, for which we propose an efficient Variational Bayes inference method. After the objects have been localized and their appearances have been learned, we can use the posterior distributions to "back-project" the assigned object features to the image and obtain segmentation at pixel level. In experiments on challenging datasets, we show that our batch method outperforms state-of-the-art batch and on-line multi-view trackers in terms of number of identity switches and proportion of correctly identified objects. We make our software and new dataset publicly available for non-commercial, benchmarking purposes.

1. Introduction

Traditional tracking-by-detection approaches contain a data association step in which detections are matched to inferred object properties such as appearance, size, and location. However, this poses problems when objects are temporarily (partially) occluded or undetected. Wrong data association can deteriorate the learned object appearances, which further affects future associations. This exposes a chicken-and-egg problem: To localize objects in a scene, image observations need to be correctly associated to candidate objects, which requires knowledge of the object-specific identifying properties. Learning such properties from the observations, however, requires prior knowledge of the presence and location of the objects in the scene. Additionally, camera calibration may be unreliable (i.e. camera orientation may have changed) or unavailable, thus predefined detection windows may not necessarily fit the targets, and objects may exhibit unusual pose or size, resulting in low confidence detections that complicate association even more.

This paper focuses on these related problems, and presents a novel graphical model to determine the number of objects, their appearance, and their location per frame, from possibly *inaccurate* detections. To exploit detections that contain multiple partially occluding objects and background, we seek to loosen the traditional one-to-one relation between detections and objects, and instead infer which of the low-level appearance features present in a detection belong to what target. Key to our method is the adaptation of Latent Dirichlet Allocation (LDA) to "unmix" a number of consistent object appearances in the comparatively large number of detected regions, which are represented as a bag-of-features. This allows us to infer the presence and appearance of an object, even when a single object is responsible for multiple detections and when a single detection spans multiple objects, as often happens in the case of partial occlusion of one object by another. Hence, occluding objects are separated at the feature level and we eliminate the need for special treatment of assignments and appearance updates under occlusion. Additionally, we exploit that in a single frame an object is local to a part of that frame, so that non-overlapping detections are unlikely to both contain the same object. This spatial constraint is enforced by modeling each frame as a mixture of objects whose feature locations have a Gaussian distribution centered at an object's image location. In post-processing, we can optionally "back-project" the feature labels in all images, and segment individual targets. These steps are illustrated in Fig. 1 on a frame from a challenging fight scene.

☆ This paper has been recommended for acceptance by Nikos Paragios.

* Corresponding author.
E-mail addresses: julian.kooij@gmail.com (J.F.P. Kooij), g.englebienne@uva.nl (G. Englebienne), d.m.gavrila@uva.nl (D.M. Gavrila).

http://dx.doi.org/10.1016/j.cviu.2015.03.012
1077-3142/

Fig. 1. Best viewed in color. Top left: example illustrating the challenges we address: detections (shown as black boxes) are inaccurate due to challenging poses, perspective and occlusions. Top right: inferred object presence and location. For each detection window the proportion of pixels associated with an object ID is indicated by a color-coded bar graph on top, black bars indicate background. Bottom left: for each pixel we sample an ID to illustrate the mixing proportions (optional post-processing). Bottom right: per pixel object IDs after global image segmentation (optional post-processing). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To demonstrate the appearance "unmixing" paradigm, we address multi-target association (up to the pixel level) in a batch of frames for scenes with static background and a reasonable upper limit on the expected object count, but where detections from an external object detector may be inaccurate. Our method is compatible with traditional tracking frameworks where motion models reduce the positional uncertainty, since our model incorporates a prior distribution over each object's location. Many different approaches have been proposed to enforce temporal consistency (e.g. merging tracklets [32], searching the space–time volume for globally consistent paths [13,22] or particle filters [11,31]), and state-of-the-art trackers have used strong motion models and the explicit specification of entry/exit regions to push performance. Therefore, our work focuses on appearance without dictating how to exploit such additional cues, and can be seen as complementing recent work on tracking under occlusion [2,22] which solely relies on temporal information. As a result, this paper treats video as an unordered collection of frames without temporal information, using the same positional prior for each individual frame, though for evaluation the output of our model will be compared to that of complete multi-target trackers.

2. Previous work

Our proposed method is related to tracking, image segmentation and object-recognition methods. The body of literature on these is extensive, and here we can only discuss a small set of papers which are most directly relevant.

Tracking-by-detection employs some method to detect target objects in video frames, and combine the detections into consistent tracks. For example, person detectors can be trained on HOG features [12], or by reasoning about spatial occupancy to explain background/foreground masks [5,13,21]. In terms of tracking, we can distinguish between on-line trackers that incorporate new observations on a frame-by-frame basis [11,15,21,31], and *batch* methods that perform global optimization for multiple frames at once [1–3,22,5,24,33].

Multi-view trackers observe targets simultaneously from various overlapping views, and are therefore more robust against occlusion in a single view. Using Probabilistic Occupancy Maps (POM) [13] for detection in calibrated views, [5] applies global appearance constraints to formulate a network flow optimization problem over all frames. This requires defining *a priori* appearance templates for distinct object classes, rather than learning the appearance of individuals from data. The on-line multi-view tracker of [21] relies on background subtraction and voxel carving to find candidate object locations in 3D. Using the Hungarian method [18], candidates are assigned to tracks or labeled as 'ghosts', i.e. faulty correspondences of foreground from different views, based on similarity scores for appearance, size, and Kalman filtered position. Back-projecting voxels to the images yields per-view object masks that take inter-object occlusion into account, which are used to learn the object appearances.

In the single view tracker of [24], appearances are first learned by clustering body part patches from a generic part-based person detector. The trained model is then used to track an individual and makes it possible to reason about self-occlusion. In [33] a part-model is used to deal with other types of occlusions, and fuses tracklets (i.e. trajectory fragments) into consistent tracks while learning a discriminative appearance model for each person. Another use of part-based models is to exploit the dynamics of parts to disambiguate tracks and recover after occlusion, e.g. [1] distinguishes multiple people seen in side-view from their articulated leg pose within a walking cycle.

Part-based models are not the only way to deal with occlusion. [2] adds occlusion reasoning to a continuous energy minimization framework [3] that globally optimizes detection-to-track associations under temporal constraints. In [22] this framework is extended to mixed discrete–continuous optimization for improved data association, with additional global constraints to consider track dynamics and exclude collisions, and keep overlapping tracks separated.

Others treat occlusions as temporary occurrences where no association can reliably be made. For instance, [15] weighs a large

set of features to construct an affinity matrix between tracks and detections, and applies the Hungarian algorithm to find an optimal assignment. Ground truth annotations are required to optimize the weights for the various features using SVM and to handle entering/leaving correctly, and only short-term occlusions are dealt with by maintaining a history of features (occlusions of about a second can result in track termination). Likewise, [31] learns scene-specific feature weights from ground-truth with SVM too, but considerably extends the set of candidate regions with predictions from particle filters. Ref. [11] does not rely on annotations, but uses detector outputs as an object confidence map to weigh a particle filter, and learn discriminative target classifiers on-line. These methods thus focus on reducing the periods where objects are undetected, and/or rely on discriminative appearance models to improve track recovery after such gaps. However, occlusions must still be short and treated as a special case where the discriminative appearance model cannot be applied or updated.

Let us now look at topic models, which originate from the unsupervised analysis for text documents, and have been successfully applied to computer vision tasks by defining these in terms of 'visual document' and a codebook of 'visual words' [26,25,23,29,19,14,17]. Latent Dirichlet Allocation (LDA) [8] is a topic model that represents a document as an unordered bag-of-words, i.e. occurrence counts of each word, and then jointly infers topics as distributions over co-occurring words, and infers per document the mixture of topics within it. There are various techniques for approximate inference in such graphical models [7], e.g. [8] presents variational inference for LDA.

In the context of image segmentation, [25] uses LDA to discover object classes and their appearance in an image database, relying on an external image segmentation algorithm to provide relevant regions. Spatial LDA [28] discovers object classes too, but also segments each image by assigning each feature to one of many overlapping regions to enforce spatially consistent labeling. In the hierarchical model for object recognition by [26], LDA was adapted for unsupervised learning of part-models in a supervised object recognition task. Note that [25,26,28] focus on recognizing distinct object classes, not distinct instances within the same class. Topic models have also been used to discover common motion patterns exhibited by moving objects in video, though none of these methods perform any tracking. Refs. [30,14] avoid the need to track individuals in far-field surveillance. Regarding individual frames as documents, a visual codebook is created by quantizing optical flow in both location and direction, such that topics capture typical [30] (or rare [14]) co-occurring regions of motion. If an external tracker can provide tracks, one can also treat these as documents, and discover typical motion patterns that co-occur within a single track [29,17]. Ref. [34] takes an intermediate approach, finding scene wide patterns from given tracklets without merging them into complete tracks. These methods either quantize the track position and motion into visual words [14,29,34], or learn continuous distributions in the spatial domain [17]. LDA has also been used to compare person appearances in a multi-camera setup [23] from externally provided trajectories, and establish probable matches among non-overlapping viewpoints. In [19] LDA is adapted to discover behavior patterns in a multi-camera CCTV setup from quantized optical flow.

3. Our approach

Our approach processes a collection of input frames simultaneously (i.e. in *batch mode*), and consists of a feature extraction step, a joint inference step, and optionally an image segmentation step. In the feature extraction step, we use the output of an external object detector to determine in all images the presence of the object class of interest (e.g. person) at a given set of (possible overlapping) candidate regions. We keep those detections for which the detector confidence is sufficiently large, though we use a low threshold to keep inaccurate detections too. In each detection we extract low-level appearance features. A feature is described by a *visual word*, obtained by feature quantization, and its *spatial position*. Our inference algorithm however will represent each detection by a bag-of-features, i.e. the word occurrence counts, and mean and variance of the spatial distribution of the features in the detection. In our experiments we regard each pixel in a detection as a feature, and use its binned color value as visual word, thus we record per detection only a color histogram. The spatial distribution is derived analytically from the detection bounding box.

