Joint Tracking and Event Analysis for Carried Object Detection

Aryana Tavanai
fy06at@leeds.ac.uk
Muralikrishna Sridhar
scms@leeds.ac.uk
Eris Chinellato
e.chinellato@leeds.ac.uk
Anthony G. Cohn
a.g.cohn@leeds.ac.uk
David C. Hogg
d.c.hogg@leeds.ac.uk

School of Computing
University of Leeds
Leeds, UK

Abstract

This paper proposes a novel method for jointly estimating the track of a moving object and the events in which it participates. The method is intended for dealing with generic objects that are hard to localise and track with the performance of current detection algorithms - our focus is on events involving carried objects. The tracks for other objects with which the target object interacts (e.g. the carrying person) are assumed to be given. The method is posed as maximisation of a posterior probability defined over event sequences and temporally-disjoint subsets of the tracklets from an earlier tracking process. The probability function is a Hidden Markov Model coupled with a term that penalises non-smooth tracks and large gaps in the observed data. We evaluate the method using tracklets output by three state of the art trackers on the new created MINDSEYE2015 dataset and demonstrate improved performance.

1 Introduction

This paper investigates whether reasoning about events that objects may participate in can facilitate tracking. For example, in domains where objects are frequently carried, dropped and exchanged, the trajectories of objects and the people carrying them have prototypical spatio-temporal relationships.

In this work, we exploit the knowledge introduced by events in order to improve the outcome of object tracking. We particularly focus on event-based tracking for domains which tend to contain generic objects, for which it is not straightforward to train class specific object detectors. The task of object tracking becomes especially challenging in such domains, as false and missing detections are highly prevalent [4, 5, 14]. This leads to false tracks, and also tracks that are heavily fragmented. We observe this phenomenon when we apply state-of-the-art trackers [1, 12, 14].

Furthermore, the prevalence of false and fragmented tracks makes the goal of incorporating events particularly challenging as event based tracking is a circular problem. For example, events such as *carry* and *drop* represent a unique spatio-temporal behaviour with respect to an object. This characteristic behaviour of events may be used to impose a constraint on the object's track, such as being spatially consistent with respect to a person (*carry*) or a vertical inconsistency (*drop*). However, in order to obtain such events we require tracks. Therefore, this circular problem involves inferring events using reasonable tracks, and then using the events to subsequently improve the tracks. Due to this challenge, event based tracking has been rarely approached [2, 14, 15] in the past (§2).

Further described in §3, given a set of tracklets, the aim is to exploit the spatio-temporal structure of tracks induced by events in order to improve object tracking. More specifically, we exploit learned domain-specific temporal transitions between events in order to target true positive tracklets, which we then use to form meaningful whole tracks. Our approach is illustrated in Fig. 1 and is performed iteratively within an optimisation similar to [11].

Our framework is not constrained to using any particular tracker to build the set of tracklets and, in principle, can be applied to any tracker. To illustrate this point, we have applied the framework to three state-of-the-art trackers [1, 12, 14], and we evaluate them by comparing object tracks, constructed with and without the use of events. We demonstrate that the use of events significantly improves over each of the aforementioned trackers (§4.1).

Our approach is also able to deal with other challenges such as highly occluded objects and different points of view (frontally, laterally and in perspective). This is shown in the experimental validation performed on the newly created MINDSEYE2015 dataset (§4), which contains a large number of events representing the changing relation between objects and people, captured from three different viewpoints.

2 Related Work

The literature on carried object detection is relatively young, but rapidly growing. Earliest works are based on silhouette and motion [3, 8]. More recently, these features have been complemented by taking into account protrusions and the way carried objects modify human appearance [4]. A limitation of the above approaches is that they rely heavily on fitting people to silhouettes and on the availability of visible protrusions, respectively. Other approaches reformulate the carried object detection problem as one of identifying pedestrians with human walking patterns, by searching for shape outliers [10, 13] or by analysing subjects' gait [9], but without localising the object itself. Very few works have tried to improve tracking by including events. Wang *et al*. [15] join pedestrian tracking and event detection into a single optimisation problem, however their events describe human motion with respect to the viewpoint (left, right, away or towards the camera) which are extrinsic rather than intrinsic events, i.e. events are defined with respect to the viewpoint rather than the actors. In [2] event recognition and tracking are not optimised jointly; tracking requires event knowledge, and only running the optimisation for each event class allows to compare the results of event choice. Carried object tracking is performed using spatial consistency between person and object in [14], but the system only considers the *carry* event and is thus able to track an object only when it is being carried.

