Delta-Dual Hierarchical Dirichlet Processes: A pragmatic abnormal behaviour detector

Tom S.F. Haines, Tao Xiang
School of Electrical Engineering and Computer Science
Queen Mary, University of London
{thaines,txiang}@eecs.qmul.ac.uk

Abstract

In the security domain a key problem is identifying rare behaviours of interest. Training examples for these behaviours may or may not exist, and if they do exist there will be few examples, quite probably one. We present a novel weakly supervised algorithm that can detect behaviours that either have never before been seen or for which there are few examples. Global context is modelled, allowing the detection of abnormal behaviours that in isolation appear normal. Pragmatic aspects are considered, such that no parameter tuning is required and real time performance is achieved.

1. Introduction

A desirable capability with automated surveillance is the detection of rare behaviours that are of interest to users. Rare behaviours that is interesting includes activities ranging from shoplifting through to terrorism, and everything in between. Human behaviour is complex, and typically subtle, especially when observed via CCTV. This subtlety can take two forms - in the first instance an actor behaving abnormally will often do so at the same time many regular actors continue normally, e.g. a driving offence committed by one driver at a busy intersection. For criminal offences, such as shoplifting, they will often conceal their behaviour, and attempt to look normal. In the second instance behaviour that appears normal in isolation may actually be abnormal given context, e.g. a car running a red light, or a person in an overcoat when others are wearing shorts. Consequently, the abnormal signal may be tiny, often no stronger than noise, and some behaviours are only detectable using a global model with context.

Many existing methods are unsupervised [11, 9, 7, 16, 18] - a single class model is learned using normal behaviour and the statistical outliers marked as abnormal. This strategy offers a practical solution to the issues raised when modelling imbalanced class distributions, and the handling of unseen classes. However, such methods are ineffective in detecting subtle anomalies because they are often indistinguishable in feature space from regular behaviour, due to the preponderance of shared normal behaviour. Furthermore, as outlier detectors, they are not able to *categorise* different types of rare behaviour.

Human knowledge can be exploited to address this problem, in the form of supervised learning. Ideally a fully annotated training data set with all anomalies of interested labelled both spatially and temporally is provided. However, this will often incur a prohibitive manual labelling cost, and introduce the inconsistencies of human interpretation. In practise as few as one example from each anomaly class will be available, with weak labels, e.g. the approximate range of video frames that contain the anomaly, but not where or precisely when it happens. In addition, since anomalies are, by definition, rare, one cannot assume that examples from all classes are available during model training.

To this end *Delta-Dual Hierarchical Dirichlet Processes* (dDHDP) is proposed[1]. dDHDP is a probabilistic topic model designed for jointly learning both normal and abnormal behaviour using weakly supervised training examples. One shot learning [5] is supported. Key to the model structure is the acknowledgement that normal behaviour does not necessarily stop because something abnormal is occurring - abnormal example are modelled as a mixture of both normal and abnormal behaviour, such that the combination of features unique to the abnormality are discovered. It is this mixture model that allows weak supervision, as the readily available examples of normal behaviour allows the unusual behaviour to be separated. Context is included by clustering behaviour temporally, which models which behaviours occur together, and, implicitly, which do not. Despite its apparent complexity no free parameters exist to be tuned and the detection of abnormalities in newly presented data proceed in real time.

[1] The source code can be obtained from http://thaines.com.

1.1. Related Work

Early behavioural models are predominantly based on dynamic Bayesian networks (DBN) [18, 19]. By modelling the dynamics of behaviour explicitly a DBN is sensitive to anomalies caused by temporal order violations. However, such models are sensitive to noise and input errors, have poor tractability and do not scale. Suck weaknesses have motivated many recent approaches to use topic models [7, 8, 12, 13, 15, 16, 17].

Key to a topic model is the bag-of-words assumption, where no relationships between features are considered - noticeably both temporal and spatial information is discarded. The advantages are considerable however - tractability, scalability and robustness are all obtained. Typically a topic model is defined over a corpora of documents, where each document contains many words - the model then discovers topics which are shared among all documents but in different ratios for each specific document. For a video sequence the documents are short *clips*, on the order of seconds, and the words discrete video features, whilst the inferred topics are the behaviours on display in each clip. Standard topic models include *latent Dirichlet allocation* (LDA) [3] and *hierarchical Dirichlet processes* (HDP) [12].

Since the semantic meaning of a behaviour is context dependent various hierarchical topic models have been proposed to compensate for the loss of temporal ordering. Hospedales et al. introduced a Markov clustering topic model [7] to model the temporal order of clips explicitly. Wang et al. [16] formulated an extension to HDP, *dual hierarchical Dirichlet processes* (DHDP), to co-cluster both words into topics and documents into contextual phases, with the number of clusters at both layers determined automatically. A contextual phase is in effect a prior over the expected behaviours in a document, allowing the detection of rare combinations. The presented dDHDP model is closely related to DHDP, in terms of context modelling and being free of parameter tuning. However, all these topic models are used for unsupervised learning only, and are incapable of supervised learning of any kind.