Since detections can contain several (occluding) objects, we do not seek to associate whole detections to a unique target. Instead, the observed features in a detection are considered to be distributed as a mixture of objects. Each object thus has a specific appearance distribution over the feature words, and per image a spatial distribution over feature positions. Hence, we wish to learn a relatively low number of object appearances from a relatively

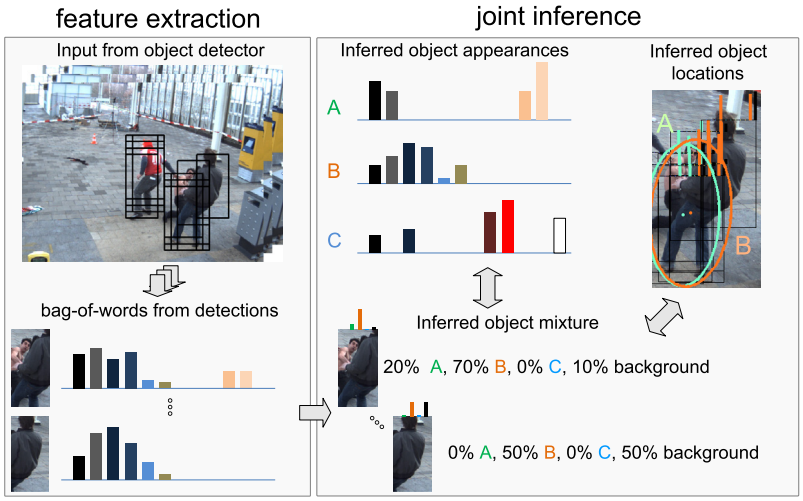


Fig. 2. Overview of our method. First, an object detector is applied to all frames, and color histograms are extracted at each detection. Then, joint inference determines per detection the mixture weights of the objects, the latent object appearances, and the corresponding object locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

high number of mixed detections with unknown mixture proportions. This "unmixing" task is analogous to LDA, where documents, described as a bag-of-words, are decomposed into a small set of topics, i.e. word distributions. However, our generative model additionally accounts for the fact that a target can only be supported in (at most) one spatial region within each frame. The model therefore has latent variables that represent each object's appearance, and location per time step, and the mixture coefficients per detection. Joint approximate Bayesian inference yields posterior distributions over these latent variables, as illustrated in Fig. 2. One can extract object tracks from this posterior (though we have not included any temporal constraints), as it describes for each time step which objects are present, and where.

As noted before, while the generative model is expressed in terms of individual features, described in Section 4, inference will exploit an efficient bag-of-features representation (similar to LDA), as discussed in Section 5. In the optional image segmentation step we can *back-project* the posterior distributions of the feature assignments to the original images to obtain object foreground masks, as will be discussed in Section 5.2.

One parameter that has to be set in advance is K, the *upper limit* of distinct objects that will be detected. Unlike traditional clustering methods, such as K-means and maximum likelihood estimates for mixture models, the variational inference scheme avoids overfitting even when K is set much larger than the true number of objects, due to the use of priors on the model parameters [7]. During the iterative learning processes, candidate objects that are redundant are assigned fewer times to observed features, which makes their appearance and spatial position less specific. This further reduces the probability of assigning such objects in future iterations, until they are not assigned at all anymore. In practice it is preferable to set K not too large either since complexity grows linearly with K, but one could test increasing values for K until the found object count stabilizes.

False positive detections (i.e. background identified as a target object) may affect our method, but in general the model copes with these by either (a) learning an appearance for the background region, regarding it as a mostly static object, or (b) identifies it as background if we include the option of a background distribution in our model (see Section 5.1). A target is not located in an image if it is fully occluded, has (temporarily) left the scene, or not yet entered, or if there are missing detections (i.e. false negatives). The presented generative model does not by itself distinguish or resolve such cases, but should (re-)identify the object in the other frames. A low detection threshold reduces the risk of false negatives. If the frame rate is high and targets move predictably, one could extrapolate an object's motion, though we do not make such assumptions here as we focus on exploiting available detections instead. Further, the proposed method does not rely on accurate detections with non-maximum suppression, no calibrated cameras [21,33], nor knowledge of part-configurations [1,24,33], but uses only on an appropriate object detector to select regions of interest.

Our experiments will focus on keeping track of multiple persons against a static background, while in principle other object types could be dealt with too. Whereas part-based methods try to explicitly model the non-rigid nature of people, our bag-of-features representation drops any rigidity assumption. Because we can use a low detection confidence threshold, we found that a trained person detector based on HOG features [12] even yields usable detections when a person is partially occluded by another person or scenery.

The proposed adaptation of LDA has similarities to [26,28], but there the goal is to recognize various object *classes* in a set of images, which results distinct models and inference schemes. Ref. [26] performs supervised image classification by learning common parts and their spatial configurations per image class. Spatial LDA [28] performs unsupervised image segmentation by learning class appearances that account for consistently labeled pixel neighborhoods. It does not however constraint the occurrence of a single appearance at various places in the image, nor segment occluding instances of the same class, nor use an object detector to focus on a specific class as our method does. And, [28] jointly Gibbs samples the labels of individual features, which we avoid with the bag-of-features representation. Topic models have also been used for video analysis to discover typical objects motion patterns for a particular environment, e.g. without tracking using optical flow at the image level [30,34], or by analyzing tracks provided by an external tracker [29,34,17]. These applications did not address the track association problem itself, nor resolve partial occlusions.

In summary, the main contributions of this paper are:

1. A novel model to jointly localize and learn appearances of an unknown number of objects in a batch of images. We introduce an efficient Variational Bayesian inference algorithm with linear complexity w.r.t. frames, objects.
2. Address scenarios with inaccurate detections, and that contain erratically moving objects and/or have irregular or low frame rates (since objects are identified in the images without considering temporal information).
3. Introducing feature-to-object instead of detection-to-object correspondences to deal with partial occlusion.
4. Pixel-level segmentation of individual objects from object detectors and color histograms only.

4. Model

We now specify the variables and full distribution of our generative model, which is depicted as a graphical model in Fig. 3. There are T time steps with D^t detections at time t obtained from the external object detector, and detection (j,t) at window (i.e. image region) j contains N_j^t features. The tuple (j,t,i) identifies the i-th feature in (j,t), which has two observed properties: a discrete visual *word*, which we represent as an integer $y_{ji}^t \in [1,V]$, and a (2D) position x_{ji}^t in the image plane. We define the word occurrence count of word v in detection window (j,t) as $\mathbf{N}_{jv}^t = \sum_i \delta(y_{ji}^t, v)$, with $\delta(a,b) = 1$ iff $a = b$ and 0 otherwise. For instance, when quantized

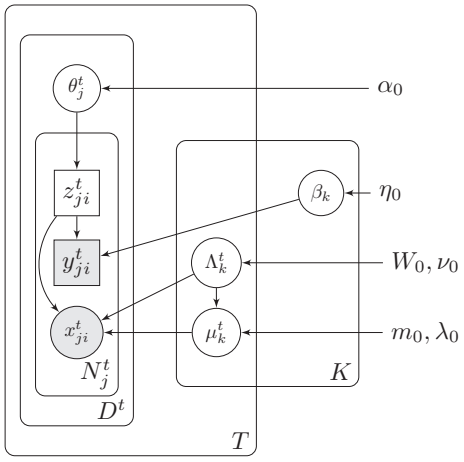


Fig. 3. Our model for tracking combines LDA and Mixture of Gaussians, round nodes are continuous, square ones discrete, observed variables are shaded. Hyperparameters have no border and plates indicate repeated variables. There are K objects, T timesteps, and D^t detections at time t. Detection (j,t) has N_j^t features represented by spatial position x_{ji}^t and discrete word y_{ji}^t. Latent variable $z_{ji}^t \in [1,K]$ indicates which object generated the feature, sampled from the detection's object distribution θ_j^t. Finally, β_k is object k's appearance distribution and $\mathcal{N}(\mu_k^t, (\Lambda_k^t)^{-1})$ its spatial distribution at time t.

pixel colors are used as words, then the vector $\mathbf{N}_{j\cdot}^t = [\mathbf{N}_{j1}^t \cdots \mathbf{N}_{jV}^t]$ is the color histogram of (j,t), where $V = c^3$ if each color channel is discretized in c bins.

We assume that the feature is generated by one out of K objects, indicated by the latent variable $z_{ji}^t \in [1,K]$. Since each detection contains a mixture of objects, the z_{ji}^t in a detection (j,t) follow a multinomial[1] distribution with parameter vector θ_j^t, e.g. element $\theta_{j(k)}^t$ is the mixture weight of object k in (j,t). Each of the K objects defines a multinomial appearance distribution β_k over the V visual words, which is shared by detections in all time steps, and each object has per time step a multivariate Gaussian distribution with mean μ_k^t and precision matrix Λ_k^t over the image positions of features. The full distribution is thus factorized into the following terms:

$$p(x_{ji}^t | z_{ji}^t, \{\mu_k, \Lambda_k\}) = \mathcal{N}(x_{ji}^t | \mu_{z_{ji}^t}^t, (\Lambda_{z_{ji}^t}^t)^{-1}) \tag{1}$$

$$p(y_{ji}^t | z_{ji}^t, \{\beta_k\}) = \mathrm{Mult}(y_{ji}^t | \beta_{z_{ji}^t}) \tag{2}$$

$$p(z_{ji}^t | \theta_j^t) = \mathrm{Mult}(z_{ji}^t | \theta_j^t). \tag{3}$$

We place conjugate priors on the latent variables,

$$p(\theta_j^t | \alpha_0) = \mathrm{Dir}(\theta_j^t | \alpha_0) \tag{4}$$

$$p(\beta_k | \eta_0) = \mathrm{Dir}(\beta_k | \eta_0) \tag{5}$$

$$p(\Lambda_k^t | W_0, v_0) = \mathcal{W}(\Lambda_k^t | W_0, v_0) \tag{6}$$

$$p(\mu_k^t | m_0, \Lambda_k^t, \lambda_0) = \mathcal{N}\left(\mu_k^t | m_0, (\lambda_0 \Lambda_k^t)^{-1}\right) \tag{7}$$

where Dir and $\mathcal{W}$ represent the (symmetric) Dirichlet and Wishart distribution respectively. Details of the used distribution can be found in Appendix A.

