

# Dynamic Hierarchical Dirichlet Process for Abnormal Behaviour Detection in Video

Olga Isupova, Danil Kuzin, Lyudmila Mihaylova

Department of Automatic Control and System Engineering, University of Sheffield  
Sheffield, UK

Email: o.isupova@sheffield.ac.uk, dkuzin1@sheffield.ac.uk, l.s.mihaylova@sheffield.ac.uk

**Abstract**—This paper proposes a novel dynamic Hierarchical Dirichlet Process topic model that considers the dependence between successive observations. Conventional posterior inference algorithms for this kind of models require processing of the whole data through several passes. It is computationally intractable for massive or sequential data. We design the batch and online inference algorithms, based on the Gibbs sampling, for the proposed model. It allows to process sequential data, incrementally updating the model by a new observation. The model is applied to abnormal behaviour detection in video sequences. A new abnormality measure is proposed for decision making. The proposed method is compared with the method based on the non-dynamic Hierarchical Dirichlet Process, for which we also derive the online Gibbs sampler and the abnormality measure. The results with synthetic and real data show that the consideration of the dynamics in a topic model improves the classification performance for abnormal behaviour detection.

## I. INTRODUCTION

Unsupervised and semi-supervised learning for various video processing applications is an active research area nowadays. In many situations supervised learning is inappropriate or impossible. For example, in abnormal behaviour detection it is difficult to predict in advance what kind of abnormality may happen, collect and label a training dataset for some supervised learning algorithm.

Within the unsupervised methods topic modeling is a promising approach for abnormal behaviour detection [1]–[3]. It allows not only to give warnings about abnormalities but also provides an information about typical patterns of behaviour or motion.

Topic modeling [4], [5] is a statistical tool for discovering a latent structure in data. In text mining it is assumed that unlabelled documents can be represented as mixtures of topics, where the topics are distributions over words. The topics are latent and the inference in topic models is aimed to discover them.

In the conventional topic models, documents are independent. They share the same set of topics, but weights in a topic mixture for a particular document are independent of weights for all other documents in a dataset. However, in some cases it is reasonable to assume dependence in topic mixtures in different documents.

Consider the analysis of scientific papers of a given conference in text mining. It is expected that if a topic is “hot” in a given year, it would be popular in the next year too. The popularity of the topics changes through the years but in

each two successive years the set of popular topics would be similar. It means that in a topic model the topic mixtures in the documents in successive years are similar to each other.

The same ideas are valid for abnormal behaviour detection. Documents are usually defined as short video clips extracted from a whole video sequence. Topics represent some local motion patterns. If the clips are sufficiently short, motions started in a given clip would continue in the next clip. Therefore it may be expected that the topic mixtures in the successive clips would be similar.

In this paper the dynamic topic model is proposed to improve the performance of abnormal behaviour detection. Two types of dynamics are considered in the topic modeling literature. In the first type the dynamics is assumed on the topic mixtures in documents [6]–[8]. This type of the dynamics is described earlier. In the second type the dynamics is assumed on the topics themselves [9]–[11], i.e. the distributions over words, which correspond to topics, change through time. There are works where both types of the dynamics are considered [12], [13].

In the proposed model the first type of the dynamics is considered. The model is constructed to encourage neighbour documents to have similar topic mixtures. The second type of the dynamics is not assumed, as in the video processing the set of words and their popularity do not change, thus the distributions over words are not expected to change.

Imagine there is an infinitely long video sequence. Motion patterns, which are typical for a scene, may appear and disappear and the total number of these patterns may be infinite. The motion patterns are modelled as topics in the topic model, hence the number of topics in the topic model may potentially be infinite. This kind of intuition may be simulated by a nonparametric model [14]. Therefore the proposed model is nonparametric.

The most related model to the proposed one is presented in [13], which is also a dynamic topic model. The main difference between this model and the proposed one is that in the later a document, although is encouraged to have a topic mixture similar to the one in the previous document, may have any of the topics used in the dataset so far.

In abnormal behaviour detection it is essential to make a decision as soon as possible to warn a human operator to react. We propose batch and online inference for the model based on the Gibbs sampler. During the batch offline set

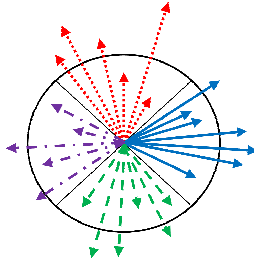


Figure 1. Quantisation of motion directions. Optical flow vectors are quantised into the four directions — up, right, down and left. The vectors of the same category have the same colour on the figure.

up the Gibbs sampler processes a training set of documents, estimating distributions of words in topics. During the online set up testing documents are processed one by one. The main goal of the online inference is to estimate a topic mixture for the current document, without reconsidering all the previous documents. We also propose an abnormality measure, which is used in the final decision making.

The rest of the paper is organised as follows. In section II visual words and documents are defined. The proposed model is described in section III. Section IV presents the inference for the model, while section V introduces the abnormality detection procedure. The experimental results are given in section VI. Section VII concludes the paper.

## II. VIDEO REPRESENTATION

In order to apply the topic modeling approach to video processing it is required to define visual words and visual documents. In this paper a visual word is defined as a quantised local motion measured by an optical flow [15]. The optical flow vector is discretised spatially by averaging among  $N \times N$  pixels. The direction of the average optical flow vector is further quantised into the four main categories — up, right, down and left (Figure 1). The location of the averaged optical flow vector and its categorised direction together form a visual word.

The whole video sequence is divided into non-overlapping clips. Each clip is a visual document. The document consists of all the visual words extracted from the frames that form the corresponding clip.