Supervised learning with topic models was first explored by Andrzejewski et al. [1]. In particular *delta latent Dirichlet allocation* (dLDA) was proposed for the purpose of statistical debugging, a situation with large quantities of shared normal behaviour and small quantities of abnormal behaviour, which are mixed in with the regular behaviour. Two major weaknesses exist for dLDA - it only has one kind of abnormality, and it has no capacity to handle new documents provided after learning the initial model. Li et al. [8] fixed these in their extension, multi-class dLDA[2], which is also used to detect rare and subtle behaviour. However it is limited by not modelling context, leaving an entire class of abnormality undetectable, and is only capable of assigning one behaviour to a clip[3]. An additional weakness is that the number of normal behaviours is not learnt, when this parameter can have a catastrophic effect on performance if set badly. dDHDP resolves all of these issues.

To summarise, we contribute a novel model capable of learning subtle behaviours given little training data, that maintains the ability to detect previously unseen behaviours as outliers. It can recognise abnormal behaviours that previous approaches, such as dLDA, cannot; has real time performance and no parameters to tune. The implementation is non-trivial, and a simple yet novel extension to a Dirichlet process that allows for pre-existing topics in unknown ratios is presented, as is a Bayesian method of estimating a multinomial given sample count vectors that are sparse.

2. Methodology

2.1. Graphical Model

The graphical model for dDHDP is given in figure 1, using plate notation. Additionally, a special representation of Dirichlet processes is used - instead of just indicating the base measure a solid plate containing the entities which are drawn from the base measure is linked to the random variable representing the Dirichlet process (DP). The base measure itself is then given by the arrows leading to the entities within the base measure plate, as they define how to draw its contents. Similarly, a draw from the DP is represented by an arrow going to a dashed plate, which contains the specific instances from the solid plate that have been drawn. The main value of this notation can be seen with regard to Q, the DP over clusters, where it allows the two parts of a cluster to be represented separately, rather than as a single random variable. It also avoids fragmenting the visualisation into multiple parts.

To give an overview the model has many documents, $d \in \mathcal{D}$, that contain many words, $n \in \mathcal{N}_d$, and each word is assigned either a regular topic, $H_t^{\mathcal{R}}$, or an abnormal topic, $H_t^{\mathcal{A}}$, as represented by variable t_{dn}. Each document has a distribution over these topics, $G_d^{\mathcal{D}}$, and shares its topics with other documents, such that they cluster. The distribution over topics has a prior, S_d, which depends on the documents assignment to a cluster, as represented by c_d - this causes a second clustering, of documents, to occur.

A complete explanation of the graphical model is now given, starting from the top of the figure and working down. Top left is plate $H_t^{\mathcal{A}}$, the set of abnormal topics, and to the right is the regular topics, $H_t^{\mathcal{R}}$. Each topic is represented by

[2] Multi-class is for convenience dropped for the rest of the document and dLDA used to refer to the Li variant rather than the original.

[3] Whilst the chances of rare behaviours occurring simultaneously may be small there is no reason to assume that only abnormalities will be learned - regular behaviours can also be learnt, to collect statistics or indicate that a specific behaviour is *not* of interest to the user.

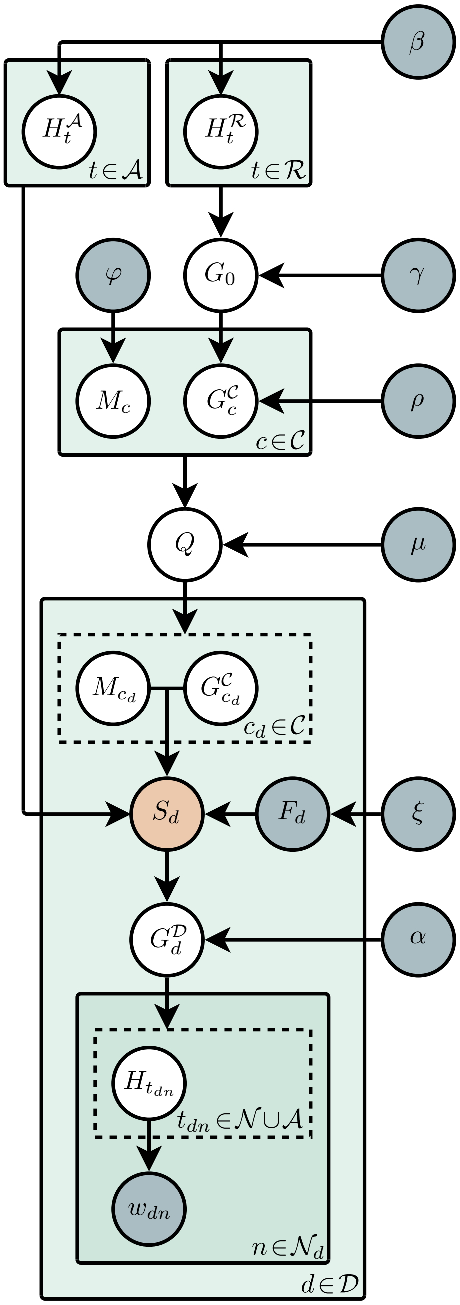

Figure 1. dDHDP graphical model. See subsection 2.1 for details.

