Recognizing Linked Events: Searching the Space of Feasible Explanations

Dima Damen, David Hogg
School of Computing, University of Leeds
Leeds, LS2 9JT, UK
{dima,dch}@comp.leeds.ac.uk

Abstract

The ambiguity inherent in a localized analysis of events from video can be resolved by exploiting constraints between events and examining only feasible global explanations. We show how jointly recognizing and linking events can be formulated as labeling of a Bayesian network. The framework can be extended to multiple linking layers, expressing explanations as compositional hierarchies. The best global explanation is the Maximum a Posteriori (MAP) solution over a set of feasible explanations. The search space is sampled using Reversible Jump Markov Chain Monte Carlo (RJMCMC). We propose a set of general move types that is extensible to multiple layers of linkage, and use simulated annealing to find the MAP solution given all observations. We provide experimental results for a challenging two-layer linkage problem, demonstrating the ability to recognise and link drop and pick events of bicycles in a rack over five days.

1. Introduction

The visual analysis of events is often ambiguous when performed locally in isolation from other events. A global analysis will generally provide a more reliable solution, exploiting constraints that exist between the different things happening during a given period. We propose a general framework for exploiting such constraints.

The term 'event recognition' refers to mapping an observation into previously modeled event types. Assuming independence from surrounding events, each observation is normally assessed separately, and the event type that best explains the observation is chosen as the recognized event.

Linking events is the process of grouping related events to represent high-level explanations. Often events are related if they involve the same agent or the same object. Global constraints such as arity and temporal ordering govern the linking process. For example, linking the event of a person entering a room to the departure event of the same person provides a high-level explanation about the complete act and its duration. A one-to-one correspondence (arity) constraint is expected and the first event must occur before the second. A feasible explanation is one that does not violate these constraints.

Event recognition and linkage could be performed separately where the event is first recognized for each observation, and the linkage can be decided next. In this paper, we propose simultaneously (i.e. jointly) recognizing and linking events into complete explanations. We apply joint event recognition and linkage to the *Bicycles* problem, first introduced in [4]. The complexity of this problem demonstrates the generality and capabilities of the framework. We refer to the act of leaving the bicycle in the rack as a 'drop', and the act of retrieving the bicycle as a 'pick'. The task is to correctly associate people to the bicycle they have dropped or picked, and to link picks to earlier drops. Two types of detections are considered; the first is of people entering and leaving the rack area, and the second is of changes within the racks that indicate the appearance and disappearance of bicycles, and are referred to as 'bicycle clusters', as each may contain multiple bicycle detections.

Ambiguities in the recognition process increase with occlusion when multiple individuals approach the racks. We refer to these time intervals, during which one or more people are simultaneously inside the rack area, as "activity units" [5]. Figure 1 illustrates an example of an activity unit by highlighting the people and the bicycle clusters. Each

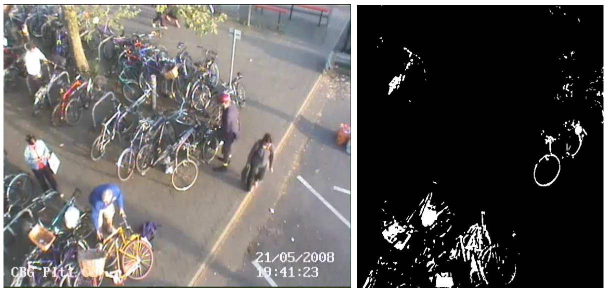

Figure 1. An example of an activity unit showing 5 individuals (left) and several bicycle clusters (right).

activity unit is an event recognition and linkage problem. The linking is constrained so each person is linked to one bicycle cluster at most. This emerges from the natural constraint that a person cannot drop/pick more than one bicycle per visit to the racks. We further link drops to subsequent picks forming a 'higher-level' linkage problem. Each drop can be connected to one pick at most from a later activity unit, and vice versa.

Searching the space of feasible explanations is of exponential complexity. We need a method to enumerate these explanations and assess their posterior probabilities given the observations. To avoid confusion, events that are observed directly from a single detection are called 'atomic events', while high-level explanations are referred to as 'compound events' as they arise from linking other events. We propose a novel framework and argue it can be used whenever,

- The task is to recognize and link related events.
- These linkages can be represented as a hierarchy of (compound) events.
- The labeling of each atomic event can be assessed given an associated observation.
- Links between events are scored, favoring some links over others, and are governed by natural constraints.

Related work is reviewed next, and Section 3 details the method. Section 3.1 explains how a dynamic Bayesian network can be built to model the posterior dependencies. Section 3.2 reviews Markov chain Monte Carlo (MCMC) sampling, and is followed by explaining how reversible moves can traverse the space of feasible explanations in Section 3.3. The selected features and the collected dataset for the *Bicycles* problem are explained in Section 4.1. The results (Section 4.2) demonstrate that maximizing the joint posterior using this proposed framework improves the accuracy over separately recognizing and then linking the events.