Topics in topic modeling are defined as distributions over words. They indicate which words appear together. In the video processing applications topics are distributions over visual words. As visual words represent local motions, topics indicate the set of local motions that frequently appear together. They are usually called *activities* or *actions* (e.g. [2], [6], [16], [17]).

Once visual documents, words and topics are defined, the topic model for video processing can be formulated.

## III. PROPOSED MODEL

There is a sequence of documents  $\mathbf{x}_{1:J} = \{\mathbf{x}_j\}_{j=1:J}$ , where each document  $\mathbf{x}_j$  consists of  $N_j$  words  $x_{ji}$ :  $\mathbf{x}_j = \{x_{ji}\}_{i=1:N_j}$ . It is assumed that words are generated from a

set of hidden distributions  $\{\phi_k\}_{k=1:\infty}$ , that are called *topics* and documents are mixtures of this shared set of topics. The number of topics is not fixed. Moreover it is assumed that observing the infinite amount of data we can expect to have an infinite number of topics.

### A. Hierarchical Dirichlet Process Topic Model

This kind of mixture models with a potentially infinite number of mixture components can be modelled with the Hierarchical Dirichlet Process (HDP) [18]. The HDP is a hierarchical extension of the Dirichlet process (DP), which is a distribution over random distributions [19]. Each document  $\mathbf{x}_j$  is associated with a sample  $G_j$  from a DP:

$$G_j \sim \text{DP}(\alpha, G_0), \quad (1)$$

where  $\alpha$  is a concentration parameter,  $G_0$  is a base measure.  $G_j$  can be seen as a vector of mixture components weights, where the number of components is infinite.

The base measure  $G_0$  itself is a sample from another DP:

$$G_0 \sim \text{DP}(\gamma, H), \quad (2)$$

with the concentration parameter  $\gamma$  and the base measure  $H$ . This shared measure  $G_0$  from a DP ensures that the documents will have the same set of topics but with different weights. Indeed,  $G_0$  is almost surely discrete [19], concentrating its mass on the atoms  $\phi_k$  drawn from  $H$ . Therefore,  $G_j$  picks the mixture components from this set of atoms.

A topic, that is an atom  $\phi_k$ , is often modelled as the multinomial distribution with a probability  $\phi_{wk}$  of choosing a word  $w$  [4], [5]. The base measure  $H$  is therefore chosen as the conjugate Dirichlet distribution, usually a symmetric one. Let  $\boldsymbol{\eta} = [\eta, \dots, \eta]$  denote a parameter of this Dirichlet distribution.

The document  $j$  is formed by repeating the procedure of drawing a topic from the mixture:

$$\theta_{ji} \sim G_j \quad (3)$$

and drawing a word from the chosen topic:

$$x_{ji} \sim \text{Mult}(\theta_{ji}) \quad (4)$$

for every token  $i$ , where  $\text{Mult}(\cdot)$  is the multinomial distribution.

1) *Chinese restaurant franchise*: There are several ways of the HDP representation (as well as the DP). In this paper the representation called Chinese restaurant franchise (CRF) is considered as it is used for the derivation of the Gibbs sampling inference scheme. In this metaphor, each document corresponds to a “restaurant”; words correspond to “customers” of the restaurant. The words in the documents are grouped around “tables”. Each table serves a “dish”, which corresponds to a topic. The “menu” of dishes, i.e. the set of the topics, is shared among all the restaurants.

Let  $t_{ji}$  denote a table assignment for the token  $i$  in the document  $j$ ,  $k_{jt}$  denote a topic assignment for the table  $t$  in the document  $j$ . Let  $n_{jt}$  denote the number of words assigned

to the table  $t$  in the document  $j$  and  $m_{jk}$  denote the number of tables in the document  $j$  serving the topic  $k$ . The dots in subscripts mean marginalisation over the corresponding dimension, e.g.  $m_{.k}$  denotes the number of tables among all the documents serving the topic  $k$ , while  $m_{j.}$  denotes the total number of tables in the document  $j$ . Marginalisation over both dimensions  $m_{..}$  means the total number of tables in the dataset.

The generative process of a dataset is as follows. A new token comes to the document  $j$  and chooses one of the occupied tables with a probability proportional to a number of words  $n_{jt}$  assigned to this table, or the new token starts a new table with a probability proportional to  $\alpha$ :

$$p(t_{ji} = t | t_{j1}, \dots, t_{ji-1}, \alpha) = \begin{cases} \frac{n_{jt}}{i-1+\alpha}, & \text{if } t = 1 : m_{j.}; \\ \frac{\alpha}{i-1+\alpha}, & \text{if } t = t^{\text{new}}. \end{cases} \quad (5)$$

If the token starts a new table it chooses one of the used topics with a probability proportional to a number of tables  $m_{.k}$  serving this topic among all the documents, or the token chooses a new topic, sampling it from the base measure  $H$ , with a probability proportional to  $\gamma$ :

$$p(k_{jt}^{\text{new}} = k | k_{11}, \dots, k_{jt-1}, \gamma) = \begin{cases} \frac{m_{.k}}{m_{..} + \gamma}, & \text{if } k = 1 : K; \\ \frac{\gamma}{m_{..} + \gamma}, & \text{if } k = k^{\text{new}}, \end{cases} \quad (6)$$

where  $K$  is a number of topics used so far.

Once the token is assigned to the table  $t_{ji}$  with the topic  $k_{jt_{ji}}$ , the word  $x_{ji}$  for this token is sampled from this topic:

$$x_{jt} \sim \text{Mult}(\phi_{k_{jt_{ji}}}) \quad (7)$$

The correspondence between two representations of the HDP (1) – (4) and (5) – (10) is based on the following equality:  $\theta_{ji} = \phi_{k_{jt_{ji}}}$ .

