Modelling Multi-object Activity by Gaussian Processes

Chen Change Loy
ccloy@dcs.qmul.ac.uk
Tao Xiang
txiang@dcs.qmul.ac.uk
Shaogang Gong
sgg@dcs.qmul.ac.uk

School of EECS
Queen Mary University of London
E1 4NS London, UK

Abstract

We present a new approach for activity modelling and anomaly detection based on non-parametric Gaussian Process (GP) models. Specifically, GP regression models are formulated to learn non-linear relationships between multi-object activity patterns observed from semantically decomposed regions in complex scenes. Predictive distributions are inferred from the regression models to compare with the actual observations for real-time anomaly detection. The use of a flexible, non-parametric model alleviates the difficult problem of selecting appropriate model complexity encountered in parametric models such as Dynamic Bayesian Networks (DBNs). Crucially, our GP models need fewer parameters; they are thus less likely to overfit given sparse data. In addition, our approach is robust to the inevitable noise in activity representation as noise is modelled explicitly in the GP models. Experimental results on a public traffic scene show that our models outperform DBNs in terms of anomaly sensitivity, noise robustness, and flexibility in modelling complex activity.

1 Introduction

Activity modelling and automatic anomaly detection in video have received increasing attention due to the recent large-scale deployments of surveillance cameras. These tasks are non-trivial because complex activity patterns in a busy public space involve multiple objects interacting with each other over space and time, whilst anomalies are often rare, ambiguous and can be easily confused with noise caused by low image quality, unstable lighting condition and occlusion.

Most existing approaches [4, 5, 6, 8, 23, 25] have been devoted to parametric models such as Hidden Markov Models (HMMs), or in a more general case, Dynamic Bayesian Networks (DBNs) due to their effectiveness in modelling temporal dynamics of activity patterns. However, learning a DBN structure with *appropriate complexity* (i.e. the number of hidden states, the state connectivity, and model topology) remains a difficult problem. In particular, most of the automatic model selection criteria [2, 22] assume the availability of sufficiently large training sample size compared to the number of model parameters. Model complexity estimation becomes inaccurate when the training data is sparse or corrupted by noise. Although one can specify a model structure based on prior knowledge, the task can be challenging with surveillance video data as the activity states and dynamics are often not apparent and nor well defined. They also change over time. For a parametric model, once the model complexity is fixed, its expressive power is hampered/limited by the initial model structure. Adjusting model structure complexity on-line is nontrivial for a DBN that requires re-learning new structure and re-estimating model parameters over time.

In this paper, we propose a novel approach based on non-parametric Gaussian Process (GP) models [15] for activity modelling and real-time anomaly detection. A complex wide-area scene (see Fig.1(a)-(d) for example) naturally consists of a set of semantic regions; each of the regions encapsulates different activity patterns which are correlated with each other either explicitly or implicitly. Our approach aims to discover these semantic regions and model non-linear relationships among activity patterns observed from the regions using Gaussian Processes. The understanding of these relationships is crucial in facilitating the detection of subtle anomalies that involve a group of objects, which are hard to detect by observing individual object alone. With the learned models, predictive distribution is computed on each region and compared with the actual observation. Anomaly is flagged if the observation deviates largely from the predictive distribution, which indicates that the learned relationship between different activity patterns is broken.

In the context of complex multi-object activity modelling, our approach has the following advantages compared to the commonly deployed DBNs: (1) Our activity models are not specified *a priori* but instead the model complexity is automatically adjusted based on the distribution and available data [17]. Hence, *the difficult task of adjusting complexity of model structure is avoided.* (2) Our models are less likely to have overfitting problem compared to DBN. Our GP models only require a small number of hyper-parameters for modelling extensive and arbitrary functions. Therefore they are less prone to the overfitting problem [16, 17]. (3) Our models are able to cope with noise explicitly, resulting in superior robustness against the inevitable noise in activity representation. This is crucial for anomaly detection for which distinguishing noise and true anomalies is always challenging. In comparison to the non-parametric approach proposed by Boiman and Irani [3], our method does not suffer from scalability issue as it does not need to store video patches over time for similarity comparison.