In this paper we illustrate the benefits of jointly and simultaneously processing event analysis and object tracking, and how we overcome most issues and limitations affecting the works cited above. First of all, our approach does not rely on the fitting of person silhouettes

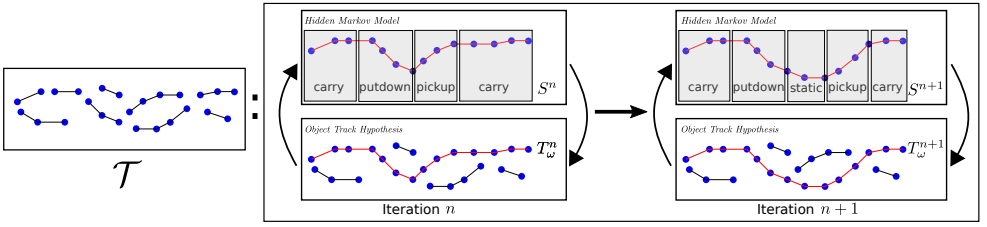


Figure 1: We illustrate two consecutive iterations within the Joint Tracking and Event Analysis optimisation. Given a set of tracklets $\mathcal{T}$ (on the left), at each iteration, a temporally-disjoint subset ω is selected and a contiguous track T_ω is produced by linearly interpolating across any gaps. The Viterbi path S^* of event labels in the HMM is inferred from T_ω (arrow up), leading to an HMM measure (arrow down) and combined with the spatio-temporal factors in Equation 3 to give an overall probability. In the next iteration, a change to the subset ω is made and the overall probability re-computed. In this case, the new configuration is accepted since the probability is increased.

or object protrusions, and we explicitly localise the carried object. Our system is also not dependent on the camera viewpoint, and it considers a large variety of events which are modelled by a standard HMM approach.

3 Formulation

In this section we introduce our Joint Tracking and Event Analysis (JTEA) framework for tracking carried objects. We generalise the definition of a carried object as any particular object that a person has interacted with in the scene, therefore not being limited to only when it is carried. The main novelty of JTEA is the improvement of carried object tracking by incorporating events to enforce spatio-temporal constraints to the tracking solution. Moreover, as a result of improved tracks, events recognition is also improved. We formulate our JTEA approach under one objective function as described below.

3.1 Joint Tracking and Event Analysis

We assume a tracker has produced a set of tracklets $\mathcal{T}$, that provide potential constituents for a single moving object within the target scene. Although there may be more than one, we are only interested in finding a single moving object. We also assume that there are other objects in the scene that interact with our target moving object (typically a person or an aspect of the scene). We refer to these objects as *reference objects*, R, and assume they have already been tracked. The goal is to find the optimal object track consisting of a continuous sequence of tracklets, influenced by spatio-temporal relationships between the target object and the reference object tracks in R.

Each tracklet $\tau \in \mathcal{T}$ is a contiguous sequence of detections, each represented by the centre of the minimal enclosing rectangle. A candidate track T_ω is defined by a subset of tracklets $\omega \subseteq \mathcal{T}$ such that there is no temporal overlap between tracklets in ω (we assume subsets are disjoint in what follows). The track T_ω is a time series of the detections that make up the tracklets of ω, linearly interpolated between the end of one tracklet and the start of the next.