### B. Dynamic Hierarchical Dirichlet Process Topic Model

In the HDP exchangeability of documents and words is assumed which means that the joint probability of the data is independent of the order of the documents and the words in the documents. However, in the video processing applications this assumption may be invalid. While the words inside the documents are still exchangeable, the documents themselves are not. All actions and motions in the real life last for some time, and it is expected that the topic mixture in the current document is similar to the topic mixture in the previous document. Some topics may appear and disappear but the core structure of the mixture components weights only slightly changes from document to document.

We propose the dynamic extension of the HDP topic model to take into account this intuition. In this model the probability of the topic  $k$  explicitly depends on the usage of this topic in the current and previous documents  $m_{jk} + m_{j-1k}$ , therefore the topic distribution in the current document would be similar to the topic distribution in the previous document. The topic probability still depends on the number of tables serving this

topic in the whole dataset  $m_{.k}$ , but this number is weighted by a non-negative value  $\delta$ , which is a parameter of the model. As in the previous case, it is possible to sample a new topic from the base measure  $H$ .

The generative process can be then formulated as follows. A new token comes to a document and, as before, chooses one of the occupied tables  $t$  with a probability proportional to the number of words  $n_{jt}$  already assigned to it, or it starts a new table with a probability proportional to the parameter  $\alpha$ :

$$p(t_{ji} = t | t_{j1}, \dots, t_{ji-1}, \alpha) = \begin{cases} \frac{n_{jt}}{i-1+\alpha}, & \text{if } t = 1 : m_{j.}; \\ \frac{\alpha}{i-1+\alpha}, & \text{if } t = t^{\text{new}}. \end{cases} \quad (8)$$

If the token starts a new table, it chooses a topic for it. One of the used topics  $k$  is chosen with a probability proportional to the sum of the number of tables having this topic in the current and previous documents  $m_{jk} + m_{j-1k}$  and the weighted number of tables among all the documents, which serve this topic,  $\delta m_{.k}$ . A new topic can be chosen for the table  $t$  with a probability proportional to the parameter  $\gamma$ :

$$p(k_{jt} = k | k_{11}, \dots, k_{jt-1}, \gamma) = \begin{cases} \frac{m_{jk} + m_{j-1k} + \delta m_{.k}}{m_{j.} + m_{j-1.} + \delta m_{..} + \gamma}, & \text{if } k = 1 : K; \\ \frac{\gamma}{m_{j.} + m_{j-1.} + \delta m_{..} + \gamma}, & \text{if } k = k^{\text{new}}. \end{cases} \quad (9)$$

Finally, the word  $x_{ji}$  is sampled for the token  $i$  in the document  $j$ , assigned to the table  $t_{ji} = t$ , which serves the topic  $k_{jt} = k$ . The word is sampled from the corresponding topic  $k$ :

$$x_{ji} \sim \text{Mult}(\phi_k). \quad (10)$$

## IV. INFERENCE

Standard inference algorithms process an entire dataset. For large or stream datasets this batch set up is computationally intractable. Online algorithms process data in a sequential manner, one data point at a time, incrementally updating the variables, corresponding to the whole dataset. It allows to save memory space and reduce the computational time. In this paper a combination of offline batch and online inference is proposed and this section describes it in details.

The Gibbs sampling scheme is used [20]. The inference procedure consists of two parts. Firstly, the traditional batch set up of the Gibbs sampling is applied to the training set of the documents. Then an online set up of the inference is applied for the testing documents. This means that the information about a testing document is incrementally added to the model, not requiring to process the training documents again.

In the Gibbs sampling inference scheme the hidden variables  $\mathbf{t} = \{t_{ji}\}_{j=1:J, i=1:N_j}$  and  $\mathbf{k} = \{k_{jt}\}_{j=1:J, t=1:m_j}$  are sampled from their conditional distributions. In the Gibbs sampler for the HDP model exchangeability of documents and words is used by treating the current variable  $t_{ji}$  as the table assignment for the last token in the last document and  $k_{jt}$  as the topic assignment for the last table in the last document. There is

no exchangeability of documents in the proposed model, but words inside a document are still exchangeable. Therefore, the variable  $t_{ji}$  can be treated as the table assignment for the last token in the current document  $j$ , and the variable  $k_{jt}$  can be treated as the topic assignment for the last table in the current document  $j$ . The documents are processed in the order they appear in the dataset.

The following notation is used below. Let  $V$  denote the size of the words vocabulary,  $\mathbf{t}_{j_1:j_2} = \{t_{ji}\}_{j=j_1:j_2, i=1:N_j}$  is the set of the table assignments for all the tokens in the documents from  $j_1$  to  $j_2$ . Let  $\mathbf{k}_{j_1:j_2} = \{k_{jt}\}_{j=j_1:j_2, t=1:m_j}$  and  $\mathbf{x}_{j_1:j_2} = \{\mathbf{x}_j\}_{j=j_1:j_2}$  denote the corresponding sets for the topic assignments and the observed data. Let  $m_{j_1:j_2 k}$  denote the number of tables having the topic  $k$  in the documents from  $j_1$  to  $j_2$ . Let also  $\mathbf{x}_{jt} = \{x_{ji}\}_{i=1:N_j}$  denote the words assigned to the table  $t$  in the document  $j$ .