2. Related Work

Explaining and linking observations by proposing global feasible explanations and assessing those explanations has been previously applied to several domains. Multi-target tracking in radar surveillance was first tackled by Reid [13] in a Bayesian framework. At each scan, the radar detects noise and target measurements. The problem is to simultaneously associate target measurements into trajectories and discard noisy measurements. Reid searched the space of explanations using the Multiple-Hypotheses Tree (MHT) algorithm, where alternative explanations are explored within a tree structure. Oh *et. al.* [11] use an MCMC approach to sample from the solution space and find the Maximum a Posteriori (MAP) explanation. This work demonstrated the remarkable performance of MCMC over MHT.

Visual tracking resembles radar tracking, as broken trajectories, tracklets and noisy observations have to be connected into complete trajectories. Traditionally, observations are associated by considering a couple of frames. A recent trend towards global solutions, despite the combinatorial complexity, uses approaches such as Bayesian network inference [8], structural EM [17] and linear programming [14]. MCMC finds an approximate solution and has been increasingly employed in visual tracking of pedestrians [1, 15, 16] but also for ants and bees [9]. Smith [15] tracks an unknown number of objects using RJMCMC. A derived work by Yu *et. al.* [16] combines segmentation along with tracking. They model both spatial and temporal moves (extending those of Smith), and search the space of possible explanations within a sliding window. One of the earliest similar problems in visual tracking was introduced by Huang and Russell [7], as part of 'Roadwatch' for tracking cars across wide-area traffic scenes. They assign each car seen upstream to its corresponding observation downstream, allowing for on-ramp and off-ramp observations. Their solution uses MHT, thus it cannot scale to tracking cars between more than two cameras due to the growing complexity. An MCMC sampling approach is proposed for a scalable solution [12].

Similar reasoning can be used to recognize and link events. Gong and Xiang learn the links between events using Dynamic Multi-linked HMMs [5]. They learn causal and temporal relationships from videos of loading and unloading planes. Their work assumes all parallel events can be dependent and can not link events with temporal gaps or enforce global constraints. Chan *et. al.* argue that recognizing and linking events provide the most likely events along with the best track fragment linkage [3]. Applied to recognizing plane re-fueling event sequences, their approach is confined to brute force search as a proof of concept. Our previous work searches the space of feasible explanations for linking dropping and picking bicycles using MHT [4], where the branch with the minimum cost represents the best explanation.

In this paper, we propose a novel framework for jointly recognizing and linking related events. Unlike [5], we focus on causal relationships allowing events to be linked across temporal gaps. Our framework assumes a natural hierarchy of events is known, and partitions the observations into plausible explanations governed by related constraints. We finally use the power of RJMCMC [6] (successfully applied in other domains [11, 15]) to sample the posterior distribution. We re-formulate the *Bicycles* problem as two-layers of event recognition and linkage, and present results that show RJMCMC with simulated annealing can better search the space when compared to greedy and MHT searches.

3. The Method

For a chosen scene domain, we suppose the composition of events forms a hierarchy. The base of the hierarchy is a set of *atomic events* that are detected directly. Higher-levels are *compound events* composed by linking a pair of simpler events (atomic or compound), providing a higher level explanation.

Figure 2 illustrates two examples of the hierarchy of events for the *Bicycles* problem. The hierarchy shows two atomic event types: people (x) and bicycle clusters (y), and two layers of linkage. The first layer links people to bicycle clusters. The link (z) can explain a drop or a pick compound event (shown in brackets), then two such linking nodes are combined into a higher-level link, that explains the drop-pick (dp) compound event.

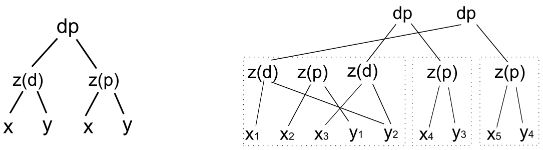

Figure 2. The basic unit for an explanation of the *Bicycles* problem (left) and a sample feasible explanation (right) for 5 people (x) and 4 bicycle clusters (y). Dotted frames surround activity units.

To explain our method, we first detail how a Bayesian network can be built for a sequence of detections based on a given event hierarchy. The complete set of labelings of the Bayesian network corresponds to the set of explanations. Though the Bayesian network is completely general and can in principle be used to discover optimal explanations, we need a tractable way to search through the set of feasible explanations for the MAP solution. We search the space of explanations using MCMC with simulated annealing. The last part of this section introduces general move types that can traverse the space of event linkages.

