Temporal Structure Models for Event Recognition

John Greenall
jhs1jpg@leeds.ac.uk
David C. Hogg
D.C.Hogg@leeds.ac.uk
Anthony G. Cohn
A.G.Cohn@leeds.ac.uk

School of Computing
University of Leeds
Leeds, UK

Throughout this work we talk about about recognising events within a *scenario*. Our method is applicable in any domain where there exists some repeated temporal structure. We introduce a novel method for efficiently combining the response from independent event detectors with a tree structured MRF prior over inter-event timings. The model is designed to cope in situations where events of different types can potentially overlap and are not strictly ordered. In our optimization, the temporal midpoints and durations of event instances in a previously unseen sequence are the latent variables to be optimized. The absence of any free parameters in our model makes it generic and easily applicable to other domains. We evaluate our method on a large real world dataset, implementing a state of the art activity classification system based on local spatiotemporal features as independent event detectors.

We assume that we have already trained a probabilistic event detector for each event class, and that these detectors will be run in parallel to evaluate every possible midpoint and duration over the discrete temporal domain, $\mathscr{T} = \{1,..,T\}$. Adopting a Bayesian perspective, the output of the independent event detectors can be treated as the likelihood of the relevant chunk of the observed video sequence, X, arising as a result of the given event scenario, which we term $p(X|C,Y,\theta)$, where θ is the set of parameters that define our model and C, Y are as described in Figure 3. Thus $p(X|c_i,t_i,d_i,\theta)$ is known for all possible $(t_i,d_i) \in \mathscr{T}^2$.

We can then concentrate on optimizing the posterior

$$p(Y|X,C,\theta) \propto p(X|C,Y,\theta)p(Y|C,\theta) \qquad (1)$$

The distribution $p(Y|C,\theta)$ denotes the prior probability of a scenario given its event set and the model. By analogy to Pictoral Structure Models [2] from the object detection literature, we define our prior to be a tree-structured Markov Random Field.

The first component of our prior which must be learnt is the set of pairwise probability distributions $p((t_i,d_i),(t_j,d_j)|c_i,c_j,\theta)$, which we define to be probability distributions on the time differences between instances of each possible pair of event classes appearing in the same sequence. These are modelled with Gaussian Mixture Models (GMMs). We follow the method of [3], which gives full Bayesian treatment to the training of GMMs.

The second component of the temporal prior which must be learnt is the edge set of the MRF. As we require our prior to be tree-structured, and the variables within Y are discrete, the MAP solution to Equation 1 may be obtained directly through the max-sum Belief Propagation (BP) algorithm [1]. Simplifying the prior to a tree structure will result in the loss of some information, so we need to ensure the edges we retain are the most informative. BP is a message passing algorithm and in this context, we believe it is reasonable to assume the most useful messages will be those of lowest entropy. Therefore to determine the structure of the tree, we evaluate the informativeness of each pairwise connection with the following equation:

$$I(i,j) = H[p(X|(t_i,d_i),c_i,\theta)\sum_{t_j} p(t_i,t_j|c_i,c_j,\theta)] \qquad (2)$$

which is equivalent to the entropy of the message that would be passed from t_i to t_j if BP were initiated at that node.

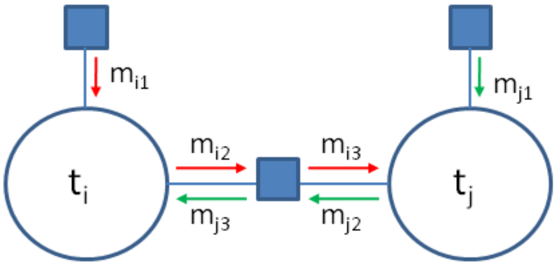

Figure 2: : Depiction of the message passing scenario between two temporal midpoint random variables. We define the usefulness of the link between i and j to be inversely proportional to the lower of the two entropies $H[m_{j3}]$,$H[m_{i3}]$.

We demonstrate the effectiveness of our method with evaluation on challenging dataset of aircraft servicing activities consisting of ~40 hours video split across 37 turnarounds. We evaluate naive detectors,'noT', versus 4 variants of our model:'RGM' has randomly-initialized MRF structure and single Gaussian for pairwise probabilities, 'DG' has dynamically allocated structure and single Gaussian for pairwise probabilities, 'DGM' is our full model with dynamic structure and GMM probabilities, 'PGM' has structure determined by same criteria as in [2] with GMM probabilities. Figures cited are Average Precision obtained with leave-one-out testing, where we require a Relative Overlap > 30% with ground truth.

Detector Type	noT	RGM	DG	DGM	PGM
Average Precision	73	77	78	**82**	**82**

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[2] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. ISSN 0920-5691.

[3] C.E. Rasmussen. The infinite Gaussian mixture model. *Advances in neural information processing systems*, 12:554–560, 2000.

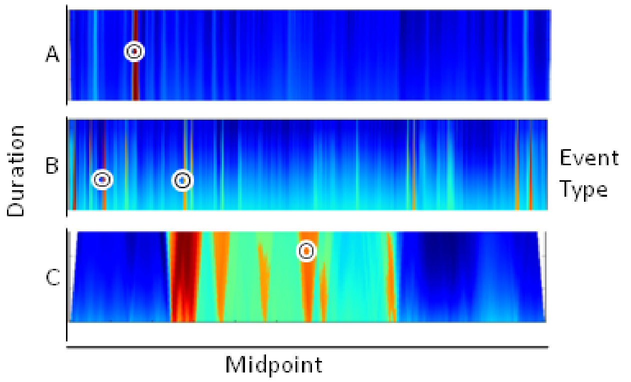

Figure 1: Example response from 3 independent probabilistic event detectors. The detectors give a response for each possible interval in a sequence, with colour indicating the probability (red being high). We show four event instances localized on these likelihood streams.

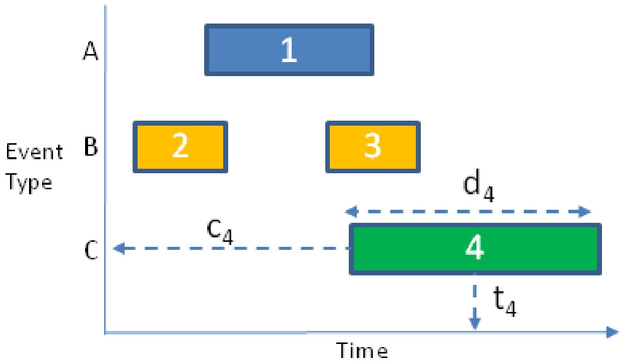

Figure 3: Illustration of sequence scenario corresponding to Figure 1. The scenario is fully specified by event class vector $C = (c_1,...,c_4)$ and temporal information $Y = ((t_1,d_1),...,(t_4,d_4))$.