5. Inference

Inference on the latent variables is performed on all detections of all time steps jointly. Since exact inference is intractable, we resort to Variational Bayesian (VB) inference to find good approximate solutions. In VB, one approximates the target distribution p by a simpler distribution q, typically by assuming independence between various (latent) variables in p. Inference then proceeds by minimizing the Kullback–Leibler (KL) divergence $\mathrm{KL}(q||p)$, see [7].

Unfortunately, inferring a variational distribution for each latent assignment label z_{ji}^t is computationally demanding. In VB for standard LDA this problem is mitigated by noting that per document all observations of the same word v are exchangeable. Instead of reasoning about individual indicators z, a multinomial distribution over the *number* of words v assigned to object k in a document can be used, which leads to a more efficient variational inference scheme [8]. In our model, however, features with the same visual word in a detection are not exchangeable, since features do not share the same position x_{ji}^t. We therefore remove the correspondence between positions x_{ji}^t and words y_{ji}^t, and instead model the x_{ji}^t in the detection window as i.i.d. normally distributed random variables. More precisely, if window j has center $\widetilde{x}_j$, width w and height h, then we analytically derive (see Appendix B for details)

$$x_{ji}^t \sim \mathcal{N}\left(\widetilde{x}_j, \widetilde{\Sigma}_j\right) \text{ with } \widetilde{\Sigma}_j = \begin{bmatrix} w^2/12 & 0 \\ 0 & h^2/12 \end{bmatrix}. \tag{8}$$

We can now avoid inferring the assignment of individual features in the variational approximation, and use instead the multinomial distribution per detection over the number of features assigned to each of the K candidate objects, as is the case with VB for standard LDA.

In Appendix C we present detailed derivations of the variational distribution, and all update equations. We group the update equations into the three parts, which must be executed iteratively since updates are coupled:

- the updates for the object priors and appearances, which are similar to those for standard LDA [8,4].
- the updates for the spatial distributions, which follow the VB updates for standard MoG (see [7], Chapter 10).
- the assignment updates which, as expected, are a combination of the assignment updates found in LDA [8] and MoG [7], but where special care has to be taken to include the uncertainty of a feature's position x_{ji}^t.

To initialize inference, we sample the parameters of the variational distributions on all feature-to-object assignment counts uniformly, and normalize the distribution for each detection (see Appendix C.6). Effectively, all observations are initially assigned almost uniformly to all K candidate objects, and therefore all objects have almost the same spatial distribution in each image, and similar appearances. At subsequent iterations however, the small random variations in the appearance and spatial distributions becoming increasingly distinctive. As more features are assigned to certain objects, the uncertainty on their appearance and spatial distributions decreases, while the distributions of objects with few features reduce to the prior which avoids overfitting. Fig. 4 illustrates the procedure at different iterations, showing for two frames in a sequence how the spatial and assignment distributions change, until eventually all features are assigned to only four candidate objects.

Direct application of the variational update equations can unfortunately get stuck in a suboptimal solution. Initially when appearances distributions are not yet representative of the true objects appearances, we observe that the positional likelihood tends to outweigh the appearance likelihood, forcing inferred objects to quickly cluster detections spatially without discovering meaningful and distinct appearances. To guide inference through this parameter space with many local optima, the variational updates initially only include the spatial distribution at every fifth iteration, until the appearance distributions converge. We found that this provides good initial distributions for the final variational updates, which include every term at each iteration, and are again performed until convergence.

The complexity of the variational inference updates is $O(T \times D \times K \times V)$, where $T \times D$ is the total number of detections. Computationally, the appearance and prior updates are mostly sums and multiplications that are fast to execute. Updating the spatial locations is the most costly operation, though these could be run in parallel for objects and time steps. Assignment updates can also benefit from parallelization.

5.1. Background modeling

The features within a detection may stem from one or more objects, but could also come from the background (or occluding foreground) of the scene. In our experience, when the background is sufficiently uniform this tends not to be a problem, especially when the detections are accurate and therefore contain many

[1] For clarity later on we use the multinomial distributions as a generalization of the categorical distribution, for instance z_{ji}^t can be represented equivalently as a one-of-K vector.

(a) iteration 0 (b) iteration 60 (c) iteration 120 (d) iteration 180

Fig. 4. Best viewed in color. Inferred object mixtures and locations for frame 10 (top row) and 260 (bottom row), at various iterations of variational updates (columns). Shown are for the K objects (in pseudo colors) their expected spatial distributions per frame, and feature assignment counts on top of each detection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

object features and few background features. In those cases, all object appearances include some low probability of observing the background features. But when the detections contain a lot of background, and/or the background varies across the image, the model will infer additional objects associated with the background features, resulting in many false positives. For this reason, and with the upcoming image segmentation step of Section 5.2 in mind, we wish to exploit background information in our model to segment fore- and background.

For each detector window location we create a background object, such that each detection (j,t) can be a mixture of $K+1$ possible object models: the K objects plus the background of window j. Background object j then has appearance distribution $\dot{\beta}_j$, and we fix its spatial distribution to $\mathcal{N}(\widetilde{x}_j, \widetilde{\Sigma}_j)$ as in Eq. (8). This means that background objects are not treated differently from the K target objects, except that we do not update their location (the most computationally expensive part of our method), and that instead of prior η_0 we use window specific background prior $\dot{\eta}_{0j}$. In our experiments we obtain $\dot{\eta}_{0j}$ by taking the input images, removing the features in the windows that will be kept as detections (i.e. erase object appearances), and count the remaining features within each window j over all images. Standard background subtraction, performed as pre-processing step, evaluates only once the observation likelihood under a background distribution, and may discard actual foreground with appearance similar to the background. Our approach (re)evaluates the feature likelihood for both background and all object appearances during inference, and background segmentation is a result of inferring the assignment labels. One may come up with different background models for different scenarios. For instance, if the background has little variation a single background appearance shared by all windows might be used. Or, if reliable foreground masks are in fact available, those could be used to remove background features and no background model would be necessary.

5.2. Image segmentation

A benefit of using a generative appearance model instead of a discriminative one (e.g. [11]) is that the model can be utilized for image segmentation. After the variational distributions converged, the distributions $q(z_{ji}^t = k|\cdot)$ express the probability that a word y_{ji}^t in detection (j,t) belongs to object k. In a post-processing step we can then compute the object appearance responses in the original input image frames by *back-projecting*. The object probabilities from overlapping detections are averaged per pixel, and pixels that are not in any detection window are fixed to background. Subsequently, we use the object responses as input for standard multi-label image segmentation software.[2] We use energy minimization [10,9,16] to determine per pixel the optimal object label, taking into account both object responses and labeling consistency between neighboring pixels. The energy minimization problem formulation makes assumptions complementary to the bag-of-features representation in our model, which does not enforce consistent neighborhood labeling. While that enables very fast variational updates, it also results in more noisy assignments of features to objects. Since the human visual system is very sensitive to edges, approximate global optimization of neighborhood consistency on the complete frame obtains visually more pleasing results.

6. Experiments

After explaining the used evaluation metric in Section 6.1, we present results on two types of datasets. Section 6.2 discusses experiments using inaccurate detections on challenging sequences containing several people fighting. In Section 6.3 we additionally show how our method is competitive with different state-of-the-art methods on publicly available data where motion models play a major beneficial role. These experiments also show how the quality of the detections affect the tracking results. We make both our software and our new dataset publicly available for non-commercial, benchmarking purposes.

6.1. Evaluation metrics

For quantitative evaluation we use common metrics for evaluating multi-target tracking performance, namely the Mostly Tracked (MT) and Mostly Lost (ML) measures from [20] (similar to [22]), and the CLEAR MOT metrics [6]: Identity switches (IDS), False Positive ratio (FP) and False Negative ratio (FN), the Multiple Object Tracking Accuracy (MOTA) and the Multiple

[2] http://vision.csd.uwo.ca/code/.

Object Tracking Precision (MOTP). The statistics are computed by comparing object bounding boxes to ground truth bounding boxes, and we keep correspondences as long as the boxes intersect. To obtain bounding boxes from our model's output, we use the estimated object locations given by the Gaussian $\mathcal{N}(\mu_k^t, (\Lambda_k^t)^{-1})$. A bounding box rectangle is created with its center at the Gaussian's mean position, and the width and height of the box are estimated from the covariance matrix with the same formula as in Eq. (8).

The complete set of measures and statistics are as follows:

- *MT*: the percentage of ground truth tracks which are matched to a tracker for more than 80% in length.
- *ML*: the percentage of ground truth tracks which are matched to a tracker for less than 20% in length.
- *IDS*: the number of ID switches, which counts the number of times a ground truth object is matched to a different track than that of the previous time instance.
- *FP*: the false positives ratio in percentage, the number of false positives divided by the number of ground truth objects.
- *FN*: the false negative ratio in percentage, the number of false negative divided by the number of ground truth objects.
- *MOTA*: an accuracy score in percentage, based on the false positives, false negatives and identity switches relative to the total number of matches.
- *MOTP*: a precision score in percentage, that measures the average overlap between ground truth and matched tracker results. In our case, we measure overlap of bounding boxes in the 2D image plane.
- *GT*: the number of ground truth tracks.

6.2. Experiments on fighting sequences

We evaluate our method both quantitatively and qualitatively on a new dataset with three challenging sequences, and compare to the tracker output obtained from the on-line multi-view tracker [21], that tracks people in 3D using multiple overlapping camera viewpoints, and from the state-of-the-art global energy minimization tracker with exclusion constraints [22].[3] Each sequence consists of $T = 300$ frames at 20 fps with a resolution of 752×560 pixels, and was simultaneously recorded from three overlapping and calibrated wide-angle viewpoints. Unlike other available tracking benchmarks with people walking normally, individuals exhibit a lot of rapid motion changes and uncommon poses (e.g. kicking, punching, falling), and many long term inter-person occlusions and interactions occur (see also Figs. 1 and 4).