3.1. The Posterior Probability

We start by transforming the set of atomic events into a single Bayesian network that represents all possible explanations. We first present a simple example for recognizing and linking a pair of atomic events within a single layer of linkage. Figure 3 (left) shows a Bayesian network with three observations; o_x, o_y and s, where s is the score of linking events x and y. Three hidden random variables, (x, y, z), explain the first and the second event types, and whether the two events are linked, respectively. The joint probability is factorized so the compound event is dependent on its constituent events.

For the *Bicycles* problem, suppose we have observed n people and m bicycle cluster events, then Figure 3 (right) shows a plate representation linking each x event to all possible y events according to the domain's event hierarchy.

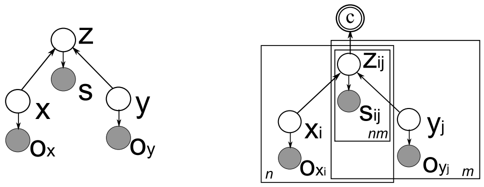

Figure 3. Directed graph linking two events (left) and a plate representation for multiple events (right).

Figure 4 is an unrolled example for $n = 3$ and $m = 2$. The different kinds of nodes in the Bayesian network are labeled on the left hand side. Each detection is represented by an observed Random Variable (RV) connected to a hidden RV. The x atomic events represent tracked people and can be labeled as dropping (e_1), picking (e_2) or passing through (e_3). Each bicycle cluster is represented by a y atomic event and can be labeled as dropped (g_1), picked (g_2) or noise (g_3). Each pair of x and y detections parents a linking node z, that can be labeled by a drop (d), pick (p) or be unlinked (f). A 'd' state, for example, indicates the person dropped a bicycle into the associated cluster. Although the labels may seem partly redundant, they will enable us to combine evidence from observations associated with each event in a consistent fashion. The linking nodes are governed by natural constraints, represented by the deterministic node c. In the *Bicycles* problem, for example, a maximum of one linking node relating the same person can be labeled as a drop or a pick within the explanation. Figure 5 shows a labeled Bayesian network corresponding to the first activity unit in the sample explanation of Figure 2.

We aim to find the MAP explanation $\omega^\star$ (a labeling of all

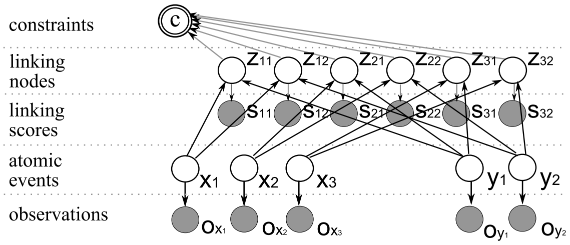

Figure 4. An unrolled Bayesian network for multiple events

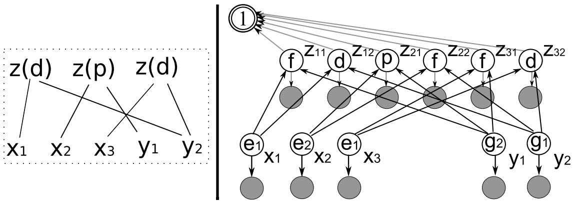

Figure 5. A sample explanation (left) and its corresponding labeling of the Bayesian network (right). The deterministic function evaluates to 1 for feasible explanations only.

hidden RVs) given all observed RVs Y where

$$\omega^\star = \arg\max_\omega p(\omega|Y) \tag{1}$$

For the graph in Figure 4, the posterior can be re-arranged as

$$p(\omega|Y) = \frac{1}{\mathcal{Z}} \prod_i p(x_i|o_{x_i}) \prod_j p(y_j|o_{y_j}) \prod_{ij} p(z_{ij}|x_i, y_j, s_{ij}) p(c|\{z_k\}) \tag{2}$$

where $\mathcal{Z}$ is the normalizing factor that need not be evaluated when searching for the maximum. $p(c|\{z_k\})$ is a deterministic function that evaluates the labels of all z linking nodes, and equals 1 if the explanation is feasible.