Let  $l_{wk}$  denote the number of times the word  $w$  is associated with the topic  $k$ ,  $l_k$  denote the number of tokens associated with the topic  $k$ :  $l_k = \sum_w l_{wk}$ , regardless the word assignments. The notation  $l_{wk}^{j_1:j_2}$  is used for the number of times the word  $w$  associated with the topic  $k$  in the documents from  $j_1$  to  $j_2$ .

The superscript  $-j_i$  indicates the corresponding variable without considering the token  $i$  in the document  $j$ , e.g. the set variable  $\mathbf{t}^{-j_i} = \mathbf{t} \setminus \{t_{ji}\}$  or the count  $n_{jt}^{-j_i}$  is the number of words, assigned the table  $t$  in the document  $j$ , excluding the word for the token  $i$ . Similarly, the superscript  $-jt$  means the corresponding variable without considering the table  $t$  in the document  $j$ .

#### A. Batch Gibbs sampling

1) *Sampling topic assignment  $k_{jt}$* : The topic assignment  $k_{jt}$  for the table  $t$  in the document  $j$  is sampled from the conditional distribution given the observed data  $\mathbf{x}$  and all the other hidden variables, i.e. the table assignments for all the tokens  $\mathbf{t}$  and the topic assignments for all the other tables  $\mathbf{k}^{-jt}$ :

$$p(k_{jt} = k | \mathbf{x}, \mathbf{t}, \mathbf{k}^{-jt}) \propto p(\mathbf{x}_{jt} | k_{jt} = k, \mathbf{k}^{-jt}, \mathbf{t}, \mathbf{x}^{-jt}) p(k_{jt} = k | \mathbf{k}^{-jt}). \quad (11)$$

The likelihood term  $p(\mathbf{x}_{jt} | k_{jt} = k, \mathbf{k}^{-jt}, \mathbf{t}, \mathbf{x}^{-jt})$  can be computed by integrating out the distribution  $\phi_k$ :

$$\begin{aligned} f_k^{-jt}(\mathbf{x}_{jt}) &\stackrel{\text{def}}{=} p(\mathbf{x}_{jt} | k_{jt} = k, \mathbf{k}^{-jt}, \mathbf{t}, \mathbf{x}^{-jt}) = \\ &\int p(\mathbf{x}_{jt} | \phi_k) p(\phi_k | \mathbf{k}^{-jt}, \mathbf{t}, \mathbf{x}^{-jt}) d\phi_k = \\ &\frac{\prod_w \Gamma(l_{wk} + \eta)}{\Gamma(l_k + V\eta)} \frac{\Gamma(l_k^{-jt} + V\eta)}{\prod_w \Gamma(l_{wk}^{-jt} + \eta)}, \end{aligned} \quad (12)$$

where  $\Gamma(\cdot)$  is the gamma-function. In the case when  $k$  is a new topic ( $k = k^{\text{new}}$ ) the integration is done over the prior distribution for  $\phi_{k^{\text{new}}}$ . The obtained likelihood term (12) is then:

$$f_{k^{\text{new}}}^{-jt}(\mathbf{x}_{jt}) = \frac{\prod_w \Gamma(l_{wk^{\text{new}}} + \eta)}{\Gamma(l_{k^{\text{new}}} + V\eta)} \frac{\Gamma(V\eta)}{(\Gamma(\eta))^V}. \quad (13)$$

The second multiplier in (11)  $p(k_{jt} = k | \mathbf{k}^{-jt})$  can be further factorised as:

$$p(k_{jt} = k | \mathbf{k}^{-jt}) \propto p(\mathbf{k}_{j+1:J} | \mathbf{k}_{1:j}^{-jt}, k_{jt} = k) p(k_{jt} = k | \mathbf{k}_{1:j}^{-jt}). \quad (14)$$

The first term in (14) is the probability of the topic assignments for all the tables in the next documents depending on the change of the topic assignment for the table  $t$  in the document  $j$ . Consider the topic assignments in the document  $j+1$  firstly. From (9) it is:

$$\begin{aligned} g_k^{-jt}(\mathbf{k}_{j+1}) &\stackrel{\text{def}}{=} p(\mathbf{k}_{j+1} | \mathbf{k}_{1:j}^{-jt}, k_{jt} = k) = \\ &\frac{\gamma^{|\mathcal{K}_{j+1}^{\text{born}}|} \prod_{s \in \mathcal{K}_{j+1}^{\text{born}}} (m_{j+1s} - 1)! (1 + \delta)^{m_{j+1s} - 1}}{\prod_{n=1}^{m_{j+1}} (m_{j\cdot} + n - 1 + \delta(m_{1:j\cdot} + n - 1) + \gamma)} \times \\ &\prod_{s \notin \mathcal{K}_{j+1}^{\text{born}}} \prod_{n=1}^{m_{j+1s}} (m_{js}^{-jt \rightarrow k} + n - 1 + \delta(m_{1:j s}^{-jt \rightarrow k} + n - 1)) \propto \\ &\prod_{s \notin \mathcal{K}_{j+1}^{\text{born}}} \prod_{n=1}^{m_{j+1s}} (m_{js}^{-jt \rightarrow k} + n - 1 + \delta(m_{1:j s}^{-jt \rightarrow k} + n - 1)), \end{aligned} \quad (15)$$

where the sign of proportionality is used w.r.t.  $k_{jt}$ ,  $\mathcal{K}_{j+1}^{\text{born}}$  is the set of the topics that firstly appear in the document  $j+1$ , the superscript  $-jt \rightarrow k$  means that  $k_{jt}$  is set to  $k$  for the corresponding counts,  $|\cdot|$  is the cardinality of the set. The similar probabilities of the topic assignments for all the next documents  $j' = j+2 : J$  depend on  $k$  only in the term  $m_{1:j'-1}^{-jt \rightarrow k}$ . It is assumed that the influence of  $k$  on these probabilities is not significant and the first term in (14) is approximated by the probability of the topic assignments in the document  $j+1$  (15) only:

$$p(\mathbf{k}_{j+1:J} | \mathbf{k}_{1:j}^{-jt}, k_{jt} = k) \approx g_k^{-jt}(\mathbf{k}_{j+1}). \quad (16)$$