Baseline method [21] relies on voxel carving with segmented background in all three views to generate volumetric detections in 3D. It uses the Hungarian algorithm [18] to on-line assign detections to tracks, or label them as 'ghosts'. For data association it uses appearance features similar to our method (color histograms), and separates appearances of inter-occluding objects by back-projecting the voxels to the images. Unlike our method, it incorporates a motion model for temporal consistency, and as it benefits from more extensive sensory data, it does not suffer from single-view occlusions. The method by [22] uses the same single view person detections as our method, and also processes all frames in a single batch. The tracker combines discrete optimization to associate detections to candidate tracks, and continuous optimization to fit tracks as splines to detections in the space–time cube, focusing on temporal consistency rather than appearance.

[3] Code was made available by the authors of [22] at https://bitbucket.org/amilan/dctracking.

6.2.1. Experimental setup

We manually annotated ground truth tracks with bounding boxes in a single view every 5 frames. The boxes completely contain a person (thus due to the unusual poses in the data, the aspect ratios of the boxes vary considerably), and we also annotated people when they are almost completely occluded in the available view, to the best of our capabilities. For our method and [22], we obtained object detections from a HOG person detector [12] of fixed size (200×100 pixels) on a fixed grid (at steps of 20 pixels), applied across each frame with a low threshold (0.2) to reduce false negatives. Detections are not well tuned to the particular size of observed targets, and a target can span several detections, i.e. detections are inaccurate. We performed non-maxmimum suppression for [22], as we found that improved the results, but not for our method (i.e. we used all detections that passed the threshold). For the quantitative comparison the 3D voxel output of [21] is projected to the same view to create track bounding boxes.

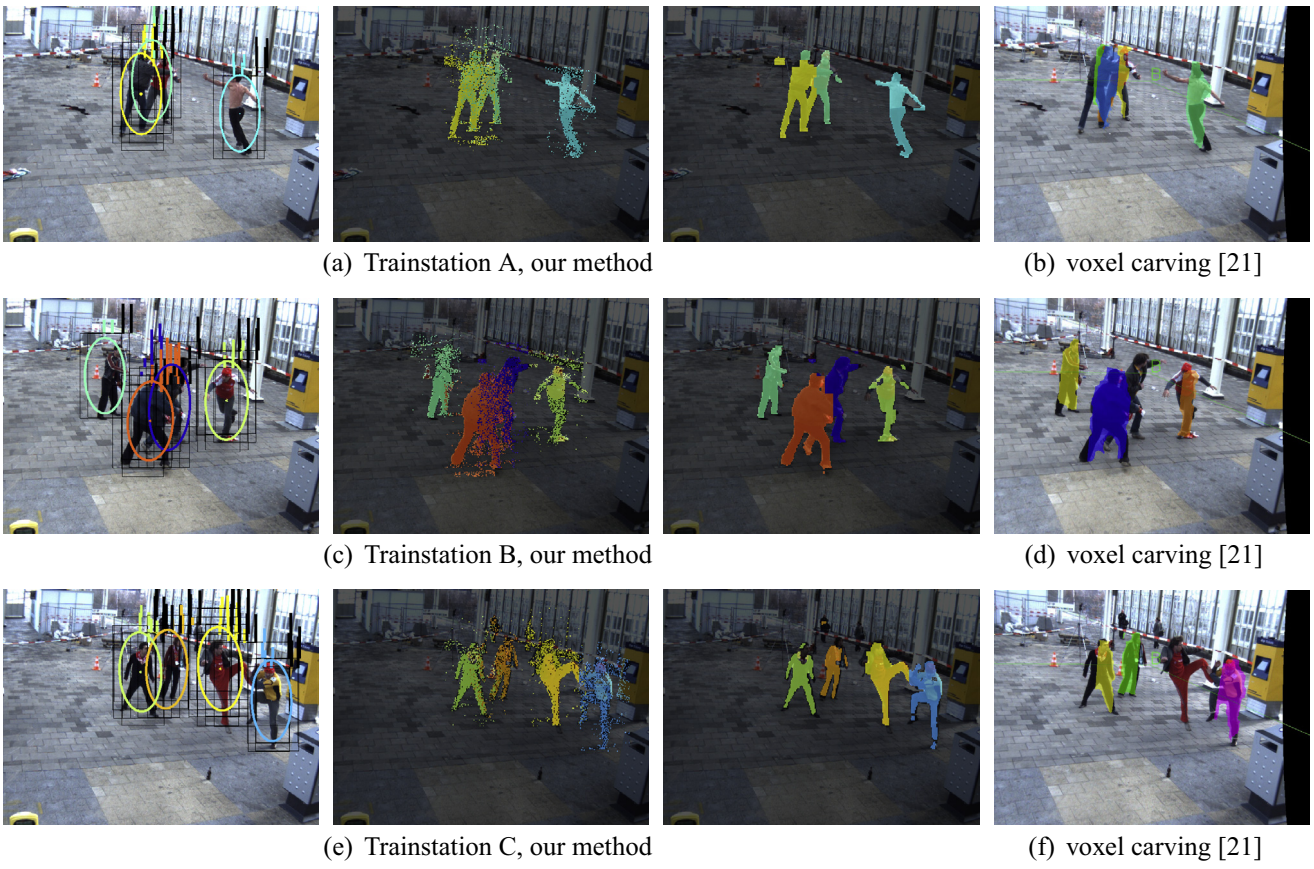
We quantized the Hue-Saturation-Value (HSV) colorspace into 6 bins per color channel ($V = 6^3 = 216$). The parameter vector of the object prior, α_0, is a K-dimensional vector, with all elements k set to $\alpha_{0(k)} = 10^{-5}/\mathrm{K}$. The word prior, η_0, is a V-dimensional vector, with all elements v set to $\eta_{0(v)} = 10^6/\mathrm{V}$. We set a weak spatial prior, $\nu_0 = 2$, with low precision and which prefers a 2:1 height to width ratio, $W_0 = 10^{-4} \times \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$. Further, $\lambda_0 = 10^{-1}$ and $\mu_0 = [I_w/2, I_h/2]^\top$ where $[I_w, I_h]$ are the image dimensions in pixels. Inference was run for 700 iterations, using $K = 10$ objects. After these iterations, we keep the object locations at time instances for which the expected number of assigned features exceeds 2500.

6.2.2. Results

Qualitative results of both methods are shown in Fig. 5, which illustrates that our method actually yields better foreground segmentation than [21] despite using only one camera.[4] The method of [22] does not perform image segmentation, but it should be noted that tracked bounding boxes exhibit smooth motion, though tracks are highly fragmented. As shown in Table 1, our method outperforms the baselines on the most important measures in terms of correct identification of the objects, namely the MT, ML, IDS and MOTA scores. In contrast, the methods of [21,22] tend to segment tracks into many small ones, leading to a lot of identity changes, or fail to correctly recuperate tracks, leading to low Mostly Tracked scores. The higher MOTP scores of [21] are due to the very accurate estimates of target size that voxel carving in the multi-view calibrated-camera setup allow. We now discuss the various sequences in more detail.

6.2.2.1. Trainstation A. In this scene, a person intervenes between two aggressive men, and keeps them apart to deescalate the situation. As seen in Table 1, our method consistently identifies and localizes the three individuals. Our method outperforms the baselines on all measures except the MOTP score. Ref. [21] benefits from accurately sized bounding boxes from voxel carving, but performs more poorly on this sequence with regard to the other measures. First, due to the multiple persons, their proximity, and the dynamic background, there is a large number of 'ghosts' in each frame. As a result, that method tends to initialize multiple tracks on a single person, or incorrectly assumes that an actual person is a ghost. The temporal model does not help much in making a correct distinction because of the erratic motion of the people. Second, since the appearance model of each track is updated after back-projecting the voxels, incorrect identification of 'ghosts' leads

[4] Associated video clips can be found on the websites of the authors.



(a) Trainstation A, our method (b) voxel carving [21]

(c) Trainstation B, our method (d) voxel carving [21]

(e) Trainstation C, our method (f) voxel carving [21]

Fig. 5. Best viewed in color. Examples from the sequences used in Section 6.2. Colors are randomly assigned to objects. First column: used detections (black boxes), inferred object location, shown by the Gaussians $\mathcal{N}(\mu_k^t, (\Lambda_k^t)^{-1})$, and per detection window the proportion of pixels associated to each object, indicated by color-coded bars (black bar indicates background). Second column: color-coded object labels sampled per pixel from the posterior. Third column: Image segmentation based on posterior. Last column: Corresponding output from [21]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Quantitative comparison to a state-of-the-art tracking method on three particularly difficult datasets. Reported measures are MT, Mostly Tracked; ML, Mostly Lost; IDS, Identity Switches; FP, False Positive ratio; FN, False Negative ratio; MOTA, Multiple Object Tracking Accuracy; MOTP, Multiple Object Tracking Precision; GT, number of Ground Truth objects. Bold numbers highlight the best-performing method. See the text for more details.

Method	MT (%)	ML (%)	IDS	FP (%)	FN (%)	MOTA (%)	MOTP (%)	GT
Trainstation A								
Our method (1 cam)	**100.0**	0.0	**0**	**0.7**	**4.7**	**94.7**	56.5	3
Liem, DAGM 2011 [21] (3 cams)	66.7	0.0	15	52.5	11.0	34.7	**64.1**	3
Milan, CVPR 2013 [22] (1 cam)	33.3	0.0	5	1.7	32.2	65.6	61.0	3
Trainstation B								
Our method (1 cam)	**100.0**	0.0	0	2.9	**4.4**	**92.7**	59.3	4
Liem, DAGM 2011 [21] (3 cams)	75.0	0.0	0	**0.8**	18.6	80.6	**69.8**	4
Milan, CVPR 2013 [22] (1 cam)	25.0	0.0	2	3.5	27.9	68.4	51.7	4
Trainstation C								
Our method (1 cam)	100.0	0.0	**0**	**0.6**	7.9	**91.5**	61.0	4
Liem, DAGM 2011 [21] (3 cams)	100.0	0.0	3	5.0	**7.2**	87.5	**64.9**	4
Milan, CVPR 2013 [22] (1 cam)	25.0	0.0	13	5.8	27.0	66.2	53.0	4

to inaccurate appearance models, which hurts tracking in subsequent frames even more.