a multinomial over words, drawn from a Dirichlet distribution, β. Whilst β is shaded to indicate that it is known, in practice it is learnt using a maximum likelihood technique. The number of abnormal topics is known - they correspond to the user provided training examples, but the number of regular ones is not, hence the DP G_0 with concentration γ. All concentration parameters have been marked as known, but actually have weak Gamma priors and are learnt alongside everything else. Plate $c \in \mathcal{C}$ contains the draws from Q, the DP over clusters, and so represents the set of clusters in the model. Each cluster consists of two parts - a DP over regular topics with base measure G_0 and concentration ρ, and a multinomial over behaviours, M_c. Behaviours are the set of abnormal topics in combination with a single entry for all regular topics. The behaviour multinomials, M_c, have the Dirichlet prior φ, which is also learnt. DP Q with concentration μ handles the variable cluster count.

Plate $d \in \mathcal{D}$ represents the many documents (video clips) that are drawn from the model. Each document has a single cluster, drawn from Q, represented by the dashed box $c_d \in \mathcal{C}$. Documents also have a set of flags, F_d, with values $\{0, 1\}$ indicating which behaviours exist within the document - it is known during training, but learnt for novel documents. Each flag is drawn from a Bernoulli distribution, the parameters of which are represented by vector ξ - this point is obviously only relevant when observing new documents, where ξ has to be provided as a parameter as there is no means to learn it. The flag for regular behaviour is set for all documents, which means that the relevant entry in ξ must be set to 1.

S_d is the key construct that combines abnormal and regular behaviour into a single distribution, from which multinomials over words, H, may be drawn. It draws a behaviour from M_{c_d}, but only considers behaviours that are marked as existing in the flags, F_d. If the behaviour is regular it then draws a topic multinomial from the clusters DP over regular behaviours, $G_{c_d}^{\mathcal{C}}$, otherwise it has drawn the multinomial for the corresponding abnormal behaviour. This may be expressed as

$$H|F, M, G^{\mathcal{C}}, H^{\mathcal{A}} \sim \hat{M}(0)G^{\mathcal{C}} + \sum_{t\in\mathcal{A}} \hat{M}(t)\delta_{H_t^{\mathcal{A}}} \quad (1)$$

where $\hat{M}$ is defined as

$$\hat{M}(i) = \frac{F(i)M(i)}{\sum_{i\in 0\cup\mathcal{A}} F(i)M(i)} \quad (2)$$

$\hat{M}$ is M renormalised after the un-flagged entries have been zeroed out. For indexing 0 is used for regular behaviour, whilst the members of $\mathcal{A}$ index abnormal behaviour.

Each document has a DP, $G_d^{\mathcal{D}}$, with base measure S_d and concentration α. A document contains $|\mathcal{N}_d|$ words, represented by plate $n \in \mathcal{N}_d$, each of which is assigned a topic drawn from $G_d^{\mathcal{D}}$. It is from the associated multinomial that its word, w_{dn}, is drawn.

2.2. Model Learning

The only information provided, ignoring weakly-informative priors, is the words and abnormalities assigned to each document - everything else is learnt. Collapsed Gibbs sampling is used for learning - all variables given in figure 1 except for the words are sampled/collapsed, with the details for each now given.

Some of the variables can be sampled using previously published methods. The DP concentration parameters are sampled using the method of Escobar & West [4], which uses a Gamma prior[4], with the extension of Teh et al. [12] where concentration parameters are shared by multiple DPs. Variables β and φ, both Dirichlet distributions, are subject to the maximum likelihood method of Minka [10]. This is the only deviation from a Bayesian approach, taken as no better method is available. Both H parameters are collapsed out, using the standard Bayesian formulation - see equation 4 (Griffiths & Steyvers [6] and others [12, 16] also do this.). The various DPs outside the documents, G_0, $G_c^{\mathcal{C}}$ and Q, are identical to the original HDP [12] and DHDP [16] models, and can be sampled identically.