Gaussian Process (GP) models have been a popular non-parametric method for performing non-linear regression [21] and classification [20]. They have been recently adopted for solving vision problems such as action recognition [26] and human motion modelling [18]. However, the visual data employed in the existing studies are collected in controlled environments. The data is therefore relatively clean as compared to the more noisy visual features typical in surveillance videos. Such data impose more challenges for regression models such as GP. To the best of our knowledge, this study is the first attempt in using non-parametric GP models for complex activity modelling and anomaly detection in a busy public scene. To better cope with noise, in this work we formulate a new one-step ahead prediction strategy for robust real-time anomaly detection. We also introduce automatic inference on correlation strength among regional activities in a wide-area scene by employing an Automatic Relevance Determination (ARD) [13] covariance function in our GP models. The proposed approach is evaluated using a public traffic scene featured with complex multi-object interactions. Experimental results show that our GP models outperform DBNs for activity modelling and anomaly detection on sensitivity to anomaly, noise robustness and flexibility in learning from scarce training data.

2 Activity Modelling and Anomaly Detection

2.1 Activity Representation

First, a method similar to that in [10, 11] is employed to automatically decompose a complex scene into N regions, $\mathbf{r} = \{r_n | n = 1, \ldots, N\}$ according to the spatial-temporal distribution of activity patterns observed in a training set of video sequences. In particular, the image space is first divided into equal-sized blocks with 8×8 pixels each. Optic flow was computed using Lucas-Kanade model [12] over the whole image space. Flow vectors that are greater than a threshold are deemed as reliable and averaged within each block to obtain local block activity patterns represented using the horizontal and vertical flow components, $\mathbf{u}'_b$ and $\mathbf{v}'_b$, where b denotes the block index. Both $\mathbf{u}'_b$ and $\mathbf{v}'_b$ are 1D vectors computed over time (i.e. time series). Correlation distances are computed among local block activity patterns to construct an affinity matrix, which is then used as an input to a spectral clustering algorithm [24] for semantic scene decomposition (see Fig. 1(e)).

Second, regional activity patterns are represented based on local block activity patterns ($\mathbf{u}'_b$ and $\mathbf{v}'_b$). Specifically, the regional activity in a region r_n is represented by two 1D vectors $\mathbf{u}_n$ and $\mathbf{v}_n$, which are obtained as $\mathbf{u}_n = \sum_{b \in r_n} \mathbf{u}'_b$, and $\mathbf{v}_n = \sum_{b \in r_n} \mathbf{v}'_b$ respectively. To obtain a more compact and precise representation of the regional activity patterns, we compute the average values of $\mathbf{u}_n$ and $\mathbf{v}_n$ at every interval of 50 non-overlapping frames (the interval is chosen to smooth out noise with minimal loss of information). As a result, the averaged regional activity pattern is represented as a bivariant time series: $\bar{\mathbf{u}}_n = (\bar{u}_{n,1}, \ldots, \bar{u}_{n,T})$, and $\bar{\mathbf{v}}_n = (\bar{v}_{n,1}, \ldots, \bar{v}_{n,T})$, where T is the total number of intervals used in the learning process. To further reduce the influence of noise, a logistic function is applied to both $\bar{\mathbf{u}}_n$ and $\bar{\mathbf{v}}_n$, *i.e.* $\hat{u}_{n,t} = (1 + \exp(-\beta \bar{u}_{n,t}))^{-1}$ and $\hat{v}_{n,t} = (1 + \exp(-\beta \bar{v}_{n,t}))^{-1}$, where β is set to 0.5. The final regional activity features are then normalised to have zero mean and unit variance, and denoted as $\hat{\mathbf{u}}_n$ and $\hat{\mathbf{v}}_n$.

2.2 Gaussian Process Activity Modelling

Two GP regression models are constructed for each region to model features $\hat{\mathbf{u}}_n$ and $\hat{\mathbf{v}}_n$ separately. Therefore we have $2N$ GP models given N decomposed regions. The output of a model is activity patterns observed at interval t from the i-th region r_i and the inputs are the activity patterns observed at previous interval $t-1$ from all the other regions $\{r_j | j = 1, \ldots, N, j \neq i\}$. Each GP model is thus a regression model that predicts the activity pattern from each region in the next time interval using activity patterns in other regions observed at present. This regression model is formally defined as $y = f(\mathbf{x}) + \varepsilon$, where $\mathbf{x}$ denotes an input vector of dimension $D = N - 1$ at $t - 1$ and y denotes a one-dimensional scalar output at t. Gaussian Process $f(\mathbf{x})$ is specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. In this study, we assume zero-mean GP; therefore we denote the process as $f(\mathbf{x}) \sim GP(0, k(\mathbf{x}, \mathbf{x}'))$. We assume the noise presented between the output observations and GP as an independent Gaussian white noise $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ with zero mean and variance σ_n^2. Note that $\mathbf{x}$ and y may refer to either one of the two features, *e.g.*, $\mathbf{x} = (x_1, \ldots, x_d, \ldots, x_D)^\top$ with $x_d = \{\hat{u}_{j,t-1} | j = 1, \ldots, N, j \neq i\}$ and $y = \hat{u}_{i,t}$. A training set $\mathscr{D}$ with M observations is denoted $\mathscr{D} = \{(\mathbf{x}_m, y_m) | m = 1, \ldots, M\}$. The column inputs for all M cases are collected into a $D \times M$ matrix $\mathbf{X}$, and the targets form a vector $\mathbf{y}$, so we denote $\mathscr{D} = (\mathbf{X}, \mathbf{y})$.