The temporal smoothing in [22] also results in multiple track fragments and identity switches, though also a reasonable MOTP score, since the spline fitting prefers smooth paths over sudden motion changes. There is a large number of false negatives however, since targets remain close together for prolonged time and the tracker outputs at most frames only two tracks.

Our method, on the other hand, does not have to deal with incorrect correspondences between multiple viewpoints as [21], and can successfully 'unmix' all three objects in only a single view, even during occlusion. Compared to the other batch method [22], we benefit from appearance cues to keep people apart, while the poor appearance models learned by [21] induce more identity switches as tracks are often lost and replaced.

6.2.2.2. Trainstation B. This is a fighting sequence containing four people, with a real-world artifact that is usually excluded from experiments: during recording, 13 frames were dropped by the recording equipment, resulting in a abrupt change in the sequence. Because we do not rely on temporal information, our method has no problem with inconsistent intervals, and thus again correctly

Table 2
Quantitative comparison to reported results of state-of-the-art tracking methods. Used measures are MT, Mostly Tracked; ML, Mostly Lost; IDS, Identity Switches; FP, False Positive ratio; FN, False Negative ratio; MOTA, Multiple Object Tracking Accuracy; MOTP, Multiple Object Tracking Precision; GT, number of Ground Truth objects.

Method	MT (%)	ML (%)	IDS	FP (%)	FN (%)	MOTA (%)	MOTP (%)	GT
TUD-campus								
Our method (*full-gt*)	37.5	12.5	0	3.0	25.1	71.9	67.4	8
Our method (*clean-gt*)	75.0	12.5	0	3.2	18.3	78.4	67.4	8
Our method (*weak-detec-full-gt*)	62.5	0.0	0	8.6	21.5	70.0	53.3	8
Our method (*weak-detec-clean-gt*)	75.0	0.0	0	9.1	16.8	74.1	53.3	8
Breitenstein, TPAMI 2011 [11]	–	–	2	0.1	26.4	73.3	67.0	–
Yan, ECCV 2012 [31]	–	–	0	0.0	15.18	84.82	67.76	–

identifies all people. The tracker of [21], on the other hand, misses one person after the unexpected gap because it violates its motion model's assumptions in 3D space. This track is not recovered because the detections in the center of the scene, a low-probability area for creating new tracks, are typically considered 'ghosts'. The batch tracker [22] creates again many fragmented tracks, but also has moments where there are too few detections due to non-maximum suppression. Without the non-maximum suppression however, we found it would create many more short spurious tracks.

6.2.2.3. Trainstation C. In this third fight sequence people punch each other, fall down, get up, kick and jump. Still, [21] and our method are capable of tracking all four persons most of the time. However, on-line tracker [21] does lose people several times, and initializes extra tracks as people change motion very quickly, re-enter the scene, or are divided into multiple detections in the voxel space due to wide spread arms and legs. This problem is even worse for [22], which cannot rely on the multiple viewpoints accessible to [21]. For our method such motion changes, scene re-entering, or multiple detections per person do not pose a problem. In fact, due to the low detector threshold even a person performing a flying kick (i.e. kick performed in mid-air) is detected and correctly identified, as can be seen in Fig. 5(e). The slightly higher FN rate is due to the heavy occlusion in the used viewpoint.

6.3. Experiments on TUD dataset

The publicly available *TUD-campus* dataset consists of $T = 71$ frames, showing several persons seen from the side walking in straight lines from left to right, or right to left, across the scene. Due to the low viewpoint and many people occurring simultaneously in the scene, partial and full occlusions occur regularly. Unlike the fighting sequences used in Section 6.2, this sequence is much shorter, contains common poses (all observed poses are from gait cycles with little to no articulation except for legs), and predictable motion patterns (fixed speed without curvature). In fact, using a person detector at various scales and a fine spatial grid, and after non-maxima suppression, quite accurate detections can be obtained for this sequence, such as those made available by the TU-Darmstadt.[5]

While this sequence is not representative of the problems our method addresses, we will demonstrate that the method in principle also handles cases that are well suited for traditional trackers. Results will be compared to those of to the state-of-the-art trackers of [11,31]. The compared trackers use an extended set of detection regions obtained from particle filters [31], temporal models [11,31], more extensive features [11,31], and training association weights on ground truth annotations [31].

[5] Detections and groundtruth for *TUD-campus* can be found at http://www.gris.informatik.tu-darmstadt.de/aandriye/data.html.

6.3.1. Experimental setup

Since our method solely focuses on identifying targets and their appearances without committing to any particular way of modeling temporal consistency, we consider four types of results for our method to relate it to the these trackers:

- *full-gt*, using the publicly available objects detections as input. The $T \times D = 220$ detections are of varying sizes and obtained with non-maxima suppression, such that detections correspond well to various targets. However, only detections with high confidence have been provided, and this translates to many false negatives, i.e. ground truth bounding boxes that do not overlap with any detection. Since our proposed method does not add missing detections, we expect to miss ground truth objects.
- *clean-gt*, using the same publicly available objects detections as input, but removing ground truth without any detector overlap. This gives a better indication of how our method copes with the available detections.
- *weak-detec-full-gt*, using an object detector of fixed size with low threshold, and no non-maxima suppression. These detections are inaccurate, and detections are not well tuned to the particular size of observed targets. Similar to the experiments of Section 6.2, we applied a HOG person detector [12] of fixed size and a 1:2 width-height ratio on a fixed grid across each frame, and set a low threshold of 0.2 on the confidence threshold to accept detections, though we have not spent time optimizing this threshold for best performance.
- *weak-detec-clean-gt*, uses the same detections as in *weak-detec-full-gt*, but ground truth without any overlapping detection has been removed, as was done in *clean-gt*.

For evaluation we compare the bounding boxes to the bounding boxes of the ground truth annotations made available by the TU-Darmstadt. For *clean-gt*, 25 of 303 (8.25%) ground truth annotations were removed because of missing detections, for *weak-detec-clean-gt* this resulted in removing 17 of 303 (5.61%) annotations. We set $K = 15$, and quantize the HSV colorspace in 8 bins per channel ($V = 512$).

6.3.2. Results

Quantitative results of our method are shown in Table 2, together with the published results of [11,31].

Comparing the statistics computed for *clean-gt* and *full-gt*, we see that missing detections in the detector output strongly affects, as expected, the Mostly Tracked (MT), False Negative (FN) and thus MOTA score. By just considering ground truth that does contains detections (i.e. *clean-gt*), we can see that our method can compare to state-of-the-art trackers, even though we rely only on simple color features and did not include temporal information. Yan [31] does report better FP and FN, and as a result a higher MOTA score. Indeed, this sequence is well suited to their constant velocity motion model, which via particle filtering also yields good additional detections, and also relies on more features. But we also

observe that our method's scores are lower because the method identifies the tight group of three people in the background as a single object, since the people's appearances occur consistently together. In the ground truth annotations, however, these persons are labeled individually, but even there not all three are fully annotated throughout the sequence due to the strong occlusion. This highlights a common difficulty with creating annotations and the choice of deciding which and when people should be included in the ground truth.

In case of *weak-detec-full-gt*, we see that even with the inaccurate detections we obtain good results, and *weak-detec-clean-gt* yields even a better MOTA score than *full-gt* due to less false negatives, although the MOTP score is relatively low. This is mainly due to the fact that the ground truth bounding boxes of the more distant people are much smaller than the detection windows we used. Since the precision matrix of the localized objects depends on the precision (i.e. inverse covariance) matrix of the detection windows, the bounding boxes are at least as large as the used detection window size. Therefore, the overlap with people much smaller than the detection window size is generally low, as expressed by the MOTP score.

6.4. Implementation

Our single core Matlab implementation ran on a 2.67 Ghz 64-bit CPU. It takes for instance ± 232 s for 700 iterations on Trainstation A with inaccurate detections ($T = 300, K = 10, V = 216$, $T \times D = 3513$). Inference on the TUD sequence with $T \times D = 220$ good detections ($T = 71, K = 15, V = 512$) takes ± 64 s for 900 iterations. Runtimes could be improved by performing variational updates in parallel over object, detections and time steps, and porting code to C/C++.

7. Discussion

In this paper we have cast the related problems of discovering the number of objects, detection to object association, occlusion handling, appearance modeling, and foreground and background segmentation, as an inference problem in a single graphical model. This formulation allows us to benefit from advancements in the actively researched area of Bayesian inference in graphical models, and provides a unified framework to deal with ambiguities and missing data. Traditional trackers take an ad hoc approach to deal with these types of problems, relying on thresholds and staged filters instead. For example, an initial background subtraction stage will make a hard binary decision at the pixel level, but is not informed by the learned foreground color distributions used in a later stage.

However, our method does not stand in direct competition with these traditional trackers, as different assumptions are made about accuracy of detections, occlusions, and pedestrian movement. Future research will investigate various ways to combine the strength of our method, which has focused on identifying objects using their appearance only, with state-of-the-art trackers and the complementary information that they exploit.

First, we currently rely on a K-dimensional symmetric Dirichlet prior, which bounds the number of possible targets. To remove this limit, one could run the procedure for increasing K until the object count converges. Alternatively, the model could be reformulated as a Dirichlet Process (DP) mixture model (our Dirichlet prior can be seen as a fixed-size approximation of the infinite DP prior [27]) to adapt K during a single inference run.

Second, there is currently no temporal relation between the frames, but appearance alone may not be sufficient to resolve certain ambiguities that are easily resolved when considering temporal constraints. How temporal information can best be included in our model is an open research question, and many possible directions can be found in the literature (e.g. Kalman filter positions, augmenting detections with predictions, merging tracklets in post-processing). On the other hand, we have seen that motion models are designed to deal with typical common movements, but that sudden changes in direction, occlusions, and people taking different poses can lead to spurious track creation and deletion. Not relying on time may also benefit other applications, such as when images are taken at long intervals and motion models are essentially useless (e.g. surveillance cameras recording images every few seconds), or when analyzing multiple frames from different viewpoints.