The remaining variables require methods unique to the model in question, with specific details now given:

Sampling a words assigned multinomial, t_{dn}: Each document has a DP, $G_d^{\mathcal{D}}$, with base measure S_d, as defined in equation 1. To sample which multinomial is assigned to a word we use the Chinese restaurant process [2], as extended to multiple levels by Teh et al. [12]. This is a metaphor where the discrete values of samples from a DP distribution are represented by *tables* in a restaurant. The process is described in terms of new *patrons* (each representing a draw) choosing either a new table with a fixed weight (the concentration) or a pre-existing table with a weight that is the number of patrons already sitting at that table. When a new table is selected it is assigned a *meal* that all patrons at the table eat - the meal represents the entity drawn from the base distribution. This metaphor directly describes how to calculate the probabilities used in Gibbs sampling.

The task is to sample which table each word is to be (re-)assigned to, including the option of a new table. From now on we refer to *instances* instead of tables[5]. Each instance earns its name from being an instance of a draw from the base measure, in this case a multinomial over words, H_t. To give a convenient definition it is a delta function for a draw from the DPs base measure, in the case of a documents DP, $i^{\mathcal{D}} = \delta_{H_t}$. In a hierarchy of DPs this relation is recursive, where an instance represents a draw from the base measure via an instance drawn from its base measure DP, and so on - it serves to define $\delta_{\delta_x} = \delta_x$. As a DP is exchangeable when resampling a word it can be removed from the current set, by subtracting its count from the relevant instance, and then reassigned based on the distribution - it is assumed that it has already been removed in the following.

The probability of an instance, $i^{\mathcal{D}}$, being assigned to word w_n is

$$P(i^{\mathcal{D}}|w_n, G^{\mathcal{D}}) \propto P(w_n|i^{\mathcal{D}})P(i^{\mathcal{D}}|G^{\mathcal{D}}) \qquad (3)$$

where the omitted division by $P(w_n)$ is constant for all terms and hence irrelevant. $P(w_n|i^{\mathcal{D}})$ is taken to equal $P(w_n|H_t)$, where H_t is the topic that $i^{\mathcal{D}}$ is instancing, such that

$$P(w_n|i^{\mathcal{D}}) = P(w_n|H_t) = \frac{c_w + \beta_w}{\sum_{v\in\mathcal{W}}(c_v + \beta_v)} \qquad (4)$$

where c_w is the number of times word w has been drawn from H_t, $\mathcal{W}$ is the set of all words and β is the vector of parameters for the Dirichlet prior. Note that H_t is integrated out and the probability given in terms of the posterior. The second term can be expanded via the Chinese restaurant interpretation of a DP as

$$i^{\mathcal{D}}|G^{\mathcal{D}} \sim \frac{1}{\alpha + \sum_{j\in I^{\mathcal{D}}} c_j}\left(\alpha S_d + \sum c_j i_j^{\mathcal{D}}\right) \qquad (5)$$

where $I^{\mathcal{D}}$ is the current set of instances in the documents DP and c_j the number of words assigned to instance j. S_d, as previously defined in equation 1, can be expanded to get

$$i^{\mathcal{D}}|G^{\mathcal{D}} \sim \frac{\left(\alpha\left[\hat{M}(0)G^{\mathcal{C}} + \sum_{t\in\mathcal{A}}\hat{M}(t)i_t^{\mathcal{A}}\right] + \sum c_j i_j^{\mathcal{D}}\right)}{\alpha + \sum_{j\in I^{\mathcal{D}}} c_j} \qquad (6)$$

where, for consistency, $i_t^{\mathcal{A}} = \delta_{H_t^{\mathcal{A}}}$. To get the final equation $G^{\mathcal{C}}$ is expanded, and then G_0 within it, using the exact same pattern.

Sampling a documents DP, $G_d^{\mathcal{D}}$: As for the other DPs it is helpful[6] to resample what each instance is instancing. This is similar to the above, and involves calculating

$$P(i^{\mathcal{C},\mathcal{A}}|W_{i^{\mathcal{D}}}, S_d) \propto P(W_{i^{\mathcal{D}}}|i^{\mathcal{C},\mathcal{A}})P(i^{\mathcal{C},\mathcal{A}}|S_d) \qquad (7)$$

where $W_{i^{\mathcal{D}}}$ is the set of all words in the document that are assigned to the instance being resampled. The superscript of i containing a $\mathcal{C}$ and a $\mathcal{A}$ indicates that it could be a normal topic from the documents cluster, or an abnormal topic, respectively, and the omitted divisor is constant for all terms, hence irrelevant. The first term expands as a multinomial distribution,

$$P(W_{i^{\mathcal{D}}}|i^{\mathcal{C},\mathcal{A}}) = (\sum_{v\in\mathcal{W}} s_v)! \prod_{v\in\mathcal{W}} \frac{p_v^{s_v}}{s_v!} \qquad (8)$$

where s_v is the number of words of type $v \in \mathcal{W}$ in the set of words assigned to the current instance, $W_{i^{\mathcal{D}}}$, and p_v is the mean probability of word v for the posterior from which the identifiers associated multinomial, H_t, is drawn, which is given in equation 4. The second term is in effect S_d, as given in equation 1 - it again has $G^{\mathcal{C}}$ and G_0 expanded out.