2.2.1 Covariance Functions

Covariance function plays an important role in GP because it encodes our assumption on continuity and smoothness of the GP function $f(\mathbf{x})$. There are many possible covariance functions [15], such as linear, Matérn, rational quadratic and neural network. As our objective is to model relationships among activity patterns from different regions, we seek for a covariance function capable of capturing the strength of influence among regions of a busy scene. To that end, we consider a covariance function that implements Automatic Relevance Determination (ARD) [13] being suitable to our problem since it can determine how relevant an input is to the prediction. An example of these covariance functions is squared-exponential covariance function, which has the following form

$$k_{\mathrm{SE}}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}'\right)^\top \Sigma \left(\mathbf{x}-\mathbf{x}'\right)\right), \tag{1}$$

where σ_f defines the magnitude. We have $\Sigma = l^{-2}I$ for an isotropic covariance function (a function only of $|\mathbf{x}-\mathbf{x}'|$) [15] and $\Sigma = \mathrm{diag}(\boldsymbol{l})^{-2}$ with $\boldsymbol{l} = \{l_d\}$ and $d \in \{1,\ldots,D\}$ for an ARD-enabled covariance function. The characteristic length-scales $\boldsymbol{l}$ are associated with the relative importance of different inputs to the prediction, *i.e.* the larger the value of a length-scale, the less relevant the input to the prediction. Adopting an ARD-enabled covariance function provides us with insights on the strength of correlations between each pair of regions.

The squared-exponential covariance function is a stationary covariance function which is invariant to translations in the input space. On the other hand, a nonstationary covariance function is more realistic because it allows a model to adapt to functions whose smoothness changes with the inputs [15]. An example is the neural network covariance function [19]

$$k_{\mathrm{NN}}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})\,(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})}}\right), \tag{2}$$

where $\tilde{\mathbf{x}} = (1,x_1,\ldots,x_D)^\top$ and $\Sigma = l^{-2}I$. Both types of covariance functions are considered in our models due to their different strengths.

Given a model with specific covariance function, the fitness of this model to the data can be evaluated using the marginal likelihood conditioned on the hyper-parameters $\boldsymbol{\theta}$, which is given as

$$\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}\right) = -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{M}{2}\log 2\pi, \tag{3}$$

where $\boldsymbol{\theta}$ denotes the hyper-parameters that define this covariance function and the uncorrelated Gaussian white noise, whilst K is the covariance matrix with $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$.

2.2.2 Hyper-parameters Estimation

The free parameters of a covariance function are known as hyper-parameters. Apart from ensuring good predictions, they also provide intuitive interpretation about the properties of the data [15]. To optimise the hyper-parameters, we maximise the marginal likelihood in Eqn. (3) using its partial derivatives w.r.t the hyper-parameters, which is given as

$$\frac{\partial}{\partial\theta_j}\log p\left(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}\right) = \frac{1}{2}\mathbf{y}^\top K^{-1}\frac{\partial K}{\partial\theta_j}K^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_j}\right), \tag{4}$$

where tr denotes the trace. In this paper, the hyper-parameters are first initialised to random values and optimised using conjugate gradient optimiser [9]. To avoid being trapped at poor local minima, multiple random initialisations are performed and the hyper-parameter set that returns the best marginal likelihood is chosen.