Unfortunately, the number of linking nodes in the constructed Bayesian network increases exponentially with the number of atomic events, while the number of feasible links increases only linearly. The product $\prod_i p(z_i|x, y, s)$, from the posterior in (2), can be replaced by a proportional expression that is independent of all links labeled f as follows (We abbreviate $p(z_i|x, y, s)$ into $p(z_i|o)$ in the derivation).

$$\begin{aligned}
\prod_i \quad p(z_i|o) &= \prod_{i:z_i=f} p(z_i = f|o) \prod_{i:z_i=t} p(z_i = t|o) && (3)\\
&= \prod_{i:z_i=f} p(z_i = f|o) \prod_{i:z_i=t} p(z_i = t|o) \frac{\prod_{i:z_i=t} p(z_i=f|o)}{\prod_{i:z_i=t} p(z_i=f|o)} &&\\
&= \prod_i p(z_i = f|o) \prod_{i:z_i=t} \frac{p(z_i=t|o)}{p(z_i=f|o)} && (4)\\
&\propto \prod_{i:z_i=t} \frac{p(z_i=t|o)}{p(z_i=f|o)} && (5)
\end{aligned}$$

After presenting the Bayesian network for the first layer, we present the complete Bayesian network for the *Bicycles* problem. Figure 6 represents this two-layered linkage problem for $n = m = 3$. Two activity units (dotted frames) are shown in the unrolled example. Notice that we only hypothesize and mutually constrain links between people and bicycle clusters within the same activity unit, thereby greatly reducing the number of possible explanations. For the second layer, the linking node v connects z nodes from different activity units, and can represent a drop-pick compound event (dp) or be unlinked (f). The linking score assesses the likelihood of linking a drop to a pick event. An additional random variable z_0 represents unobserved events. Some drops remain unlinked indicating the bicycle is still within the racks, and some picks are related to drops that occurred before the observation period. The posterior probability can be retrieved from the graphical model, where different explanations imply different labelings. The posterior at both linking layers is rewritten according to Equation 5 to be independent of false links.

This section has shown how a Bayesian network can be constructed for two levels of event linking. The same method of construction could be used for any binary hierarchy of atomic and compound events, given a different set of labels and constraints that arise from the domain.

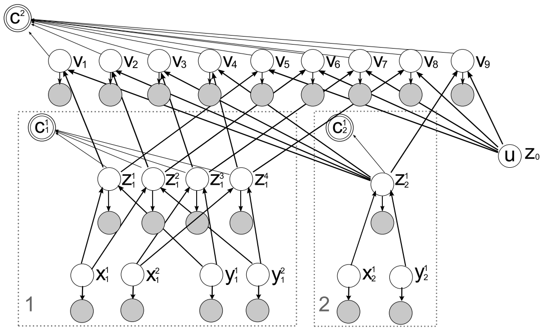

Figure 6. An unrolled Bayesian network for the *Bicycle* Problem showing 2 activity units. Detected people (x) and bicycle clusters (y) are linked within activity units to explain drops and picks. Events are linked in a second layer to explain drop-picks. Explanations at each layer are constrained by deterministic RVs c^1, c^2.

3.2. MCMC

Instead of exhaustively searching the space, MCMC samples the posterior distribution using a Markov chain. The set of possible states in the Markov chain Ω is the set of all feasible explanations, and a conditional proposal distribution $q(\omega, \omega')$ defines the probability of proposing state ω' given the current state is ω. After a state is proposed using q, the move to that state is made with the probability $\alpha(\omega, \omega')$ known as the 'acceptance probability'. A thorough review of MCMC techniques can be found in [2]. We use the Metropolis-Hastings algorithm and define the acceptance probability α as proposed by Green's Reversible Jump MCMC (RJMCMC) [6], where the proposal distribution is split into two steps: j_m for selecting a move type and g_m for selecting a specific move within that type. Green's formulation allows introducing a pair of reversible moves instead of self-reversible moves only, maintaining the detailed balance for convergence.

For finding the MAP solution, adding simulated annealing is in principle a better alternative [2], although previous related work has not used this [1, 9, 11, 15, 16]. MCMC is a sampling technique that is not designed to search for the global maximum. Adding annealing is a minor modification where the Markov chain is non-homogeneous and its invariant distribution φ at each step i in the chain depends on a 'temperature' T that is decreased according to a 'cooling schedule' $\varphi(\omega) = \pi(\omega)^{\frac{1}{T_i}}$.

3.3. Designing Markov Chain Moves

When using RJMCMC to traverse the space of feasible explanations, a different explanation is proposed at each step along the Markov chain based on the current one. For discrete search spaces, multiple types of moves are needed to traverse the search space [6]. We designed 4 move types to traverse the search space (Figure 7). These connect or disconnect a link, change one of the linked events or switch two links. It should be noted that this is not the minimal set of move types. A change move for example can be constructed from a disconnect move followed by a connect move. Disconnecting would decrease the posterior probability significantly, which makes it a less probable move along the chain. Accordingly, change and switch move types enable efficient search of the space and faster convergence. Other complex changes can be constructed from a sequence of these moves.

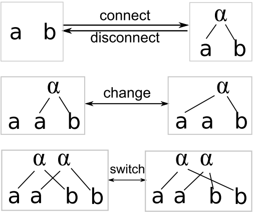

Figure 7. Four moves are proposed to link events, break links, change linked events and switch linkages.