Another option is to combine our method with existing trackers that already account for temporal consistency. For instance, the appearance unmixing paradigm may be useful to resolve the partial occlusions that are skipped by the tracker, or deal with dropped frames in the video. Indeed, our experiments on the TUD dataset show that our method is compatible with and benefits from having accurate object locations, which a traditional tracker may provide, but can also deal with the degenerate cases with occlusions in inaccurate detections.

Furthermore, we modeled appearance as a single latent distribution over binned color values. Adding dynamics to the appearance distribution may improve performance under changing lighting conditions. Other types of features could also be envisioned, e.g. by coupling color information to body regions to distinguish for instance black trousers from a black shirt. Another source of information is the detector confidence, which is currently not exploited (low confidence detections could also result from, say, an unusual pose). It may not even be necessary to use an object detector at all, but just use all detection windows and trust that the background appearance model deals with the false positives. From this point of view, the object detector is just a convenient way to focus computational resources on areas of interest, but at the risk of introducing false negatives.

8. Conclusion

We have described a novel method for identifying multiple objects in a collection of images, which relaxes the one-to-one correspondence between an object's detection in an image and the object's identity. Data association and the related problems of discovering the number of objects, appearance modeling, occlusion handling, and foreground and background segmentation, are treated in a principled way as a joint inference problem in a single graphical model.

Experiments on challenging video sequences of people fighting, and a public benchmark with detections of varying quality, show that the proposed method is particularly effective when the object detection itself is challenging or when temporal modeling is difficult. Such conditions are common in real-life situations, and occur due to occlusion, to fast and irregular motion of the targets, to perspective distortion, to unexpected poses of individuals, or to irregular time intervals between frames in the recording (i.e. frame drops). On sequences with inaccurate detections we show substantially improved identification results in terms of tracked and lost identities, and of identity switches, compared to both global batch-mode and on-line multi-view state-of-the-art trackers, despite the lack of a temporal model.

Future work will extend the method with temporal motion constraints, with dynamics for the appearance model, removing the prior target upper-limit K, and by support for multiple object classes.

Appendix A. Joint distribution details

In this section we provide a more detailed description of the distribution of our model before we describe inference in Appendix C.

A.1. Likelihoods

The likelihood of an observed location x_{ji}^t is

$$p(x_{ji}^t|z_{ji}^t,\{\mu_k^t,\Lambda_k^t\}) = \mathcal{N}\left(x_{ji}^t|\mu_{z_{ji}^t}^t,(\Lambda_{z_{ji}^t}^t)^{-1}\right) \tag{A.1}$$

This means that the likelihood of all observed locations $x_j^t=\{x_{ji}^t\}$ in detection (j,t) is

$$p(x_j^t|z_{ji}^t,\{\mu_k^t,\Lambda_k^t\}) \tag{A.2}$$

$$= \prod_i \mathcal{N}\left(x_{ji}^t|\mu_{z_{ji}^t}^t,(\Lambda_{z_{ji}^t}^t)^{-1}\right) \tag{A.3}$$

$$= \prod_k\prod_i \mathcal{N}(x_{ji}^t|\mu_k^t,(\Lambda_k^t)^{-1})^{\delta(z_{ji}^t,k)} \tag{A.4}$$

However, we have uncertainty about the true locations x_{ji}^t and instead assumed that these are distributed according to Eq. (8) as $x_{ji}^t\sim\mathcal{N}\left(\widetilde{x}_j,\widetilde{\Sigma}_j\right)$. Therefore,

$$\mathbb{E}\left[p(x_j^t|z_j^t,\{\mu_k^t,\Lambda_k^t\})\right] \tag{A.5}$$

$$= \prod_k\prod_i \mathcal{N}(\widetilde{x}_j|\mu_k^t,(\Lambda_k^t)^{-1}+\widetilde{\Sigma}_j)^{\delta(z_{ji}^t,k)} \tag{A.6}$$

$$= \prod_k \mathcal{N}\left(\widetilde{x}_j|\mu_k^t,(\Lambda_k^t)^{-1}+\widetilde{\Sigma}_j\right)^{\sum_i\delta(z_{ji}^t,k)} \tag{A.7}$$

So because of this assumption, we can forget about the exact locations x_{ji}^t and instead take the detection window location $\widetilde{x}_j$. But during inference, which will be discussed in Appendix C, we will need to account for the added uncertainty.

The likelihood of observing words $y_j^t=\{y_{ji}^t\}$ in detection (j,t) is expressed as

$$p(y_{ji}^t|z_{ji}^t,\{\beta_k\}) = \mathrm{Mult}(y_{ji}^t|\beta_{z_{ji}^t}) \tag{A.8}$$

$$= \prod_v (\beta_{z_{ji}^t(v)})^{\delta(y_{ji}^t,v)} \tag{A.9}$$

$$p(y_j^t|z_j^t,\{\beta_k\}) = \prod_i p(y_{ji}^t|z_{ji}^t,\{\beta_k\}) \tag{A.10}$$

$$= \prod_i\prod_v (\beta_{z_{ji}^t(v)})^{\delta(y_{ji}^t,v)} \tag{A.11}$$

$$= \prod_k\prod_v (\beta_{k(v)})^{\sum_i\delta(y_{ji}^t,v)\delta(z_{ji}^t(v),k)} \tag{A.12}$$

Indicator variables $z_j^t=\{z_{ji}^t\}$ are distributed as

$$p(z_{ji}^t|\theta_j^t) = \mathrm{Mult}(z_{ji}^t|\theta_j^t) = \prod_k (\theta_{j(k)}^t)^{\delta(z_{ji}^t,k)} \tag{A.13}$$

$$p(z_j^t|\theta_j^t) = \prod_i p(z_{ji}^t|\theta_j^t) = \prod_i\prod_k (\theta_{j(k)}^t)^{\delta(z_{ji}^t,k)} \tag{A.14}$$

$$= \prod_k (\theta_{j(k)}^t)^{\sum_i\delta(z_{ji}^t,k)} \tag{A.15}$$

A.2. Priors

We use the following conjugate prior distributions

$$p(\theta_j^t|\alpha_0) = \mathrm{Dir}(\theta_j^t|\alpha_0) \tag{A.16}$$

$$= C(\alpha_0)\prod_k (\theta_{j(k)}^t)^{\alpha_{0(k)}-1} \tag{A.17}$$

$$p(\beta_k|\eta_0) = \mathrm{Dir}(\beta_k|\eta_0) \tag{A.18}$$

$$= C(\eta_0)\prod_v (\beta_{k(v)})^{\eta_{0(v)}-1} \tag{A.19}$$

And finally, (μ_k^t,Λ_k^t) are drawn from the Normal–Wishart distribution, which can be written as the product

$$p(\mu_k^t,\Lambda_k^t) = p(\mu_k^t|m_0,\Lambda_k^t,\lambda_0)p(\Lambda_k^t|W_0,\nu_0) \tag{A.20}$$

where

$$p(\Lambda_k^t|W_0,\nu_0) = \mathcal{W}(\Lambda_k^t|W_0,\nu_0) \tag{A.21}$$

$$p(\mu_k^t|m_0,\Lambda_k^t,\lambda_0) = \mathcal{N}(\mu_k^t|m_0,(\lambda_0\Lambda_k^t)^{-1}). \tag{A.22}$$

Appendix B. Spatial uncertainty

We now show how to derive the spatial uncertainty,

$$x_{ji}^t\sim\mathcal{N}\left(\widetilde{x}_j,\widetilde{\Sigma}_j\right) \quad \text{where} \tag{B.1}$$

$$\widetilde{\Sigma}_j = \begin{bmatrix} w^2/12 & 0 \\ 0 & h^2/12 \end{bmatrix}. \tag{B.2}$$

The diagonal components in the covariance matrix $\widetilde{\Sigma}_j$ are obtained as follows: Given a (univariate) variable x uniformly distributed on the range $[-r/2,+r/2]$, i.e. a range of size r with $\mathbb{E}(x)=0$, then the variance of x is

$$var(x) = \mathbb{E}[(x-\mathbb{E}[x])^2] = \mathbb{E}[x^2] \tag{B.3}$$

$$= \int_{-r/2}^{r/2} x^2\frac{1}{r}dx = \frac{1}{3r}x^3\Big|_{-r/2}^{r/2} \tag{B.4}$$

$$= \frac{1}{3r}(r/2)^3 - \frac{1}{3r}(-r/2)^3 = \frac{1}{3\times 8}r^2 + \frac{1}{3\times 8}r^2 \tag{B.5}$$

$$= \frac{r^2}{12} \tag{B.6}$$

Hence, the covariance matrix of the locations within a rectangle (detection window) of width w and height h is

$$\widetilde{\Sigma}_j = \begin{bmatrix} w^2/12 & 0 \\ 0 & h^2/12 \end{bmatrix}. \tag{B.7}$$

Appendix C. Inference

We start this section with a short description of variational inference in Appendix C.1, after which we continue to derive the update equations for our model in Appendix C.2.

C.1. Variational Bayesian inference

For completeness, we summarize here the explanation of approximate inference for directed graphical models using Variational Bayesian (VB) inference, as found in [7], Chapter 10. Assume we wish to estimate the posterior distribution $p(Z|X)$ of some distribution $p(X,Z)$ with parameters Z and observed variables X. If an exact solution is intractable, we can approximate $p(Z|X)$ with a variational distribution $q(Z)$, for which estimating the optimal parameters is easier, and minimize the Kullback–Leibler (KL) divergence between $\mathrm{KL}(q||p)$. This is achieved by maximizing the lower-bound $\mathcal{L}(q)$, since [7]

$$\ln p(X) = \mathcal{L}(q) + \mathrm{KL}(q||p), \quad \text{where} \tag{C.1}$$

$$\mathcal{L}(q) = \int q(Z)\ln\left\{\frac{p(X,Z)}{q(Z)}\right\}dZ \tag{C.2}$$

$$\mathrm{KL}(q||p) = -\int q(Z)\ln\left\{\frac{p(Z|X)}{q(Z)}\right\}dZ. \tag{C.3}$$

Now suppose that distribution $q(Z)$ factorizes into M distributions, each for a different subset of parameters Z_i, such that $Z = (Z_1 \dots Z_M)$, i.e.