The second term in (14) is the prior for  $k_{jt}$ :

$$p(k_{jt} = k | \mathbf{k}_{1:j}^{-jt}) \propto \begin{cases} m_{jk}^{-jt} + m_{j-1k} + \delta m_{1:j k}^{-jt}, & \text{if } k = 1 : K; \\ \gamma, & \text{if } k = k^{\text{new}}. \end{cases} \quad (17)$$

As a result, (14) is computed as follows:

$$p(k_{jt} = k | \mathbf{k}^{-jt}) \propto \begin{cases} g_k^{-jt}(\mathbf{k}_{j+1}) (m_{jt} + m_{j-1k} + \delta m_{1:j k}^{-jt}), & \text{if } k = 1 : K; \\ g_{k^{\text{new}}}^{-jt}(\mathbf{k}_{j+1}) \gamma, & \text{if } k = k^{\text{new}}. \end{cases} \quad (18)$$

Combining (12) – (13) and (18) the topic assignment sampling distribution can be expressed as:

$$p(k_{jt} = k | \mathbf{x}, \mathbf{t}, \mathbf{k}^{-jt}) \propto f_k^{-jt}(\mathbf{x}_{jt}) p(k_{jt} = k | \mathbf{k}^{-jt}). \quad (19)$$

2) *Sampling  $t_{ji}$* : The table assignment  $t_{ji}$  for the token  $i$  in the document  $j$  is sampled from the conditional distribution given the observed data  $\mathbf{x}$  and all the other hidden variables, i.e. the topic assignments for all the tables  $\mathbf{k}$  and the table assignments for all the other tokens  $\mathbf{t}^{-ji}$ :

$$p(t_{ji} = t | \mathbf{x}, \mathbf{k}, \mathbf{t}^{-ji}) \propto p(x_{ji} | \mathbf{t}^{-ji}, t_{ji} = t, \mathbf{x}^{-ji}, \mathbf{k}) p(t_{ji} = t | \mathbf{t}^{-ji}) \quad (20)$$

The first term in (20) is the likelihood of the word  $x_{ji}$ . It changes depending on whether  $t$  is one of the previously used table or it is a new table. For the case when  $t$  is the table which is already used the likelihood is:

$$f_{k_{jt}}^{-ji}(x_{ji}) = p(x_{ji} | t_{ji} = t, \mathbf{t}^{-ji}, \mathbf{k}, \mathbf{x}^{-ji}) = \frac{l_{x_{ji} k_{jt}} + \eta}{l_{k_{jt}} + V\eta} \quad (21)$$

Consider now the case when  $t_{ji} = t^{\text{new}}$ , i.e. the likelihood of the word  $x_{ji}$  being assigned to a new table. This likelihood can be found by integrating out the possible topic assignments  $k_{jt^{\text{new}}}$  for this table:

$$r_{t^{\text{new}}}(x_{ji}) \stackrel{\text{def}}{=} p(x_{ji} | \mathbf{t}^{-ji}, t_{ji} = t^{\text{new}}, \mathbf{x}^{-ji}, \mathbf{k}) = \sum_{k=1}^K f_k^{-ji}(x_{ji}) p(k_{jt^{\text{new}}} = k | \mathbf{k}) + f_{k^{\text{new}}}^{-ji}(x_{ji}) p(k_{jt^{\text{new}}} = k^{\text{new}} | \mathbf{k}), \quad (22)$$

where  $p(k_{jt^{\text{new}}} = k | \mathbf{k})$  is as (18).

The second term in (20) is the prior for  $t_{ji}$ :

$$p(t_{ji} = t | \mathbf{t}^{-ji}) \propto \begin{cases} n_{jt}, & \text{if } t = 1 : m_j; \\ \alpha, & \text{if } t = t^{\text{new}}. \end{cases} \quad (23)$$

Then the conditional distribution for sampling a table assignment  $t_{ji}$  is:

$$p(t_{ji} = t | \mathbf{x}, \mathbf{k}, \mathbf{t}^{-ji}) \propto \begin{cases} f_{k_{jt}}^{-ji}(x_{ji}) n_{jt}, & \text{if } t = 1 : m_j; \\ r_{t^{\text{new}}}(x_{ji}) \alpha, & \text{if } t = t^{\text{new}}. \end{cases} \quad (24)$$

If a new table is sampled, then a topic for it is sampled from (19).

### B. Online inference

In online or distributed implementations of inference algorithms in topic modeling the idea is to separate global variables, i.e. those that depend on the whole set of data, and local variables, i.e. those that depend only on the current document [21]–[23].

For the proposed dynamic HDP model the global variables are the distributions  $\phi_k$ , which are approximated by the counts  $l_{wk}$ , and the global topic popularity, which is estimated by the counts  $m_{\cdot k}$ . Note, that the relative relationship between counts is important, rather than the absolute values of the counts. The local variables are the topic mixture weights for each document, governed by the counts  $m_{jk}$ . The training dataset is assumed to be large enough such that the global variables are

well estimated by the counts available during the training stage and a new document can only slightly change the obtained ratios of the counts.