Our objective is to find an optimal set of tracklets $\omega^* \subseteq \mathcal{T}$ and an associated optimal sequence of event states S^* from the set of all possible event sequences $\mathcal{S}$, expressed as:

$$(\omega^*, S^*) = \underset{\omega \subseteq \mathcal{T}, S \in \mathcal{S}}{\operatorname{argmax}} P(\omega, S|R, \mathcal{T}) \quad (1)$$

$$= \underset{\omega \subseteq \mathcal{T}, S \in \mathcal{S}}{\operatorname{argmax}} P(\omega|S, R, \mathcal{T}) P(S|R, \mathcal{T}) \quad (2)$$

In Equation 1 the term $P(\omega, S|R, \mathcal{T})$ evaluates the probability of each hypothesis set of tracklets ω and a sequence of event states S, given reference tracks R and tracklet set $\mathcal{T}$. The conditional probability for ω (the first factor in Equation 2) is defined in Equation 3 and is a product of three parts: the first two penalise non-smooth tracks and large gaps between tracklets, and the third is an event-state dependent Gaussian observation density over position and velocity. Although the spatial (smoothness) term may seem redundant alongside the Gaussian observation density, it is not localised to the current time instant by virtue of its construction with a smoothing function F, and we have found that it improves results. We express $P(\omega|S, R, \mathcal{T})$ as three terms, namely *spatial*, *temporal* and *Gaussian observation*, each defined in Equation 3:

$$P(\omega|S, R, \mathcal{T}) = \underbrace{\left(\prod_{i=1:||T_\omega||} \sigma_1(|T_\omega^i - F(T_\omega)^i|) \right)}_{\text{Spatial}} \underbrace{\left(\sigma_2\left(\frac{\sum_{\tau_j \in \omega} ||\tau_j||}{||T_\omega||} \right) \right)}_{\text{Temporal}} \underbrace{\left(\prod_{i=1:||T_\omega||} \mathcal{N}(\mathbf{x}^i | \mu^{s^i}, \Sigma^{s^i}) \right)}_{\text{Gaussian observation}} \quad (3)$$

The *spatial* and *temporal* terms express the probability of a trajectory from ω, independent of the reference track R and capture standard tracking measures. The *spatial* term measures the degree of spatial association between temporally consecutive detections $T_\omega^i \in T_\omega$. It is calculated by taking the product of probabilities of a generalised logistic function σ_1 based on the absolute euclidean distances between each detection T_ω^i and $F(T_\omega)^i$. σ_1 returns a value of 1 for shorter distances and decreases to 0 for larger distances. F is a smoothing function applied to T_ω and $F(T_\omega)^i$ returns the smoothed corresponding point of T_ω^i. This term penalises the use of outlier tracklets in non-smooth tracks.

The *temporal* term penalises the gaps between the tracklets that make up the track T_ω. We obtain this measure by applying a generalised logistic function σ_2 on the ratio of non interpolated detections from tracklets in ω over the total length of the track T_ω. The larger the ratio, σ_2 returns a value closer to one and vice versa. This measure promotes the use of observed tracklets over interpolated points.

The *Gaussian observation* term is where the events are taken into account. It enforces a prior distribution, based on bilateral relations between object track T_ω and reference tracks R. We calculate these relations using a function L with which we obtain an observation matrix $\mathbf{x}$ i.e. $\mathbf{x} = L(T_\omega, R)$. We therefore calculate the Gaussian observation term as a product of the probabilities obtained from the normal distribution of individual observations $\mathbf{x}^i$ with respect to multiple events E, modelled by a mean μ^{s^i} and a covariance matrix Σ^{s^i} for an event s^i.

For the second term in Equation 2, $P(S|R, \mathcal{T})$, we assume S is independent of R and $\mathcal{T}$ and define as a Markov chain on the state sequence. Thus, this term and the Gaussian observation term in Equation 3 effectively define an HMM, which is then coupled with the smoothness and gap-penalty terms to give the overall probability. This HMM provides a measure of how likely pairs of an object track and reference tracks of possessive entities, conform to a model sequence of event states. Modelling the HMM is further described in the next section.

3.2 Modelling Events

We define our HMM model by a set of discrete events E, an event variable $s_n \in E$ at time n, transition probabilities between events $P_{u|v} = P_{s_n=u|s_{n-1}=v}$, $1 < u, v < |E|$, prior probabilities for the initial event $P_{s_1=u}$ and output probabilities for each event $P_u(\mathbf{x}) = P_{s_n=u}(\mathbf{x})$.