The *Bicycles* example uses the designed 4 move types for each layer. In the initial explanation ω_0, all people are passing through and all detected clusters are noise. This is a valid explanation, though unlikely to be the best. At each step of the Markov chain, a move is applied to the current explanation. Figure 8 shows a sequence of moves applied successively. Each applied move creates a new feasible explanation ω', and can change multiple labels in the Bayesian network. Moves of type 'change drop', for example, change the states of two hidden RVs of type v: ($dp \rightarrow f, f \rightarrow dp$).

We now discuss how we propose a move at each step of the Markov chain $q(\omega, \omega')$. RJMCMC splits proposing a new explanation into two steps: choosing the move type j_m then choosing a specific move g_m. Randomly choosing a move type does not efficiently search the space of explanations. We thus estimate the number of distinct moves of each type that can be applied to the current explanation. For example, the number of possible 'disconnect' moves in the first layer equals the number of dropping and picking people in the current explanation. These counts are used as weights in choosing the move type. Weighting increases the acceptance rate ρ_{accept} and speeds convergence as will be shown in the experiments. The acceptance rate ρ_{accept} is the ratio of the number of accepted moves to the length of the Markov chain.

Next, a specific move of that type is chosen and applied to the current explanation. This 'within-type' choice can also be performed uniformly at random. Alternatively, we can design a customized 'within-type' proposal distribution for each proposed move type. These are application-specific and depend on the expected ambiguities in the observations. We use a distance measure for each move type that weights the preference for choosing moves. For example, the 'connect' move type in the first layer prefers connecting people to bicycle clusters without alternative links. Assume $B(x_i)$ yields the set of clusters that could be connected to person x_i, while $T(y_j)$ yields the set of people that could be connected to cluster y_j, then the distance measure for this move type $\delta_{connect}$ is defined in Equation 6.

$$\delta_{connect}(x_i) = \sum_{y_j \in B(x_i)} \frac{1}{|T(y_j)|} \tag{6}$$

We do not explain the proposed distance measures for the other move types due to space limitation. These are domain specific and their choices do not affect the framework.

4. Experiments and Results

Three aspects of the framework are evaluated. We investigate the advantage of jointly recognizing and linking events versus performing each task alone. We also compare three search techniques for finding the MAP solution: MHT, MCMC and MCMC with the addition of simulated annealing. Then, the MAP solution is compared with ground-truth revealing the ability of the complete framework to explain all observations. We first discuss how the visual features were obtained and introduce the dataset.

4.1. Features and Dataset

In the *Bicycles* scenario two types of detections are identified from a CCTV camera mounted high above the ground: people trajectories (x) and bicycle clusters (y). Trajectories were retrieved by an off-the-shelf background subtraction tracker [10]. Changes to the bicycle rack before the person approaches it and after departing are grouped into connected components representing bicycle clusters. Four observations and linking scores are required: $p(x|o_x)$, $p(y|o_y)$, $p(z|s_z)$ and $p(v|s_v)$ (see Figure 6). Supervised training is used to estimate Gaussian class conditional densities for each likelihood.

$p(x|o_x)$ assesses whether the person is dropping, picking or passing through by comparing the blob size before entering and after exiting the racks. An increase in the blob size signifies a pick and vice versa. Noise or broken trajectories will produce poor assessments.

$p(y|o_y)$ is measured by comparing the number of pixels representing new and removed edges. Assuming the background is relatively free of edges, a significant increase in edges within the changed pixels indicates a dropped bicycle, and vice versa. The remaining clusters are expected to be heterogeneous or noise clusters.

$p(z|s_z)$ assesses the linkage of a person to a bicycle cluster by measuring the maximum degree of overlap between

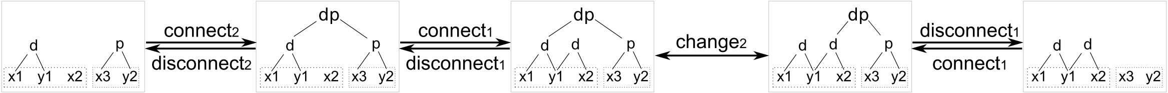

Figure 8. A sequence of {connect drop-pick → connect drop → change drop → disconnect pick} moves was applied. The last move affects both layers as disconnecting a pick cancels the drop-pick linked to that pick. The subscript next to the move type indicates the layer at which the move is applied.

the bounding box of the cluster and the bounding boxes of the foreground regions representing the person across the whole trajectory.