2.3 Activity Prediction and Anomaly Detection

Given a test vector $\mathbf{x}_*$ that consists of the past observations at $t-1$ from $N-1$ regions $\{r_j\}$, where $j = 1,\ldots,N, j \neq i$, the one-step ahead predictive distribution of region r_i at t is computed as

$$\begin{aligned} \overline{f}(\mathbf{x}_*) &= \mathbf{k}_*^\top \left(K+\sigma_n^2 I\right)^{-1} \mathbf{y}, \\ \mathbb{V}(f_*) &= k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top \left(K+\sigma_n^2 I\right)^{-1} \mathbf{k}_*, \end{aligned} \tag{5}$$

where $\overline{f}(\mathbf{x}_*)$ is the mean and $\mathbb{V}(f_*)$ is the variance of the predictive distribution, whilst $\mathbf{k}_*$ denotes the vector of covariance between the test vector and the M training cases.

Subsequently, we want to compute a local anomaly score to measure the deviation of the actual observation from the predictive distribution in each region. A straightforward method is to compute the squared residual $\left(y_* - \overline{f}(\mathbf{x}_*)\right)^2$ between the actual observation y_* and the mean prediction at each time interval. However, the squared residual does not take predictive uncertainty into account. We consider a more conservative local anomaly score in the form of predictive log-likelihood defined as

$$\mathrm{score}_n = -\log p\left(y_*|\mathscr{D},\mathbf{x}_*\right) = \frac{1}{2}\log\left(2\pi\sigma_*^2\right) + \frac{\left(y_* - \overline{f}(\mathbf{x}_*)\right)^2}{2\sigma_*^2}, \tag{6}$$

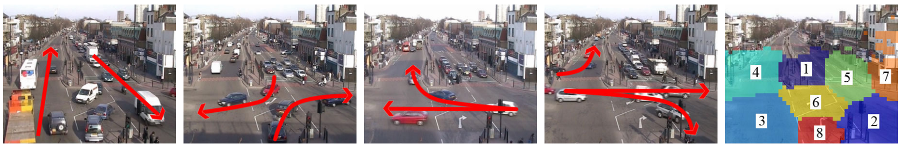
where predictive variance is computed as $\sigma_*^2 = \mathbb{V}(f_*) + \sigma_n^2$. From Eqn. (6), it is clear that a low local score will be obtained if σ_*^2 has a large value. This occurs when the model is uncertain about prediction when the function enters an area with limited training data.

To formulate a global anomaly score, we normalise and compute the sum of the local scores calculated from each region, that is $\mathrm{score}_{\mathrm{global}} = \sum_{n=1}^{N} \overline{\mathrm{score}}_n$ where $\overline{\mathrm{score}}_n$ is the normalised local score at region r_n. For anomaly detection, we set a detection threshold Th according to the detection rate/false alarm rate requirement for specific application scenarios. In particular, an interval (50 non-overlapping frames) is detected as anomaly if $\mathrm{score}_{\mathrm{global}} > \mathrm{Th}$. In each detected interval, a region is identified as being abnormal if $\overline{\mathrm{score}}_n > \frac{1}{N}\mathrm{score}_{\mathrm{global}}$.

Note that although a one-step ahead prediction strategy is employed, multi-step ahead prediction is possible by repeating single prediction iteratively with uncertainty propagated [7]. Though earlier prediction is enalble, multi-step ahead prediction is slower and less reliable than one-step ahead prediction. For reliable real-time anomaly detection, our one-step ahead prediction strategy is appropriate. A common issue of using GP is that the time complexity is $\mathrm{O}\left(M^3\right)$ due to the inversion of $M \times M$ matrix. The complexity can be reduced by adopting Cholesky decomposition [15] instead of inverting the matrix directly. In the case where we have a fixed set of training data, we can perform Cholesky decomposition offline using the training set to enable real-time anomaly detection.

3 Experiments

Datasets - The dataset features a public road junction controlled by traffic lights and dominated by four traffic flows (see Fig.1(a)-(d)). Figure 1(e) depicts the scene decomposition result, showing the eight semantic regions discovered. The length of the video is approximately 60 minutes (89999 frames) captured with 360 × 288 frame size at 25 fps. We used the first 10000 frames of the video (approximately 10% of the data) for training and the rest (79999 frames) for testing. Activity modelling and anomaly detection in this scene is challenging due to: (1) complex interactions among vehicles and pedestrians, (2) changing complexity of activity patterns caused by the changing traffic volume over time, (3) noisy observations caused by illumination change, shadows and video capturing noise.



(a) (b) (c) (d) (e)

Figure 1: A traffic scene dominated by four different traffic flows (arranged in order): (a) vertical flow, (b) left and right turn, (c) rightward flow, (d) leftward flow. The semantic scene decomposition according to the spatial distribution of activity patterns is shown in (e).