[4] Set to `Gamma`$(1, 1)$ in all cases.

[5] Otherwise in a hierarchical structure you get the absurdity of tables associated with tables - Teh et al. [12] attempted resolve this with their *Chinese restaurant franchise*, but it only handles one extra level.

[6] Whilst it would theoretically converge without doing this it would take too long. However, resampling $G^{\mathcal{C}}$ can be skipped without consequence, and G_0 does not need resampling as its base measure is collapsed.

Collapsing a clusters behaviour multinomial, M_c: Each cluster has a multinomial on behaviour - it is in effect the ratio of words in a document assigned to each behaviour, assuming the behaviour exists in the current document. When inferring this multinomial instead of sample counts for all entries, leading to the standard Bayesian formulation with a Dirichlet distribution, sample counts for the subset of behaviours in each document are provided. A Bayesian solution starts with a multinomial PDF where the unknown counts have been summed out

$$P(\mathbf{c}, k|M) = n! \prod_{i \in k} \frac{M_i^{c_i}}{c_i!} \sum_{\forall j \in u; c_j \in [0,\infty)} \prod_{j \in u} \frac{M_j^{c_j}}{c_j!} \tag{9}$$

where k is the set of outcomes for which counts are known, $\mathbf{c}$ that set of counts, $n = \sum \mathbf{c}$ and u is the set of outcomes for which counts are unknown. This has made the assumption that the missing behaviours are in effect drawn, but have not been observed - another option would be to take the view of deleting the missing terms from the multinomial and renormalising. Mathematically the only difference between these two solutions is a '+1', which proves to be of no consequence. By repeated application of

$$\sum_{i \in [0,\infty)} \frac{(i+n)!}{i!n!} x^i = \frac{1}{(1-x)^{n+1}} \tag{10}$$

it simplifies to

$$P(\mathbf{c}, k|M) = n! \prod_{i \in k} \frac{M_i^{c_i}}{c_i!} \left\{ \sum_{i \in k} M_i \right\}^{-(n+1)} \tag{11}$$

Given we have multiple draws, one for each document in the cluster, we need to multiply many such terms together. Examination indicates that this multiplication gives the data probability as

$$P(D|M) \propto \prod_{c,k \in \mathcal{K}} \left\{ \sum_{i \in k} M_i \right\}^{c} \tag{12}$$

where we have a set, $\mathcal{K}$, containing associated pairs of counts, c, and known-sets, k, where the counts are exponents for the sum of terms from M selected by k. The terms for known counts, $M_i^{c_i}$, are represented using known-sets containing a single entry, i. This limits the number of parameters to the number of behaviour combinations, independent of the number of documents - a desirable property. It may be noted that this is a generalisation of the Dirichlet distribution, which is equivalent when only complete sets of sample counts are observed. This makes the application of Bayes with a Dirichlet prior, φ from figure 1, trivial.

```
$l \leftarrow 0$
for $n \in \mathcal{N}_d$:
  $p \leftarrow 0$
  # Omitted loop would go here
  for $r \in [1, R]$:
    $p \leftarrow p + \sum_s P(w_n, s = t^r_{dn} | \{\forall t^r_{dm}; m < n\}, \ldots)$
    $t^r_{dn} \sim P(t^r_{dn} | \{\forall t^r_{dm}; m < n\}, \ldots)$
  $p \leftarrow p/R$
  $l \leftarrow l + \log(p)$
$\log(P(W|\ldots)) \simeq l$
```

Figure 2. The left to right algorithm, modified for speed. R is the number of particles and t is taken to index the instance from the documents DP that a word is assigned to.

Draws of M_c are used when sampling several other variables, but in all cases the drawn multinomial has its individual terms multiplied by independent terms - this suggests that by calculating the expectation we can collapse out M_c instead of making a draw. This is done using importance sampling with a uniform sampling distribution[7].

Sampling a documents cluster, c_d: Due to the introduction of a distribution over behaviour for each cluster the cluster resampling method of Wang et al. [16] needs to be trivially modified. Specifically, the probabilities of assigning each cluster, including the option of a new cluster (The two parts of Wang's equation 22.) need to be multiplied by a further term, the probability of drawing the behaviours seen in the document from the clusters associated multinomial, M_c.

Given the above sampling methods one or more samples may be drawn from the model using Gibbs sampling. For initialisation an incremental approach identical to Griffiths & Steyvers [6] is used for the words; the documents are all initialised to belong to the same cluster[8]. A further note is that the model becomes sensitive to how the concentration parameters are initialised when there is a lot of data - this is resolved by resampling them more often than the other variables.