The pixel matches between the dropped and picked clusters is used to compute $p(v|s_v)$. It assumes bicycles do not change their shape or position between being dropped and picked. Figure 9 shows how these matches are established. The ratio of the intersection of the two areas to the minimum area is used to estimate Gaussian conditional densities for correct and incorrect links. We use this new estimate of the bicycle's bounding box to refine $p(z|s_z)$.

The dataset consists of 7 sequences collected from two sites (1-5: first site, 6-7: second site). Sequences 1-3 are those used in our previous work [4]. Sequences 6-7 are recorded by a CCTV camera outside a busy UK train station, and are more challenging with a much greater level of activity and uncertainty (Figure 10). The rack area was manually delimited with a polygon. Table 1 summarizes statistics of these sequences. Priors and conditional probabilities were estimated from the first sequence and the corresponding hand-generated ground truth. These were kept constant for all other sequences across both sites (Table 2). Supervised training of the likelihoods was also performed using the first sequence and fixed for all the sequences, as all the features are designed to be scale and viewpoint independent.

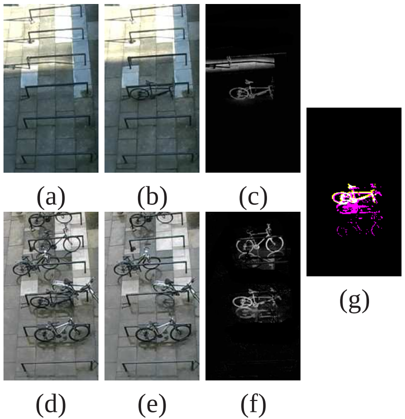

Figure 9. Two images of the racks (a) and (b) are compared to reveal changes (c) representing a dropped bicycle, and a noise cluster due to lighting changes. Later, two consecutive reference images (d) and (e) are also compared to reveal two picked bicycles (f). By matching dropped (yellow) and picked (pink) clusters (g), white pixels signify the match.

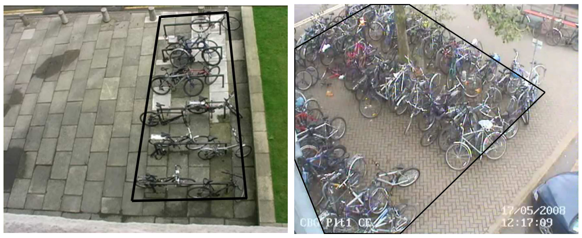

Figure 10. The two sites of the *Bicycles* dataset. Manually labeled polygons delimit the rack area

	Dataset Sequences						
	1	2	3	4	5	6	7
Duration	1h	1h	11h	12h	12h	15h	15h
$\|X\|$	58	27	128	126	137	112	197
$\|Y\|$	59	25	72	175	128	206	1847
|Drops|	24	11	20	20	14	28	39
|Picks|	20	12	19	20	13	17	41
|Drop-Picks|	20	11	18	20	13	14	22
avg(exp/x)	21.7	8.3	19.6	3.2	1.7	10.21	63.4
max(exp/x)	76	24	83	83	50	56	197

Table 1. Dataset statistics; $|X|$: number of detected people, $|Y|$: number of detected bicycle clusters, exp/x: number of different explanations involving each person, and gives a measure of the dataset's inherent ambiguity.

$p(x = e_1) = p(x = e_2) =$	0.495
$p(x = e_3) =$	0.01
$p(z = d\|x = e_1, y = g_1) =$	0.5
$p(z = f\|x = e_1, y = g_1) =$	0.5
$p(z = p\|x = e_2, y = g_2) =$	0.5
$p(z = f\|x = e_2, y = g_2) =$	0.5
$p(v = dp\|z_1 = d, z_2 = p) =$	0.4
$p(v = dp\|z_1 = d, z_2 = u) =$	0.2
$p(v = dp\|z_1 = p, z_2 = u) =$	0.05

Table 2. Estimated priors and conditional probabilities.

4.2. Results

The framework proposes a set of move types and weighted choices of these move types to search the space. We first compare convergence using the minimal set of move types (connect and disconnect moves only) to that using the full set. For the 7 sequences, the mean of ρ_{accept}, over 100 Markov chains, increased by a factor of between 1.9 and 7.4 when incorporating the switch and change moves. This is because both move types enable larger jumps within the search space. Next, we compare weighted versus uniform choice of moves. Figure 11 shows the performance of one MCMC chain (3^{rd} sequence) under different choices of proposal distributions. ρ_{accept} increases from 0.2 for uniformly selected move types to 0.4 for weighed choices, and

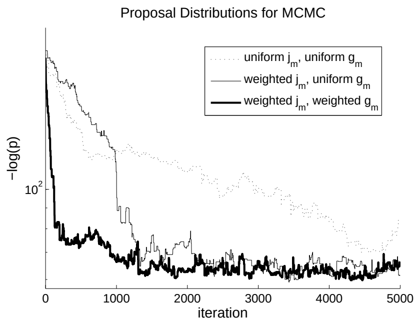

Figure 11. Convergence under various uniform and weighted move-type proposal distribution (j_m) and ‘within-move’ proposal distribution (g_m) using MCMC.

convergence is significantly faster.