The observation vector $\mathbf{x}$ is composed of the relative position and velocity between the target object and the reference objects R. Thus it has $4 \times |R|$ dimensions. We estimate the parameters of the HMM using maximum likelihood. For this we use a training dataset that is labelled with ground truth for the reference object track, carried object track and the events. Given a set of video sequences, for each event, we obtain an observation matrix where for each observation type (bilateral relation), we model a Gaussian defined by a mean and a full covariance matrix μ_e and Σ_e respectively.

We also create a transition matrix based on the occurrence of an event u following an event v for all frames in the videos. Similarly we learn a prior for the occurrence of an event u. We therefore represent our HMM model as the set of Gaussians for each event, the transition matrix, and the prior.

Thus, to test a hypothesis track T_ω, given a set of reference tracks R, we can construct a new observation matrix similar to above. By applying a Viterbi algorithm using this matrix and the above HMM model, we obtain an HMM measure for the *Gaussian observation* term along with a generated sequence of events for $P(S)$. We further describe the HMM model and the events used in §4.2.

3.3 Optimisation

Illustrated in Fig. 1, our optimisation process is similar to [□], where we apply a set of *moves*, namely *add*, *remove* and *replace*, to construct successive track hypotheses. Given a set of object tracklets $\mathcal{T}$, obtained from any tracker, we initialise our object track T_ω by including only the first and last observed tracklets of $\mathcal{T}$ in our track hypothesis ω and obtain an initial probability using the objective function in Equation 2. Note that if these two tracklets are not suitable and do not belong to the optimal track hypothesis ω^*, they may be removed or replaced in the optimisation. In each iteration of the optimisation, a tracklet $\tau \in \mathcal{T}$ is randomly sampled, weighted by a normalised distribution of tracklet lengths.

A new hypothesis can be constructed in three ways depending on the sampled tracklet τ and the set of tracklets in ω: (i) if $\tau \in \omega$, we construct a new hypothesis by *removing* it from ω; (ii) if $\tau \notin \omega$ and it does not temporally overlap with any other tracklets in ω, we *add* it to ω and (iii) if it does temporally overlap, we *replace* any overlapping tracklets with τ in ω. Based on the *moves* above, at each iteration we construct a new track hypothesis T along with a new probability based on the objective function. If this probability is higher than the previous iteration's probability, we use the new hypothesis as the current best track hypothesis, if not, we continue with the previous best track hypothesis.

By using this hill climbing approach, using a stopping criterion of a large number of iterations, the optimisation terminates and provides the track hypothesis with the highest probability. Although this approach may only reach a local optimum, in practice we have found that in most cases there is only a minor difference from the global optimum.

In this optimisation, events play a significant role as they are constructed from the track hypothesis in each iteration using the HMM, and they affect the suitability of new hypothesis tracks in future iterations through the objective function. To an extent, this solves the circular nature of the problem of joint tracking and event analysis, where both tracking and event

analysis influence and improve each other.

As previously described, any tracker may be used to obtain the set of tracklets $\mathcal{T}$. To further improve our JTEA framework, the next section describes our own tracker, which has been specifically constructed to provide more suitable tracklets for carried object tracking.

3.4 Spatial Consistency Tracker

The goal of the Spatial Consistency Tracker (SCT), an extension of our earlier work [14], is to obtain a set of tracklets $\mathcal{T}$, while maximising the number of true positive tracklets and minimising the false positive ones. To achieve this goal, the SCT tracker takes advantage of relationships based on spatial consistency between the tracklets and reference objects in R. Encoding relationships at this early stage can remove a large number of false positive tracklets from $\mathcal{T}$, the space of forming the object track hypothesis ω in JTEA. This becomes especially important for carried objects as they can vary dramatically in size, shape and colour, leading to a significant rise in false positives in addition to weak and partial detections due to high levels of occlusion. This makes tracking systems prone to false tracks and heavy fragmentation, as evidenced by applying state-of-the-art trackers to these detections.