Following this assumption, the learning procedure is organised as follows. The batch Gibbs sampler is run for the training set of the documents. After this training stage the global counts  $l_{wk}$  and  $m_{\cdot k}$  for all  $w$  and  $k$  are stored and used for the online inference of the testing documents. For each testing document the online Gibbs sampler is run to sample table assignments and topic assignments for this document only. The online Gibbs sampler updates the local counts  $m_{jk}$ . After the Gibbs sampler converges, the global counts  $l_{wk}$  and  $m_{\cdot k}$  are updated with the information obtained by the new document.

The equations for the online version of the Gibbs sampler slightly differ from the batch ones (19) and (24). Namely, the conditional probability  $p(k_{jt} = k | \mathbf{k}^{-jt})$  in the topic assignment sampling distribution (19) differs from (14). As next documents are not observed during processing the current document, this probability consists only of the prior term  $p(k_{jt} = k | \mathbf{k}_{1:j}^{-jt})$ :

$$p_{\text{online}}(k_{jt} = k | \mathbf{k}^{-jt}) = \begin{cases} m_{jk}^{-jt} + m_{j-1k} + \delta m_{1:j k}^{-jt}, & \text{if } k = 1 : K; \\ \gamma, & \text{if } k = k^{\text{new}}. \end{cases} \quad (25)$$

Substituting this expression into (19) the obtained sampling distribution for the topic assignment in the online Gibbs sampler is:

$$p_{\text{online}}(k_{jt} = k | \mathbf{x}, \mathbf{t}, \mathbf{k}^{-jt}) \propto \begin{cases} f_k^{-jt}(\mathbf{x}_{jt}) (m_{jt} + m_{j-1k} + \delta m_{1:j k}^{-jt}), & \text{if } k = 1 : K; \\ f_{k^{\text{new}}}^{-jt}(\mathbf{x}_{jt}) \gamma, & \text{if } k = k^{\text{new}}. \end{cases} \quad (26)$$

The updating distribution for the topic assignment in the online Gibbs sampler remains the same as in the batch version (24).

## V. ABNORMALITY DETECTION

Topic models provide a probabilistic framework for abnormality detection. Under this framework the abnormality measure is the likelihood of data. The low value of the likelihood means the built model cannot explain the current observation, i.e. there is something atypical in the observation, which is not fitted to the typical motion patterns, learnt by the model.

From the Gibbs sampler we have estimates of the distributions  $\phi_k$  and posterior samples of the table and topic assignments. This information can be used to estimate the predictive likelihood of a new clip. The predictive likelihood, normalised by the length  $N_j$  of the clip in terms of visual words, is used as an abnormality measure in this paper.

The predictive likelihood is estimated via a harmonic mean [24], as it allows to use the information from the

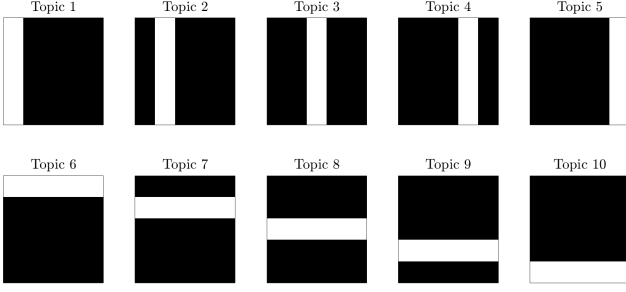


Figure 2. Graphical representation of the topics in the synthetic dataset. There are 25 words, organised into a  $5 \times 5$  matrix, where a word corresponds to a cell in this matrix. The topics are represented as the coloured matrices, where the colour of the cell indicates the probability of the corresponding word in a given topic, the lighter the colour the higher the probability value.

posterior samples:

$$p(\mathbf{x}_j | \mathbf{x}_{1:j-1}) = \left( \sum_{\mathbf{t}_{1:j}, \mathbf{k}_{1:j}} \frac{p(\mathbf{t}_{1:j}, \mathbf{k}_{1:j} | \mathbf{x}_j, \mathbf{x}_{1:j-1})}{p(\mathbf{x}_j | \mathbf{t}_{1:j}, \mathbf{k}_{1:j}, \mathbf{x}_{1:j-1})} \right)^{-1} \approx \left( \frac{1}{S} \sum_{s=1}^S \frac{1}{p(\mathbf{x}_j | \mathbf{t}_{1:j}^s, \mathbf{k}^s, \mathbf{x}_{1:j-1})} \right), \quad (27)$$

where  $S$  is the number of the posterior samples,  $\mathbf{t}_{1:j}^s$  and  $\mathbf{k}_{1:j}^s$  are from the  $s$ -th posterior sample obtained by the Gibbs sampler, and

$$p(\mathbf{x}_j | \mathbf{t}_{1:j}^s, \mathbf{k}^s, \mathbf{x}_{1:j-1}) = \prod_{k=1}^K \frac{\prod_w \Gamma(l_{wk}^{1:j} s + \eta)}{\Gamma(l_k^{1:j} s + V\eta)} \frac{\Gamma(l_k^{1:j-1} s + V\eta)}{\prod_w \Gamma(l_{wk}^{1:j-1} s + \eta)}. \quad (28)$$

The superscript  $s$  on the counts means these counts are from the  $s$ -th posterior sample.

The abnormality detection procedure is then as follows. The batch Gibbs sampler is run on the training dataset. Then for each clip from the testing dataset first the online Gibbs sampler is run to obtain the posterior samples of the hidden variables corresponding to the current clip. Afterwards the abnormality measure:

$$a(\mathbf{x}_j) = \frac{1}{N_j} p(\mathbf{x}_j | \mathbf{x}_{1:j-1}) \quad (29)$$

is computed for the current clip. If the abnormality measure is below than some threshold, the clip is labelled as abnormal, otherwise as normal. And the next clip from the testing dataset is processed.