2.3. Online Behaviour Classification

With a model learnt the next task is to detect abnormalities. This can be done either unsupervised, by detecting outliers as in many past works [3, 12, 16], or supervised, by classifying novel documents in terms of their behaviour. In classifying a documents behaviour its uncertainty is calculated as an intermediate step, so we now consider classification only. A modified version of Wallach's left-to-right algorithm [14] is used. Specifically, the innermost loop is dropped, giving the algorithm given in figure 2. This is done as the original algorithm is $O(n^2)$, where n is the number of words in a document - by losing the inner loop it becomes

[7] Importance sampling using a Dirichlet distribution was also tried, but found to confer no great advantage with an excessive computational cost; plus reliable parameter estimation for the Dirichlet proved problematic.

[8] Several cluster initialisation methods were tried - the choice does not appear to matter, but initialising to one makes the initial iterations faster.

$O(n)$, which achieves real time performance, e.g. the 30000 words per 5 second document of the mile end data set (See section 4) will typically take around 3 seconds to process (Including the repeated runs required to determine the values of interest.)[9]. As an added bonus the algorithm is online, processing words as they become available, and hence can be run as the clips features are being collected, so there is minimal delay between the clip ending and the result being available, unlike the importance sampling method of dLDA. Little if any loss in accuracy is observed when running without the innermost loop.

To expand the probability term calculated it is $P(W|c_d, F_d, \ldots)$ (The $\ldots$ represents the many terms that make up the model), i.e. the probability is dependent on the documents abnormality flags and cluster assignment. For a novel document the abnormality flags are unknown and to be inferred, so $P(F_d|W, \ldots)$ is required. The probability of a cluster can be calculated from the model, so that $P(W, c_d|F_d, \ldots) = P(W|c_d, F_d, \ldots)P(c_d|\ldots)$, meaning the cluster variable can be summed out, $P(W|F_d, \ldots) = \sum_{c_d} P(W, c_d|F_d, \ldots)$. Bayes rule can then be applied to switch W and F_d, making use of the prior over F_d, which is ξ. To optimise this prior each Bernoulli parameter is optimised whilst the others remain constant, by adjusting to get a specified false positive rate for that behaviour versus all others. For the experiments we use 10%, on the principal that a system that generates too many false positives will soon be ignored [8]. This is iterated for ξ until the overall inlier rate stops improving. With a large number of abnormalities checking all possible combinations proves expensive - to resolve this a greedy approach is taken where, starting with no flags set, all combinations with a single flag changed from the current state are considered. At each step the best option is taken until a local extrema is found. As most documents belong to the normal group they only require work proportional to the number of abnormalities. Also, in most cases there will never be more than one abnormality, so the vast majority of the time the local minima will be the global minima, as regular behaviour and all cases of one abnormality are always checked.

3. Synthetic Example

To demonstrate the working of the algorithm and clearly highlight the primary difference between it and dLDA [8] results for a synthetic data set are presented. Figure 3 gives the output of dDHDP, which is very close to ground truth - the dataset consists of documents with 5×5 grids, where each grid cell represents a discrete word, and each document contains mixtures of topics consisting of either vertical or horizontal lines. For visualisation the grid is treated as an image, with brightness proportional to the number of words. dLDA's equivalent results are given in figure 4. Two simulations are run, with the results given in figure 5. The first uses abnormal topics - dLDA can detect these and does so, to the same standard as dDHDP. However, for the second simulation abnormalities that consist of normal topics happening at an unusual time are used. dLDA can not solve this problem and fails, whilst dDHDP is successful.

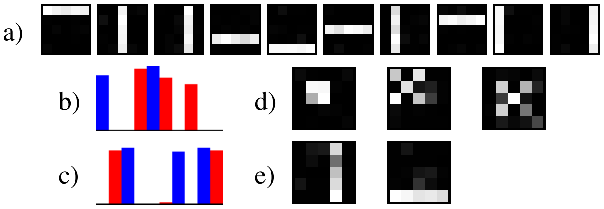

Figure 3. The output of a dDHDP run with 8 training examples for each abnormality, with both the unusual topics and wrong phase topics. a) are the normal topics whilst b & c) are the topic histograms for the two clusters - b) contains all the horizontal lines whilst c) the vertical lines. The abnormal topics are given in d) - these are abnormal behaviours, detectable without context, whilst in e) are the two cases of a normal topic occurring in the wrong cluster - these require temporal context to detect as they are otherwise identical to normal topics given in a).

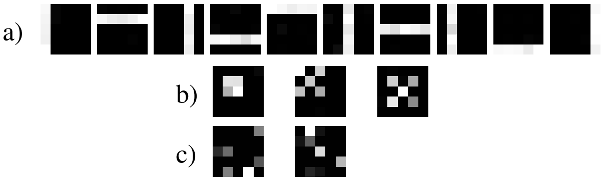

Figure 4. dLDA results for the same test run on dDHDP in figure 3. As it has no clusters, hence no topic histograms, but the normal topics are on row a) and the abnormal topics are on row b) - it has done equally well. For the normal topics occurring at unusual times however it has failed, and row c) simply contains noise.