To better capture true positive tracklets, we extend and generalise our earlier work [14] by only using spatially consistent events (e.g. carry, static), to enforce a strong spatial prior distribution (heatmap) that encodes spatial consistency between a carried object and a reference object. It must be noted that only spatial consistency can capture true positive tracklets, as only they follow a consistent behaviour relative to their interacting entities during a spatially consistent event. This is not true however for spatially sporadic events (e.g. drop, pickup), as in addition to true positives during these events, false positives also follow a sporadic behaviour.

We therefore not only include the spatial relations between a carried object and a person, as done in [14], but also with respect to the scene or potentially any other interacting entity from R. We capture these relations as heatmaps using the same optimisation in [14]. Due to this, the SCT tracker is more suitable for carried objects compared to generic object trackers.

4 Evaluation

Dataset Carried object detection datasets typically include only people walking with or without carrying objects. Our system however, is designed and expected to perform when people interact with the objects in a variety of ways (§4.2). Therefore, we created the MINDSEYE2015 dataset, using a subset of videos from the MINDSEYE Year 2 dataset [1].

MINDSEYE2015 consists of 15 videos (5 recordings captured from 3 viewpoints), each lasting approximately 6000 frames. These videos were converted to a resolution of 360×640 at 20 frames per second. Excluding frames where no event occurs, there are approximately over 45 minutes of events and interactions. The videos are taken from three different viewpoints, illustrated in Fig. 2, which allows a better evaluation of the capabilities of our approach in dealing with object occlusions. The viewpoints offer different types of challenges to a tracker: viewpoint C1 has medium levels of object occlusion and medium levels of scene depth; viewpoint C2 has medium levels of object occlusion (when the object is held in front of the person) and high levels of scene depth; viewpoint C3 has high levels of object occlusion (depending on which side of the person the object is carried) and low levels of scene depth.

[1] http://www.visint.org/datasets

(a) C1 (b) C2 (c) C3

Figure 2: Sample images from the MINDSEYE2015 dataset illustrating the 3 different viewpoints C1, C2 and C3, including a person (cyan bounding box), a carried object (red bounding box) and the event the person is performing using the object (yellow text).

Videos in the dataset show a variety of people interacting with various different objects. In the majority of frames there are at most one person and one object, but there are cases of more or less than one person or object present in the scene. It is also worth noting that, since the dataset was captured outdoors, the movement of trees and cloths on the table, as well as the change in brightness of the video due to clouds and distance of the person to the camera cause challenges for object detectors.

As described in 4.2, we have defined 7 events to allow for a full description of the scene with regards to the state of the carried object, from the start of its appearance to its disappearance. This representation enables us to show the potential of using events in the tracking process. Ground truth for person tracks, carried object tracks and events are fully annotated. This dataset is publicly available with all ground truth annotations, carried object detections, tracklets and final tracks[2]. Relevant code including the carried object detector in [14] can be found on the JTEA webpage [3].

Preprocessing To obtain foreground masks we apply background subtraction for each frame against a single frame containing only the background. We then threshold the subtracted mask using the mask's average brightness. This mostly removes noise from background motion and gives a reliable foreground mask. To obtain people detections we use [6] with the release version of [7]. We then apply the tracker [12] on the person detections to obtain person tracks.

As part of our approach, we initially divide the videos into clips defined by the start and end of each ground truth carried object track. We then apply our state-of-the-art carried object detector [14] on these clips, starting with person regions defined by previously obtained person tracks. The detector is then run on any foreground that is not covered by a person region. After obtaining the set of all detections, we apply three trackers (§4.1) to obtain the set of all tracklets. For each tracklet set, we then apply JTEA (§3.1) to obtain the final tracks of the carried objects. We evaluate our approach in terms of both tracking and event recognition.