## VI. EXPERIMENTS

In this section the proposed method is applied to abnormality detection<sup>1</sup>. The method is compared with the one, based on the HDP topic model, where for the HDP topic model the online version of the Gibbs sampler and the abnormality measure are derived similarly to the dynamic HDP (for the batch Gibbs sampler of the HDP topic model the implementation

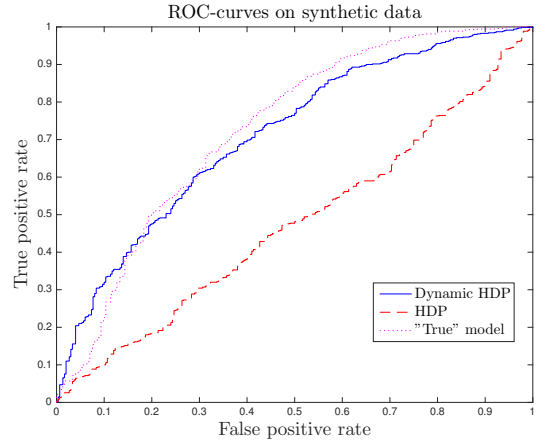


Figure 3. The ROC-curves for the synthetic data obtained by both models. The ROC-curve, obtained by the likelihood, computed with the known true hidden variables, is labelled as a “true” model.

by Chong Wang is used<sup>2</sup>). Each of the algorithms has 5 runs with different initialisations to obtain 5 independent posterior samples. Both batch and online samplers are run for 1000 “burn-in” iterations.

The methods are compared on both synthetic and real data. The abnormality classification accuracy is used for the quantitative comparison of the methods. For computing classification accuracy the ground truth about abnormality should be provided. For the synthetic data the ground truth is known from the generation, for the test real data the clips are labelled manually as normal or abnormal. Note, the methods use only unlabelled data, labels are applied for performance measure.

In statistics the following measures are used for binary classification: *true positive* (TP) is the number of observations which are correctly detected by an algorithm as positive, *false negative* (FN) is the number of observations which are incorrectly detected as negative, *true negative* (TN) is the number of observations which are correctly detected as negative, and *false positive* FP is the number of observations which are incorrectly detected as positive [25].

For the quantitative comparison the area (AUC) under the receiver operating characteristic (ROC) curve is used in this paper. The curve is built by plotting the true positive rate versus the false positive rate while the threshold varies. The true positive rate (TPR), also known as recall, is defined as:

$$TRP = \frac{TP}{TP + FN}. \quad (30)$$

The false positive rate (FPR), also known as fall-out, is defined as:

$$FPR = \frac{FP}{FP + TN}. \quad (31)$$

### A. Synthetic data

The popular “bar” data is used as a synthetic data (introduced in [24]). In this data the vocabulary consists of  $V = 25$

<sup>1</sup>The code is available on <https://github.com/OlgaIsupova/dynamic-hdp>

<sup>2</sup>It is available on <https://github.com/Blei-Lab/hdp>





Figure 4. QMUL-junction dataset snapshots. (a) is an example of a normal motion, (b) is an example of jay-walking abnormality, (c) is an example of a car moving on the wrong lane in the opposite to normal direction, (d) is an example an emergency service car disrupting a normal traffic flow.

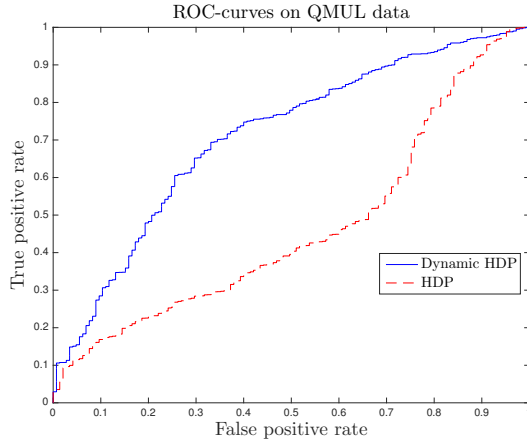


Figure 5. The ROC-curves for the QMUL data.

Table I  
AUC RESULTS

Dataset	Dynamic HDP	HDP	“True” model
Synthetic	0.7118	0.4751	0.7280
QMUL	0.7100	0.4644	—

$\mathbf{k}$ , i.e. it corresponds to the model that can perfectly restore the latent variables. Table I contains the obtained AUC values.

The results show that the proposed dynamic HDP can detect the simulated abnormalities and its performance is competitive to the “true” model. The original HDP method should not detect this kind of abnormalities, as they do not contradict to its generative model, it is confirmed by the experimental results.

words, organised into a  $5 \times 5$  matrix. There are 10 topics in total, the word distributions  $\phi_k$  of these topics form vertical and horizontal bars in the matrix (Figure 2).

The training dataset consisting of 2000 documents is generated from the proposed model (8) – (10), where 1% noise is added to the distributions  $\phi_k$ . Each of the documents has 20 words. The hyperparameters are set to the following values for the generation:  $\alpha = 1.5$ ,  $\gamma = 2$ ,  $\delta = 0.5$ .

Similarly, the testing dataset consisting of 1000 documents is generated, but where 300 random documents are generated as “abnormal”. In the proposed model it is assumed that topic mixtures in neighbour documents are similar. Contrarily to this assumption topics for an abnormal document are chosen uniformly from the set of all the topics except those used in the previous document.