4. Experiments

Real world experiments are performed using the publicly available[10] *mile end* video sequence. Example frames are given in figure 6. Five second non-overlapping video clips are used as documents - as it runs at 30fps each document consists of 150 frames. The resolution is 360×288, and for the purpose of feature extraction it is divided into a grid 45×36, made up of 8×8 pixel cells. A feature (word) is optionally extracted from each cell - if nothing is happening then no feature is generated, but if optical flow detects motion then a word representing the direction, as quantised to the four compass directions, is generated. Additionally, if background subtraction detects a stationary object then a

[9] On one core of a 2Ghz processor.

[10] Available from http://www.eecs.qmul.ac.uk/~jianli/Junction.html

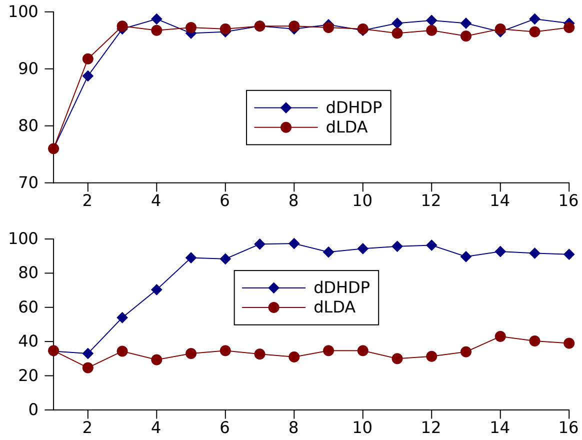

Figure 5. Two graphs, giving the percentage of inliers against the number of training examples for dLDA and dDHDP. The top graph uses the three abnormal topics - in this scenario the two algorithms get identical results. The second graph uses the two cases of normal behaviour happening at the wrong time, here dLDA can not solve the problem, with the graph remaining flat at around $33\frac{1}{3}\%$, which is no better than guessing, whilst dDHDP rises to $90\%+$ with sufficient training examples. These were trained with 256 normal documents, and all documents contain 256 words.

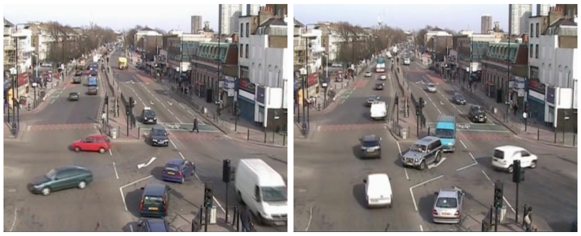

Figure 6. Frames from the video sequence, illustrating both the video and the abnormal behaviours used for the experiment.

fifth word may be generated. The words generated encode the grid position, resulting in $45 \times 36 \times 5$ discrete words that can present in each frame - the words from all the frames in a clip are combined to form the final document.

The frames selected for figure 6 illustrate the abnormalities used for weakly-supervised learning. The left frame is a u-turn being performed by a red hatchback, whilst the right frame is a white van turning right from the middle area whilst cars continue to travel vertically. These behaviours are relatively rare, but not so rare as to impede testing - whilst neither are examples of behaviour that would be of interest in real life they are both subtle, hard to detect behaviours that occur at the same time as other activities. In addition, the second abnormality, turning right though vertical traffic, includes temporal context, as it should not be detected when cars are not travelling vertically.

83.7%

364	13	37	87.9%
4	6	1	54.5%
22	3	45	64.3%

74.2%

351	22	41	84.8%
0	11	0	100.0%
56	8	6	8.6%

Figure 7. Confusion matrices: dDHDP is on the left, dLDA on the right. Rows index the true assignment for each clip, columns the estimate, with the first of each being normal, the second u-turns and the third is cutting across traffic. Above each grid is the overall inlier percentage, and adjacent to each row is the inlier percentage for the behaviour in question. Within each grid are the document counts assigned to each combination of ground truth/estimate.

Comparisons are performed against dLDA[11], as it is the only other topic model capable of supervision[12] [8]. Two experiments are performed - first classification is demonstrated, then the detection of abnormalities that have not been trained for. To give dLDA the greatest chance possible its topic count is set to the number of topics detected by dDHDP - in practice it would not have this advantage.

4.1. Classification

Figure 7 gives confusion matrices for the task of classifying behaviours that have already been observed. For training only 4 abnormal documents, equivalent to 20 seconds, have been used for each behaviour. As each clip extends over multiple documents, and most of the clips only contain part of an example that is split over two or more clips this is equivalent to around 2 actual examples[13] To learn normal behaviour 96 clips are used, equalling 8 minutes, whilst the remaining 42 minutes are used for testing.