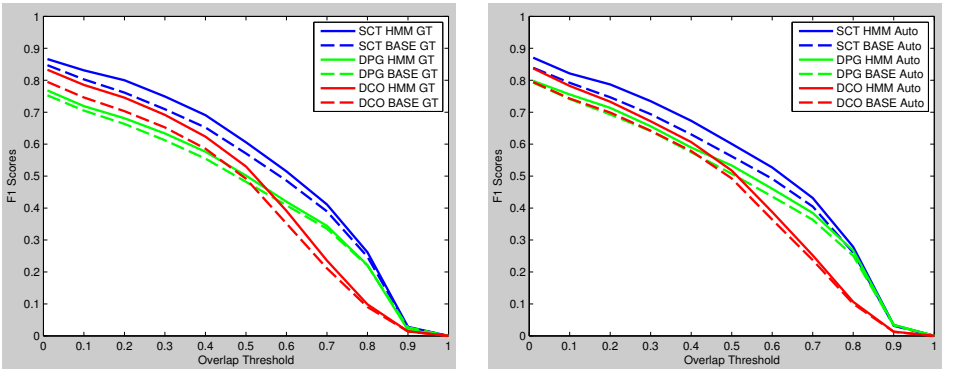
4.1 Track Evaluation

For evaluating the tracking performance we compare the extracted tracks with the ground truth, and calculate the F1 score on a frame by frame basis. We repeat this evaluation for different values of overlap thresholds that defines how much a detection in a track needs to overlap with the ground truth to be considered as a true positive.

We use three trackers: (i) our Spatial Consistency Tracker (SCT) described in §3.4, (ii) an unmodified version of Pirsiavash *et al*. tracker [12] (DPG) and (iii) an unmodified version

[2] http://doi.org/10.5518/9

[3] http://www.engineering.leeds.ac.uk/joint-tracking-and-event-analysis/



(a) (b)

Figure 3: Performance of trackers using detections from: (a) ground truth (GT) and (b) automatic (Auto) person tracks. For each of the trackers, the significant increase in performance as a result of incorporating events (HMM) over the baseline (BASE) can be clearly seen.

of Andriyenko *et al*. tracker [1] (DCO). Fig. 3 illustrates the performance of two types of approaches for each of the aforementioned trackers, (i) the full Joint Tracking and Event Analysis approach (HMM) and (ii) the approach without event analysis (BASE) which only uses the *spatial* and *temporal* terms in equation 3.

The results are presented in terms of an average F1 score across all videos calculated on a frame by frame basis. We run two sets of experiments for tracking, using carried object detections obtained from ground truth person tracks (GT) and automatic person tracks (Auto). This is to illustrate that JTEA is not heavily dependent on highly accurate person tracks. Our automatic person tracks obtain an average F1 score of 0.79, 0.73 and 0.58 with 50%, 60% and 70% overlap thresholds respectively.

From the results in Fig. 3 it can be observed that: (1) the performance of tracking is significantly improved when influenced by events; (2) our SCT tracker outperforms other trackers for the purpose of carried object tracking. It must be noted that even a 5% improvement corresponds to approximately 3000 more true positives due to the large number of frames in the dataset. It is also worth highlighting how the performance of the system does not drop quickly for higher values of overlap, showing the potential of our carried object detector [14].

For a better comparison with related work, in Table 1, we provide a summary of performance indexes computed at 20% overlap, a value typically employed in the literature for carried object tracking [4, 5]. Since there is no major difference between the GT and Auto results, also showing the robustness of our approach to the quality of person tracks, we only provide detailed information for the GT evaluation.

	F1	Precision	Recall	Accuracy	Run Time
SCT HMM	**0.80**	**0.81**	**0.79**	**0.67**	< 25 min
SCT BASE	0.76	0.77	0.75	0.61	< 25 min
DCO HMM	0.75	0.76	0.73	0.59	< 5 min
DCO BASE	0.70	0.72	0.69	0.54	< 5 min
DPG HMM	0.68	0.69	0.67	0.52	< 1 min
DPG BASE	0.66	0.67	0.66	0.50	< 1 min

Table 1: Performance indexes of carried object tracks at 20% overlap using GT person tracks.

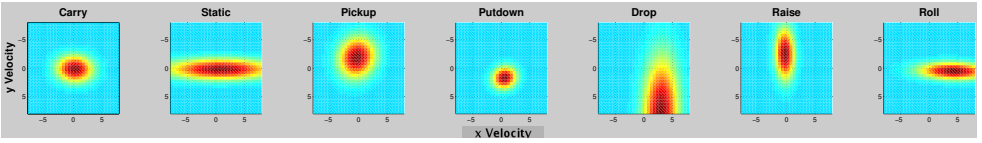


Figure 4: Gaussians in the HMM model for each event, based on the object-person velocities relation only, in the x and y dimensions. The direction and magnitude of events is captured relative to the $(0,0)$ coordinate, defining absolute consistency. For example, *raise* is mostly horizontally consistent, however, it is vertically inconsistent in an upward direction.