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i), \quad \text{then} \tag{C.4}$$

$$\mathcal{L}(q) = \int q_j(Z_j) \ln \widetilde{p}(X, Z_j) dZ_j - \int q_j(Z_j) \ln q_j(Z_j) dZ_j + \text{const} \tag{C.5}$$

where the new distribution $\widetilde{p}(X, Z_j)$ is defined as

$$\widetilde{p}(X, Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const} \tag{C.6}$$

$$\mathbb{E}_{i \neq j}[\ln p(X, Z)] = \int \ln p(X, Z) \prod_{i \neq j} q_i(Z_i) dZ_i \tag{C.7}$$

and $\mathbb{E}_{i \neq j}[\cdot]$ is the expectation with respect to q over all variables Z_i for $i \neq j$.

While we did not assume any particular form for the distributions $q(Z_j)$, we can keep all $Z_{i \neq j}$ fixed and maximize (C.5) if we use the optimal solution [7],

$$\ln q_j^\star(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const} \tag{C.8}$$

Thus each distribution $q_j^\star(Z_j)$ depends on the other distributions $q_i^\star(Z_i)$ required to solve the expectation $\mathbb{E}_{i \neq j}[\ln p(X, Z)]$. This leads to an iterative scheme where we compute the parameters of each distribution $q_j^\star(\cdot)$ in turn, each step improving the lower-bound $\mathcal{L}(q)$.

Typically, if Z_j has a conjugate prior in $p(X, Z)$, the variational posterior $q_j^\star(Z_j)$ will have the form of the corresponding factorized prior in $p(X, Z)$ [7].

Note that in C.2 we will simplify the notation by simply writing $\mathbb{E}_q[\cdot]$ instead of $\mathbb{E}_{i \neq j}[\cdot]$, where the context will make clear which variables are estimated, and thus that all other variables are fixed.

C.2. Update equations for our model

Recall that we know from our observations the word occurrence counts in detection (j, t),

$$\mathbf{N}_{jv}^t = \sum_i \delta(y_{ji}^t, v) \tag{C.9}$$

but throughout the derivations of the update equations, we also use the following *expected word occurrence counts* which we will be able to compute after we have determined $q^\star(z_j^t)$:

$$N_{jkv}^t \triangleq \mathbb{E}_q\left[\sum_i \delta(z_{ji}^t, k)\delta(y_{ji}^t, v)\right] \tag{C.10}$$

$$N_{jk}^t \triangleq \mathbb{E}_q\left[\sum_i \delta(z_{ji}^t, k)\right] \tag{C.11}$$

$$N_k^t \triangleq \mathbb{E}_q\left[\sum_j \sum_i \delta(z_{ji}^t, k)\delta(y_{ji}^t, v)\right] \tag{C.12}$$

$$N_{kv} \triangleq \mathbb{E}_q\left[\sum_t \sum_j \sum_i \delta(z_{ji}^t, k)\delta(y_{ji}^t, v)\right] \tag{C.13}$$

We can now derive the update equations of the parameters of the variational distribution q.

C.3. Appearance and prior updates

Using Eq. (C.8) to find the optimal variational posterior for θ_j^t, we derive

$$\ln q^\star(\theta_j^t) = \mathbb{E}_q[\ln p(\theta_j^t|\alpha_0)] + \mathbb{E}_q[\ln p(z_j^t|\theta_j^t)] \tag{C.14}$$

$$= \mathbb{E}_q[\ln \text{Dir}(\theta_j^t|\alpha_0)] + \mathbb{E}_q[\ln \text{Mult}(z_j^t|\theta_j^t)] \tag{C.15}$$

$$= \sum_k (\alpha_{0(k)} - 1) \ln \theta_{j(k)}^t + \sum_k E_q[\sum_i \delta(z_{ji}^t, k)] \ln \theta_{j(k)}^t + \text{const} \tag{C.16}$$

Note that due to the use of a conjugate prior for θ_j^t, the optimal variational posterior is indeed the conjugate posterior

$$q^\star(\theta_j^t) = \text{Dir}(\theta_j^t|\gamma_j^t) \tag{C.17}$$

with its parameters computed as

$$\gamma_{j(k)}^t = \alpha_{0(k)}^t + \mathbb{E}_q[\sum_i \delta(z_{ji}^t, k)] \tag{C.18}$$

$$= \alpha_{0(k)}^t + N_{jk}^t \tag{C.19}$$

Using the standard properties of the Dirichlet [7], we define

$$\ln \tilde{\theta}_{j(k)}^t \triangleq \mathbb{E}_q[\ln \theta_{j(k)}^t] = \Psi\left(\gamma_{j(k)}^t\right) - \Psi\left(\sum_{k'} \gamma_{j(k')}^t\right). \tag{C.20}$$

which will be used in the association updates. Similarly, we find

$$q^\star(\beta_k) = \text{Dir}(\beta_k|\eta_k) \tag{C.21}$$

$$\eta_{k(v)} = \eta_{0(v)} + \mathbb{E}_q\left[\sum_t \sum_j \sum_i \delta(z_{ji}^t, k)\delta(y_{ji}^t, v)\right] \tag{C.22}$$

$$= \eta_{0(v)} + N_{kv} \tag{C.23}$$

$$\ln \tilde{\beta}_{k(v)} \triangleq \mathbb{E}_q[\ln \beta_{k(v)}] = \Psi\left(\eta_{k(v)}\right) - \Psi\left(\sum_{v'} \eta_{k(vv)}\right) \tag{C.24}$$

C.4. Spatial updates

Due to the use of conjugate priors, we can again find that the variational posterior $q^\star(\mu_k^t, \Lambda_k^t)$ is again a Normal–Wishart distribution,

$$q^\star(\mu_k^t, \Lambda_k^t) = \mathcal{N}(\mu_k^t|m_k^t, (\lambda_k^t \Lambda_k^t)^{-1}) \mathcal{W}(\Lambda_k^t|W_k^t, v_k^t). \tag{C.25}$$

The posterior parameters $\{m_k^t, \lambda_k^t, W_k^t, v_k^t\}$ of object k at time t are computed from the shared prior parameters $\{m_0, \lambda_0, W_0, v_0\}$, the occurrence counts of the features N_k^t and N_{jk}^t, and the mean and covariance $\hat{x}_k^t, S_k^t$ of the assigned feature locations,

$$\hat{x}_k^t = \frac{1}{N_k^t} \sum_j N_{jk}^t \widetilde{x}_j \tag{C.26}$$

$$S_k^t = \frac{1}{N_k^t} \sum_j N_{jk}^t \left[(\widetilde{x}_j - \hat{x}_k^t)(\widetilde{x}_j - \hat{x}_k^t)^\top + \widetilde{\Sigma}_j\right]. \tag{C.27}$$

With these statistics, we compute the posterior parameters as [7],

$$\lambda_k^t = \lambda_0 + N_k^t \tag{C.28}$$

$$m_k^t = \frac{1}{\lambda_k^t}(\lambda_0 m_0 + N_k^t \hat{x}_k^t) \tag{C.29}$$

$$(W_k^t)^{-1} = W_0^{-1} + N_k^t S_k^t + \frac{\lambda_0 N_k^t}{\lambda_0 + N_k^t}(\hat{x}_k^t - m_0)(\hat{x}_k^t - m_0)^\top \tag{C.30}$$

$$v_k^t = v_0 + N_k^t. \tag{C.31}$$

C.5. Association updates

Due to the coupling of the likelihoods of the standard LDA and MoG model at the mixture indicators $z_j^t = \{z_{ji}^t\}$, derivation of $q^\star(z_j^t)$ is less straightforward. From Eq. (C.8) we know that the optimal variational distribution for z_j^t takes the following form,

$$\ln q^\star(z_j^t) = \mathbb{E}_q[\ln p(z_j^t|\theta_j^t)] + \mathbb{E}_q[\ln p(x_j^t|z_j^t, \{\mu_k^t, \Lambda_k^t\})]. + \mathbb{E}_q[\ln p(y_j^t|z_j^t, \{\beta_k\})] + \text{const} \tag{C.32}$$

Let us first define some intermediate results that are needed to express the second (spatial) term in Eq. (C.32). From the standard properties of the Wishart distribution [7] we find

$$\ln \tilde{\Lambda}_k^t \triangleq \mathbb{E}_q[\ln|\Lambda_k^t|] = \sum_{d=1}^{D} \Psi\left(\frac{\nu_k^t + 1 - d}{2}\right) + D\ln 2 + \ln|W_k^t| \tag{C.33}$$

and also

$$\mathbb{E}_q[(x_{ji}^t - \mu_k^t)^\top \Lambda_k^t (x_{ji}^t - \mu_k^t)] = \frac{D}{\lambda_k^t} + \nu_k^t \mathbb{E}\left[(x_{ji}^t - m_k^t)^\top W_k^t (x_{ji}^t - m_k^t)\right] \tag{C.34}$$

Remember from Eq. (A.7) that the positions $x_j^t = \{x_{ji}^t\}$ have spatial uncertainty $x_{ji}^t \sim \mathcal{N}(\widetilde{x}_j, \widetilde{\Sigma}_j)$, which depends on the location and size of window j. Expanding Eq. (C.34), we therefore obtain

$$\mathbb{E}_q[(x_{ji}^t - \mu_k^t)^\top \Lambda_k^t (x_{ji}^t - \mu_k^t)] = \frac{D}{\lambda_k^t} + \nu_k^t \text{trace}\left[W_k^t\left(\widetilde{\Sigma}_j + (\widetilde{x}_j - m_k^t)(\widetilde{x}_j - m_k^t)^\top\right)\right] \tag{C.35}$$

where we make use of the property,

$$\mathbb{E}_{x\sim\mathcal{N}(\mu,\Sigma)}[x^\top A x] = \text{trace}[A(\Sigma + \mu\mu^\top)]. \tag{C.36}$$

for which a proof can be found in Appendix C.7.