The both algorithms are run for these datasets, computing the abnormality measure for all the testing documents. The hyperparameters  $\alpha$ ,  $\gamma$ ,  $\delta$  are set to the same values as for the generation,  $\eta = 0.2$  ( $\eta$  is not used in the generation as the word distributions in topics are set manually).

In Figure 3 the ROC-curves for the obtained abnormality measures are presented. There is also presented the ROC-curve for the “true” abnormality measure. The “true” abnormality measure is computed using the likelihood given the true distributions  $\phi_k$  and the true table and topic assignments  $\mathbf{t}$  and

### B. Real data

The algorithms are applied to the QMUL-junction real data [6]. This is a 45-minutes video captured a road junction (Figure 4a). The frame size is  $360 \times 288$ . The  $8 \times 8$ -pixel grid cells are used for spatial averaging of the optical flow. For the optical flow estimation the sparse pyramidal version of the Lucas-Kanade optical flow algorithm is used [26] (the implementation is available in the opencv library). The resulting vocabulary size is  $V = 6480$ . Non-overlapping clips, 1-second length, are treated as visual documents. A 5-minute video sequence is used as a training dataset.

The algorithms are run with the following hyperparameters:  $\alpha = 1$ ,  $\gamma = 1$ ,  $\eta = 0.5$ . The weight parameter  $\delta$  for the dynamic HDP is set to 1.

The data is manually labelled as normal/abnormal to measure classification accuracy, where abnormal event examples are jay-walking (Figure 4b), driving wrong direction (Figure 4c), disruption in traffic flow (Figure 4d).

The ROC-curves for the methods are presented in Figure 5. The corresponding AUC values can be found in Table I. The proposed dynamic HDP method outperforms the other one. The provided results show that consideration of dynamics in a topic model may improve the classification results in abnormality detection.

## VII. CONCLUSIONS

In this paper a novel Bayesian nonparametric dynamic topic model is proposed, denoted as dynamic HDP. The Gibbs sampling scheme is applied for inference. The online set up for the inference is designed, allowing to incrementally train the model when the data is processed sequentially. The model is applied for abnormal behaviour detection in video. The abnormality decision rule is based on the predictive likelihood of the data that is developed in this paper. We show that the proposed method, based on the dynamic topic model, improves the classification performance in comparison to the method, based on the model without dynamics. We compare the proposed dynamic HDP method with the method based on the HDP, introduced in [18]. The experiments both on synthetic and real data confirm the superiority of the proposed method.

## ACKNOWLEDGMENTS

Olga Isupova and Lyudmila Mihaylova would like to thank the support from the EC Seventh Framework Programme [FP7 2013-2017] TRACKing in compleX sensor systems (TRAX) Grant agreement no.: 607400. Lyudmila Mihaylova acknowledges also the support from the UK Engineering and Physical Sciences Research Council (EPSRC) via the Bayesian Tracking and Reasoning over Time (BTaRoT) grant EP/K021516/1.

## REFERENCES

- [1] H. Jeong, Y. Yoo, K. M. Yi, and J. Y. Choi, "Two-stage online inference model for traffic pattern analysis and anomaly detection," *Machine Vision and Applications*, vol. 25, no. 6, pp. 1501–1517, 2014.
- [2] J. Varadarajan and J. Odobez, "Topic models for scene analysis and abnormality detection," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, Sept 2009, pp. 1338–1345.
- [3] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.
- [4] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [6] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International Journal of Computer Vision*, vol. 98, no. 3, pp. 303–323, 2012.
- [7] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 1951–1958.
- [8] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin, "Hierarchical Bayesian modeling of topics in time-stamped documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 996–1011, June 2010.
- [9] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*. Corvallis, Oregon: AUAI Press, 2008, pp. 579–586.
- [10] X. Fu, J. Li, K. Yang, L. Cui, and L. Yang, "Dynamic online HDP model for discovering evolutionary topics from Chinese social texts," *Neurocomputing*, vol. 171, pp. 412–424, 2016.
- [11] C. Chen, N. Ding, and W. Buntine, "Dependent hierarchical normalized random measures for dynamic topic modeling," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ser. ICML '12, J. Langford and J. Pineau, Eds. New York, NY, USA: Omnipress, July 2012, pp. 895–902.
- [12] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.
- [13] A. Ahmed and E. Xing, "Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream," in *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*. Corvallis, Oregon: AUAI Press, 2010, pp. 20–29.
- [14] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 81–89.
- [15] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.
- [16] X. Wang and X. Ma, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.
- [17] O. Isupova, L. Mihaylova, D. Kuzin, G. Markarian, and F. Septier, "An expectation maximisation algorithm for behaviour analysis in video," in *Proceedings of the 18th International Conference on Information Fusion (Fusion) 2015*, July 2015, pp. 126–133.
- [18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.
- [19] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.
- [20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.
- [21] K. Vorontsov, O. Frei, M. Apishev, P. Romov, and M. Dudarenko, "BigARTM: Open source library for regularized multimodal topic modeling of large collections," in *Analysis of Images, Social Networks and Texts*. Springer, 2015, pp. 370–381.
- [22] P. Smyth, M. Welling, and A. U. Asuncion, "Asynchronous distributed learning of topic models," in *Advances in Neural Information Processing Systems*, 2009, pp. 81–88.
- [23] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," in *AISTATS*, vol. 2, no. 3, 2011, p. 4.
- [24] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, 2004.
- [25] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [26] J.-Y. Bouguet, "Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, no. 1-10, p. 4, 2001.