Again, it can be easily verified that, for all the different indexes, the SCT tracker is consistently better, and that event detection always improves the tracking performance. Note that the run times in Table 1 are for the trackers only, and that JTEA ran an additional 5 minutes for each tracker. If the SCT tracker is optimised however, it is expected to run within a much shorter time. All given times are calculated using a single core on an Intel Xeon E5-2665 Processor @2.40GHz.

4.2 Event Evaluation

We initially describe the modelling of events within an HMM, followed by quantitative results.

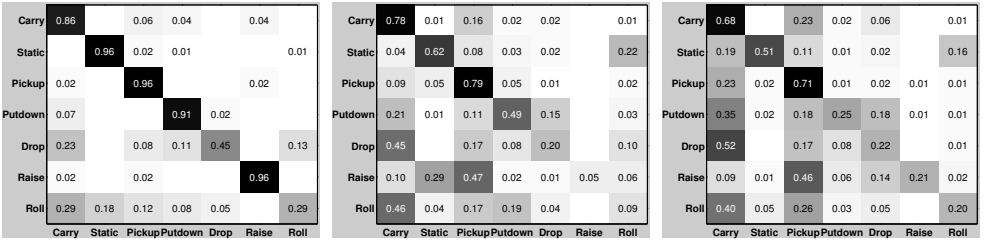
HMM Modelling We use seven types of events in which an object participates in, relative to two types of reference entities, the person and scene. These events are *Carry*, *Static* (object is stationary), *Pickup*, *Putdown*, *Drop*, *Raise* and *Roll* (object is moving on the ground).

To train the HMM model and learn the above events, we take as input ground truth object and person tracks along with a scene bounding box which covers the entire image frame. As described in section 3.2, we construct an observation matrix for each event based on position and velocity, in the x and y dimensions relative to two reference entities, person and scene. For each event, we obtain a mean and covariance matrix which we use in our final HMM model along with prior probabilities of events and a transition matrix constructed from event ground truth. Sample Gaussian models are illustrated in Fig. 4. The above HMM training is performed within five folds (one for each recording in the dataset). To test an object track from a video, we use the HMM from the fold that the video was not trained on.

Results In addition to the evaluations illustrating improvement in tracking using events in our JTEA framework, we also present an evaluation on the event analysis aspect of JTEA. We present the results of event classification obtained with three different approaches: (a) events obtained from ground truth object and person tracks in folds against ground truth events; (b) events obtained from final tracks generated by the SCT HMM GT tracker; (c) events obtained from final tracks given by SCT BASE GT.

In Fig. 5 (a), the confusion matrix shows that the HMM approach to modelling events for carried objects is suitable for the problem. Confusion matrix (b) shows that HMM based event classification notably improves over the baseline (c) as a direct result of using improved tracks using JTEA. Since the improvement of tracks was due to the use of events, by using our JTEA framework, the tracks and the events jointly influence and improve each other.

Some of the misclassifications in the confusion matrices in Figures 5 (b) and (c) are caused by incorrect HMM predictions due to the similarity in behaviour of events. For example, the *carry* and *roll* events have a similar horizontal motion and only differ in terms of their possessive entities, person and scene respectively. However, another reason for



(a) HMM Train Test (b) SCT HMM GT (c) SCT BASE GT

Figure 5: Event classification confusion matrices for three different approaches using a frame by frame evaluation.

misclassifications is due to a frame by frame evaluation. For example, *drop* being misclassified as *carry*, or *raise* being misclassified as *pickup*, is due to the *drop* event occurring directly after a *carry* event or a *raise* after a *pickup*. Therefore, it is extremely challenging to predict when an event exactly finishes or starts, allowing these misclassifications to occur due to the next event being predicted slightly earlier or later.