We compare MCMC alone with adding annealing using both exponential and linear cooling schedules showing two chains of each case (Figure 12). The temperature was reduced from 4 to 0.01 along all annealing chains.

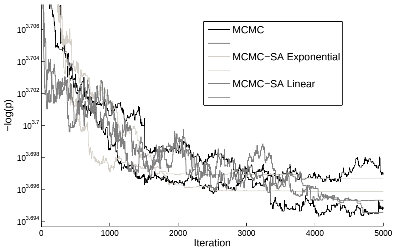

Figure 12. From the 4^{th} sequence, two runs of standard MCMC, two runs of exponential annealing and two runs of linear annealing are compared. One linear annealing and one standard run achieved the best performance.

Table 3 compares the negative log of the MAP solution across the different techniques for all the recorded sequences. Each run consists of 10 parallel and independent chains ($n_{mc} = 5000$), where the MAP solution is the maximum of the MAP solutions across chains. We ran each 40 times and recorded the mean and standard deviation of the MAP. The table reveals that adding annealing enables finding a higher or equal posterior (lower -log(p)) for all 7 sequences. Linear cooling was used for annealing. The table shows the advantage of jointly sampling the space of event recognition and linkage over performing each task separately. The baseline greedy approach maximizes each observation locally and then selects the best link based on the linking scores iteratively, keeping the solution feasible, until the posterior can no longer be increased.

The MAP solution is then compared to a manually obtained partial ground truth. The ground truth labels each person with the type of event accomplished, and records drop-pick pairs. The accuracy is defined as the ratio of correctly labeled events to the overall number of tracked people. It was noticed that the MAP solution might not result in the highest-possible accuracy. This could result from an incorrect modeling of the posterior and the priors, or noise in the features selected. Table 4 compares the accuracy values for the MAP solutions presented in Table 3. It is expected that the accuracies for sequences (6-7) are lower due to the increase in clutter. The 7^{th} sequence suffers from frequent abrupt lighting changes that result in bicycle clusters being poorly detected. Figure 13 gives some examples of recognized and linked drop and pick events across the dataset.

	Greedy	**MHT** **k=50**	**MHT** **k=500**	**MCMC**		**MCMCM-SA**	
				μ	σ	μ	σ
1	102.25	58.78	57.86	57.90	0.11	57.86	0.00
2	23.54	4.64	4.64	4.64	0.00	4.64	0.00
3	609.66	493.18	468.80	429.30	3.23	423.98	2.36
4	6272.69	6149.95	6144.30	6079.88	3.43	6078.40	3.23
5	5034.46	4998.39	4975.82	4943.71	3.59	4939.33	1.87
6	860.37	812.96	812.96	814.71	1.69	811.50	2.36
7	934.36	608.92	-	451.92	9.29	433.50	7.76

Table 3. $-\log(p)$ compared across greedy, MHT, 40 runs of MCMC and 40 runs of MCMC with simulated annealing. The result was not available for the last MHT search (k=500) due to our implementation running out of memory.

	Greedy	**MHT** **k=50**	**MHT** **k=500**	**MCMC**		**MCMC-SA**	
				μ	σ	μ	σ
1	72.41	91.38	91.38	88.36	1.09	87.46	1.79
2	85.19	100.00	100.00	100.00	0.00	100.00	0.00
3	58.59	84.38	84.38	87.68	0.89	83.36	1.65
4	73.81	74.60	75.40	83.93	1.09	83.15	1.31
5	89.05	82.48	88.32	91.90	0.79	92.65	0.90
6	66.07	60.71	60.71	68.53	1.68	70.98	1.04
7	45.69	44.67	-	47.28	1.18	47.61	0.88

Table 4. The accuracy results (%) for the MAP solutions.

Even though all the results presented above utilize the data in a batch mode, an online version of the solution has been developed. This runs a shorter chain at the end of each activity unit, and finds the best explanation for all the observations up to the current time stamp. The MAP solution initializes the Markov chain for the next activity unit.