Now we can resolve Eq. (C.32),

$$\ln q^\star(z_j^t) = \sum_k N_{jk}^t \mathbb{E}_q[\ln\theta_{j(k)}^t] + \sum_k N_{jk}^t \left(-\frac{D}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}_q[\ln|\Lambda_k^t|] - \frac{1}{2}\mathbb{E}_q[(x_{ji}^t - \mu_k^t)^\top \Lambda_k^t (x_{ji}^t - \mu_k^t)]\right) + \sum_k\sum_v N_{jkv}^t \mathbb{E}_q[\ln\beta_{k(v)}] + \text{const} \tag{C.37}$$

We can simplify this equations by combining the summations,

$$\ln q^\star(z_j^t) = \sum_k\sum_v N_{jkv}^t \ln\rho_{jkv}^t + \text{const} \tag{C.38}$$

where we define

$$\ln\rho_{jkv}^t = \mathbb{E}_q[\ln\theta_{j(k)}^t] + \mathbb{E}_q[\ln\beta_{k(v)}] - \frac{D}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}_q[\ln|\Lambda_k^t|] - \frac{1}{2}\mathbb{E}_q[(x_{ji}^t - \mu_k^t)^\top \Lambda_k^t (x_{ji}^t - \mu_k^t)] \tag{C.39}$$

$$= \ln\tilde{\theta}_{j(k)}^t + \ln\tilde{\beta}_{k(v)} - \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\tilde{\Lambda}_k^t| - \frac{1}{2}\mathbb{E}_q[(x_{ji}^t - \mu_k^t)^\top \Lambda_k^t (x_{ji}^t - \mu_k^t)]. \tag{C.40}$$

From the form of Eq. (C.38) we see that $q^\star(z_j^t)$ takes the form of a multinomial distribution (the constant term is part of the normalization factor). The parameters ϕ of this normalized distribution are therefore expressed in terms of (C.40) as

$$q^\star(z_j^t) = \prod_v\prod_k (\phi_{jv(k)}^t)^{N_{jkv}^t} = \text{Mult}(z_j^t|\phi_{jv}^t) \tag{C.41}$$

$$\phi_{jv(k)}^t = \frac{\rho_{jkv}^t}{\sum_{k'}\rho_{jk'v}^t}. \tag{C.42}$$

Substituting the results from Equations (C.20), (C.24), (C.33) and (C.35) in (C.40), we find the update

$$\phi_{jv(k)}^t = \frac{1}{c} \times \exp\left\{\Psi(\eta_{k(v)}) - \Psi\left(\sum_{v'}\eta_{k(v')}\right)\right\} \exp\{\Psi(\gamma_{j(k)}^t)\}(\widetilde{\Lambda}_k^t)^{1/2} \times \exp\left\{-\frac{D}{2\lambda_k^t} - \frac{\nu_k^t}{2}\text{trace}\left[W_k^t(\widetilde{\Sigma}_j + (\widetilde{x}_j - m_k^t)(\widetilde{x}_j - m_k^t)^\top)\right]\right\} \tag{C.43}$$

where c normalizes the distribution.

From the distribution $q^\star(z)$, we can finally compute the expected word counts,

$$N_{jkv}^t = \mathbf{N}_{jv}^t \phi_{jv(k)}^t \tag{C.44}$$

$$N_{jk}^t = \sum_v \mathbf{N}_{jv}^t \phi_{jv(k)}^t = \sum_v N_{jkv}^t \tag{C.45}$$

$$N_k^t = \sum_j\sum_v \mathbf{N}_{jv}^t \phi_{jv(k)}^t = \sum_j\sum_v N_{jkv}^t \tag{C.46}$$

$$N_{kv} = \sum_t\sum_j \mathbf{N}_{jv}^t \phi_{jv(k)}^t = \sum_t\sum_j N_{jkv}^t \tag{C.47}$$

With the updated counts, we can perform again new appearance, prior and spatial updates. Hence, all the discussed updates are coupled, and will need to be executed iteratively.

C.6. Initialization

We initialize inference by uniformly sampling values for all $\rho_{jv(k)}^t$ from $\mathcal{U}(0,1)$, and obtain normalized $\phi_{jv(k)}^t$ according to Eq. (C.42). Then, we can compute the expected word counts (C.44), and proceed to iteratively apply the updates.

C.7. Derivation of $\mathbb{E}_{x\sim\mathcal{N}(\mu,\Sigma)}[x^\top A x]$

Given that vector x is distributed according to a Normal distribution $x \sim \mathcal{N}(\mu, \Sigma)$, we wish to compute $\mathbb{E}[x^\top A x]$, where A is can be written as $A = B^\top B$.

Let us define $y = Bx$, then

$$\mathbb{E}[x^\top A x] = \mathbb{E}[x^\top B^\top B x] = \mathbb{E}[y^\top y] \tag{C.48}$$

and y is distributed as $y \sim \mathcal{N}(\mu_y, \Sigma_y)$ with

$$\mu_y = B\mu \quad \Sigma_y = B\Sigma B^\top \tag{C.49}$$

From the definition of the covariance we know that

$$\Sigma_y = \mathbb{E}[yy^\top] - \mu_y\mu_y^\top, \tag{C.50}$$

therefore

$$\mathbb{E}[yy^\top] = \Sigma_y + \mu_y\mu_y^\top. \tag{C.51}$$

Now we note that $\mathbb{E}[y^\top y] = \text{trace}[\mathbb{E}[yy^\top]]$. From all of the above, it then follows that

$$\mathbb{E}[x^\top A x] = \text{trace}[\Sigma_y + \mu_y\mu_y^\top] \tag{C.52}$$

$$= \text{trace}[B\Sigma B^\top + B\mu\mu^\top B^\top] \tag{C.53}$$

$$= \text{trace}[B(\Sigma + \mu\mu^\top)B^\top] \tag{C.54}$$

$$= \text{trace}[B^\top B(\Sigma + \mu\mu^\top)] \tag{C.55}$$

$$= \text{trace}[A(\Sigma + \mu\mu^\top)]. \tag{C.56}$$

This result was used in Eq. (C.35) to obtain the association updates for our model.

Appendix D. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cviu.2015.03.012.

References

[1] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2008, pp. 1–8.

[2] A. Andriyenko, S. Roth, K. Schindler, An analytical formulation of global occlusion reasoning for multi-target tracking, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1839–1846.

[3] A. Andriyenko, K. Schindler, Multi-target tracking by continuous energy minimization, in: Proceediongs of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1265–1272.

[4] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh, On smoothing and inference for topic models, in: Proceedings of the UAI, 2009.

[5] H. Ben Shitrit, J. Berclaz, F. Fleuret, P. Fua, Tracking multiple people under global appearance constraints, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 137–144.

[6] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the CLEAR MOT metrics, J. Image Video Process. 1 (2008) 2008.

[7] C.M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, Inc., Secaucus, NJ, USA, 2006.

[8] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[9] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1124–1137.

[10] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.

[11] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. Van Gool, Online multiperson tracking-by-detection from a single, uncalibrated camera, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1820–1833.

[12] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.

[13] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 267–282.

[14] T. Hospedales, J. Li, S. Gong, T. Xiang, Identifying rare and subtle behaviours: a weakly supervised joint topic model, IEEE Trans. Pattern Anal. Mach. Intell. 33 (99) (2011) 2451–2464.

[15] S. Kim, S. Kwak, J. Feyereisl, B. Han, Online multi-target tracking by large margin structured learning, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2012, pp. 98–111.

[16] V. Kolmogorov, R. Zabin, What energy functions can be minimized via graph cuts?, IEEE Trans Pattern Anal. Mach. Intell. 26 (2) (2004) 147–159.

[17] J.F.P. Kooij, G. Englebienne, D.M. Gavrila, A non-parametric hierarchical model to discover behavior dynamics from tracks, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, Berlin Heidelberg, 2012, pp. 270–283.

[18] H.W. Kuhn, The Hungarian method for the assignment problem, Naval Res. Logist. Quart. 2 (1–2) (1955) 83–97.

[19] J. Li, S. Gong, T. Xiang, On-the-fly global activity prediction and anomaly detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2009, pp. 1330–1337.

[20] Y. Li, C. Huang, R. Nevatia, Learning to associate: hybridboosted multi-target tracker for crowded scene, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2953–2960.

[21] M. Liem, D.M. Gavrila, Multi-person localization and track assignment in overlapping camera views, in: Proceedings of the DAGM Symposium on Pattern Recognition, 2011, pp. 173–183.

[22] A. Milan, K. Schindler, S. Roth, Detection-and trajectory-level exclusion in multiple object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3682–3689.

[23] L.L. Presti, S. Sclaroff, M. La Cascia, Object matching in distributed video surveillance systems by LDA-based appearance descriptors, in: International Conference on Image analysis and Processing (ICIAP), Springer, 2009, pp. 547–557.

[24] D. Ramanan, D.A. Forsyth, A. Zisserman, Tracking people by learning their appearance, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 65–81.

[25] B.C. Russell, W.T. Freeman, A.A. Efros, J. Sivic, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, 2006, pp. 1605–1614.

[26] E.B. Sudderth, A. Torralba, W.T. Freeman, A.S. Willsky, Learning hierarchical models of scenes, objects, and parts, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, 2005, pp. 1331–1338.

[27] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical Dirichlet processes, J. Am. Statist. Assoc. 101 (476) (2006) 1566–1581.

[28] X. Wang, E. Grimson, Spatial latent Dirichlet allocation, in: Advances in Neural Information Processing Systems (NIPS), 2008, pp. 1577–1584.

[29] X. Wang, K.T. Ma, G.W. Ng, W.E. Grimson, Trajectory analysis and semantic region modeling using a nonparametric Bayesian model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[30] X. Wang, X. Ma, E. Grimson, Unsupervised activity perception by hierarchical Bayesian models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2007, pp. 1–8.

[31] X. Yan, X. Wu, I. Kakadiaris, S. Shah, To track or to detect? An ensemble framework for optimal selection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 594–607.

[32] B. Yang, R. Nevatia, An online learned CRF model for multi-target tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2034–2041.

[33] B. Yang, R. Nevatia, Online learned discriminative part-based appearance models for multi-human tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 484–498.

[34] B. Zhou, X. Wang, X. Tang, Random field topic model for semantic region analysis in crowded scenes from tracklets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 3441–3448.