To further clarify our event evaluation, we report on the number of correct event classifications in Table 2. In this table, the ground truth based classification in (a) allows for consistently better results in all classes, but the HMM event recognition also improves substantially over the baseline's performance (64.2% vs. 53.9% correct classifications respectively).

	Carry	Static	Pickup	Putdown	Drop	Raise	Roll	Total	Total %
Sum GT Frames	10787	36578	1402	1521	83	530	259	51160	100
HMM Train Test	9559	35337	1340	1379	37	508	74	48234	94.3
SCT HMM	8382	22537	1108	742	17	28	24	32838	64.2
SCT BASE	7398	18630	994	375	18	113	53	27581	53.9

Table 2: Total number of true positive event detections based on a frame by frame evaluation.

5 Conclusions

Despite the increasing efforts put by the computer vision community into tracking objects and people from videos, few approaches have investigated the benefit of performing tracking jointly with event analysis. In this work we have presented a novel approach to the problem of tracking objects which people may interact with in various ways. In our JTEA optimisation procedure an optimal subset of tracklets are chosen that define the carried object track. The search procedure leading to this object track is influenced by events, modelled via an HMM which uses the object track to infer events.

We have shown that the inclusion of event analysis in this circular optimisation process significantly improves the performance of three different trackers. Moreover, our SCT tracker outperforms the other two trackers, due to its robustness to false positives as a results of using spatial consistency between the object and the possessive entity. We have also shown that while improving tracking, event classifications are simultaneously obtained and improved.

While our approach is reasonably robust to multiple people and objects being present in the scene, if the number of people and object further increases, it introduces challenging complexities. We are thus currently working on extending our approach to include multi-person, multi-object events such as giving, exchanging or replacing objects.

The financial support of DARPA Mind's Eye project VIGIL (W911NF-10-C-0083) and the EU projects RACE (FP7-ICT-287752) and STRANDS (FP7-ICT-600623) is gratefully acknowledged.

References

[1] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR*, pages 1926–1933. IEEE, 2012.

[2] A. Barbu, N. Siddharth, A. Michaux, and J. M. Siskind. Simultaneous object detection, tracking, and event recognition. *Advances in Cognitive Systems*, 2:203–20, 2012.

[3] C. BenAbdelkader and L. S. Davis. Detection of people carrying objects: A motion-based recognition approach. In *IEEE International Conference on Automatic Face and Gesture Recognition FGR*, pages 378–383. IEEE Computer Society, 2002.

[4] D. Damen and D. Hogg. Detecting carried objects from sequences of walking pedestrians. *PAMI*, 34(6):1056–1067, 2012.

[5] R. Dondera, V. I. Morariu, and L. S. Davis. Learning to detect carried objects with minimal supervision. In *CVPRW*, 2013.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and Ramanan D. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[7] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[8] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis. Backpack: Detection of people carrying objects using silhouettes. *Computer Vision and Image Understanding*, 81(3): 385 – 397, 2001.

[9] J. Lamar-León, R. A. Baryolo, E. B. García Reyes, and R. González-Díaz. Gait-based carried object detection using persistent homology. In *Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 8827, pages 836–843, 2014.

[10] H. Nanda, C. Benabdelkedar, and L. Davis. Modelling pedestrian shapes for outlier detection: a neural net based approach. In *IEEE Intelligent Vehicles Symposium*, pages 428–433, June 2003.

[11] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for multi-target tracking. *Transactions on Automatic Control,*, 54(3):481–497, March 2009.

[12] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, pages 1201–1208, 2011.

[13] D. Tao, X. Li, S. J. Maybank, and X. Wu. Human carrying status in visual surveillance. In *CVPR*, pages 1670–1677. IEEE Computer Society, 2006.

[14] A. Tavanai, M. Sridhar, F. Gu, A. G. Cohn, and D. C. Hogg. Carried object detection and tracking using geometric shape models and spatio-temporal consistency. In *Computer Vision Systems*, volume 7963 of *LNCS*, pages 223–233. Springer, 2013.

[15] R. Wang and T. Huang. A framework of joint object tracking and event detection. *Pattern Analysis and Applications*, 7(4), 2004. ISSN 1433-7541.