5. Conclusion and Future Work

This paper proposes a novel framework for jointly recognizing and linking visually ambiguous events. The approach combines observations along with linkage and global constraints in one probabilistic graphical model. We propose a set of reversible moves to traverse the search space using RJMCMC. Adding annealing and weighted proposal distributions assists in finding the MAP solution. The framework can in principle be extended to multiple layers of linkage. We have evaluated the approach on the *Bicycles* problem for a challenging dataset. The same approach could be applied to other domains for linking relates events with a wide temporal gap. We are currently evaluating the method to recognize and link people entering

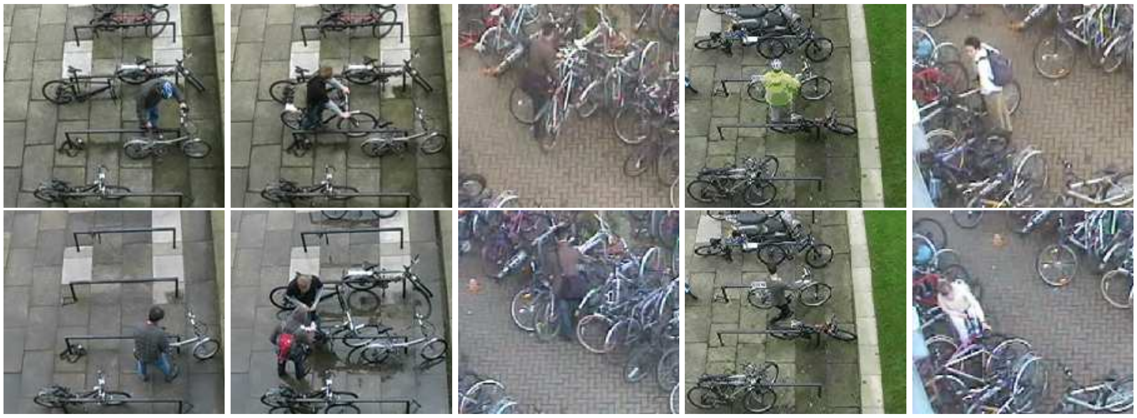

Figure 13. Five examples of connected events. The first four are correctly connected. The fourth column represents a simulated theft. The fifth example shows an incorrect connection. Recall that no clothing color comparison is performed. Individuals are connected by linking the person to a cluster and correctly linking dropped to picked bicycle clusters.

a building to those departing.

Acknowledgements

We would like to thank Mark Everingham for valuable discussions and assisting in formulating the posterior probability, and National Express East Anglia for providing the dataset off the second site.

References

[1] V. Ablavsky, A. Thangali, and S. Sclaroff. Layered graphical models for tracking partially-occluded objects. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2008.

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[3] M. T. Chan, A. Hoogs, R. Bhotika, A. Perera, J. Schmiederer, and G. Doretto. Joint recognition of complex events and track matching. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1615–1622, 2006.

[4] D. Damen and D. Hogg. Associating people dropping off and picking up objects. In *Proc. British Machine Vision Conference (BMVC)*, 2007.

[5] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *Proc. International Conference on Computer Vision (ICCV)*, 2003.

[6] P. Green. Trans-dimensional Markov chain Monte Carlo. In P. Green, N. Lid Hjort, and S. Richardson, editors, *Highly structured stochastic systems*. Oxford University Press, Oxford, 2003.

[7] T. Huang and S. Russell. Object identification: A Bayesian analysis with application to traffic surveillance. *Artificial Intelligence*, 103(1-2):77–93, 1998.

[8] P. Jorge, J. Marques, and A. Abrantes. On-line tracking groups of pedestrians with Bayesian networks. In *Proc. Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2004.

[9] Z. Khan, T. Balch, and F. Dellaert. MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(12):1960–1972, 2006.

[10] D. Magee. Tracking multiple vehicles using foreground, background and motion models. In *Proc. ECCV Workshop on Statistical Methods in Video Processing*, pages 7–12, 2002.

[11] S. Oh, S. Russell, and S. Sastry. Markov chain Monte Carlo data association for general multiple-target tracking problems. In *Decision and Control, (CDC)*, volume 1, pages 735–742, 2004.

[12] H. Pasula, S. Russell, M. Ostland, and Y. Ritov. Tracking many objects with many sensors. In *Proc. International Joint Conferences on Aritificial Intelligence (IJCAI)*, pages 1160–1171, 1999.

[13] D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979.

[14] K. Shafique, L. Mun Wai, and N. Haering. A rank constrained continuous formulation of multi-frame multi-target tracking problem. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2008.

[15] K. Smith. *Bayesian Methods for Visual Multi-object Tracking with Applications to Human Activity Recognition*. PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), 2007.

[16] Q. Yu, G. Medioni, and I. Cohen. Multiple target tracking using spatio-temporal Markov chain Monte Carlo data association. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[17] W. Zajdel and B. J. A. Krose. A sequential Bayesian algorithm for surveillance with nonoverlapping cameras. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 19(8):977–996, 2005.