The presented algorithm outperforms dLDA with an inlier rate of 83.7% versus 74.2%. For correctly classifying normal behaviour a slight performance increase is seen, and for the u-turn category dLDA comes out on top. It does so at the expense of having 30 false positives in that category versus 19 for dDHDP however. The key difference occurs with detecting vehicles cutting across traffic from the central area of the junction, the abnormality that can only be reliably differentiated given context. dLDA essentially fails at this task, unable to distinguish when cars are and are not allowed to make the turn, whilst dDHDP can correctly model the phasing of the traffic lights, and know which behaviour is being presented.

[11] Thanks go to the authors for providing their implementation.

[12] The dLDA papers compare against a LDA-C model, which is simply LDA trained for each behaviour - it is hardly competitive, as has been demonstrated [8].

[13] The code to select documents does not split instances of behaviour between the training/testing sets, and tries to put all documents from a specific example in the same set.

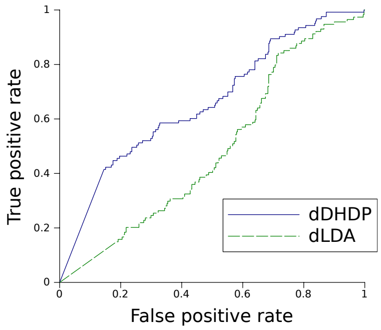

Figure 8. ROC curves indicating the detection of unknowns.

4.2. Detection

In an effort to illustrate a semi-realistic scenario the two approaches have their ability to detect both known and unknown behaviours simultaneously tested. The assumption is made that if a document is tagged as being one of the two abnormalities it is treated as such, but if it is tagged as normal it is then subject to the outlier detection that is regularly used by single class models. For the purpose of this test the two abnormal classes are used as before, under the same testing conditions, but in addition many further abnormalities are marked in the video sequence. A *receiver operating characteristic* (ROC) curve is used to represent the results in figure 8. Such a curve does not initially work in this scenario, as the data marked as abnormal by the behavioural classification is not subject to the detection threshold that is being varied - to resolve this the curve is extended for both algorithms to points $(0,0)$ and $(1,1)$. The area under the curve for dLDA is 0.49, whilst for dDHDP it is 0.66. Most noticeable is how the much improved classification rate of dDHDP allows it to jump ahead of dLDA.

5. Conclusions

A technically sophisticated approach to the problem of detecting abnormal behaviours in video has been presented. It allows for two modes of detection (known and unknown abnormalities) and includes a model of global context, allowing it to learn behaviours for which it was previously not possible to do so. Pragmatic considerations have been met - weak supervision with extremely limited examples has been demonstrated to work, new documents can be categorised in real time and there are no parameters that need tuning for a given input. Whilst limited the results clearly demonstrate the algorithms ability to learn abnormalities that require a model of the global context of a scene.

Future work needs to include further domains, as the approach is in principal quite general, and applicable to other kinds of rare abnormality detection, e.g. detecting abnormal social interactions online. Further practical concerns exist - for incremental learning when new behaviours are found, and for active learning to improve a model by finding new behaviours of interest. Detecting new behaviours as outliers often fails to differentiate noise from novel behaviour, and methods to differentiate the two are essential.

References

[1] D. Andrzejewski, A. Mulhern, B. Liblit, and X. Zhu. Statistical debugging using latent topic models. *ECML*, 18:6–17, 2007. 2

[2] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973. 4

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Machine Learning Research*, 3:993–1022, 2003. 2, 5

[4] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. American Statistical Association*, 90(430):577–588, 1995. 4

[5] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006. 1

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc. Nat. Academy of Sciences, USA*, 2004. 4, 5

[7] T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behaviour in video. *ICCV*, 2009. 1, 2

[8] T. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying rare and subtle behaviours: A weakly supervised joint topic model. *PAMI*, pages 2140–2157, 2011. 2, 6, 7

[9] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. *CVPR*, 2010. 1

[10] T. P. Minka. Estimating a dirichlet distribution. *Technical Report, MIT*, 2000. 4

[11] I. Saleemi, L. Hartung, and M. Shah. Scene understanding by statistical modeling of motion patterns. *CVPR*, 2010. 1

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. American Statistical Association*, 101(476):1566–1581, 2006. 2, 4, 5

[13] H. M. Wallach. Topic modeling: Beyond bag-of-words. *ICML*, 23:977–984, 2006. 2

[14] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. *ICML*, 26:1105–1112, 2009. 5

[15] X. Wang and E. Grimson. Spatial latent Dirichlet allocation. *NIPS*, 2007. 2

[16] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *PAMI*, 31(3):539–555, 2009. 1, 2, 4, 5

[17] S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. *CVPR*, pages 1–6, 2007. 2

[18] T. Xiang and S. Gong. Video behaviour profiling for anomaly detection. *PAMI*, 30(5):893–908, 2008. 1, 2

[19] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. *CVPR*, 2:819–826, 2004. 2