Ground truth - Ground truth was extracted from the testing set based on manual labelling, which includes atypical activities such as traffic interruptions by emergency vehicles, illegal u-turn, jaywalking, etc. The ground truth is summarised in Table 1 with examples being shown in Fig. 2. The average duration of anomalies is 4.3 seconds with the shortest interval being 2.64 secs (66 frames) and the longest interval being 8.8 secs (220 frames). Some of the anomalies occurred across several regions, therefore being visually obvious due to the large global visual changes (*e.g.*, cases 3 and 4), whilst others took place in a single region and was supported by very weak visual evidence due to the short occurrence duration, small object size, and severe occlusion (*e.g.*, jaywalking cases and illegal u-turns).

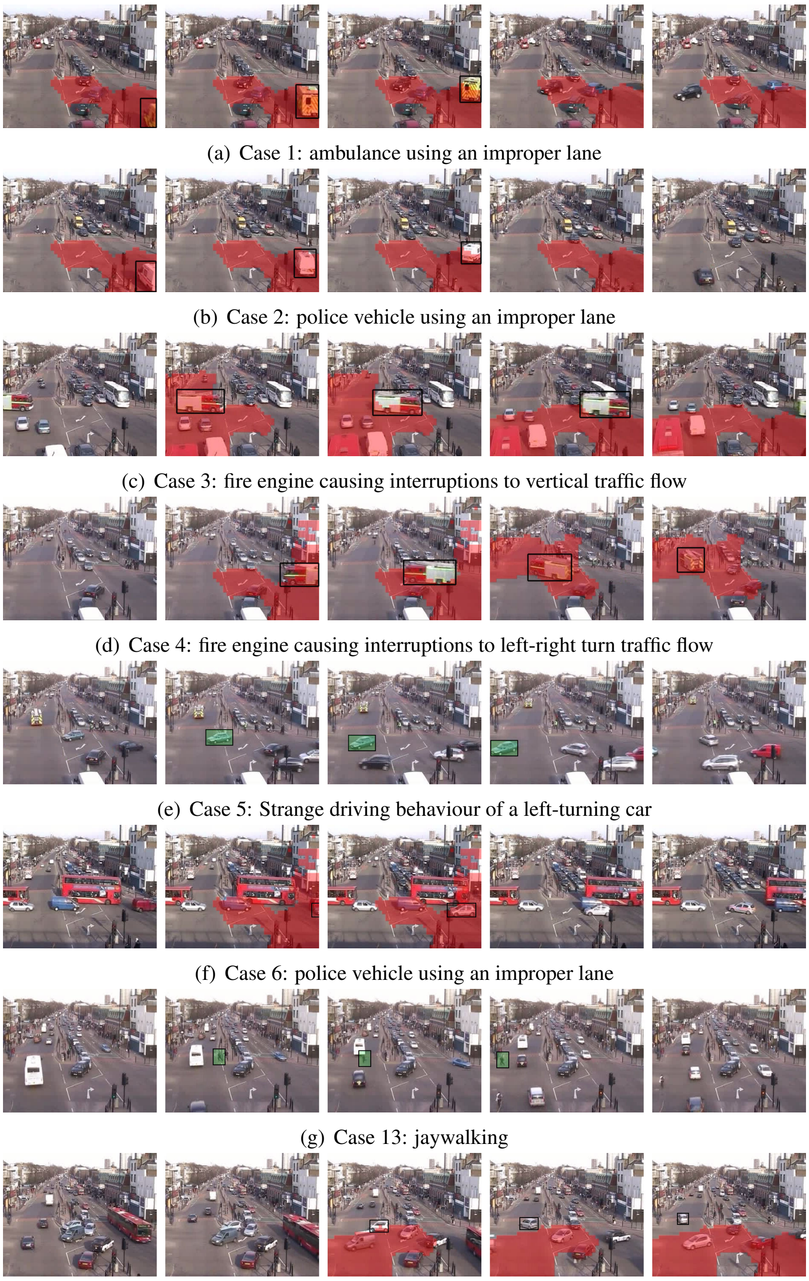
Case(s)	Anomaly description	Total frames (% from total)
1	An ambulance entered the junction using an improper lane of traffic (Fig. 2(a))	88 (0.0978)
2	A police vehicle entered the junction using an improper lane of traffic (Fig. 2(b))	95 (0.1056)
3	A fire engine entered the junction from the left entrance and caused interruptions to the vertical traffic at both directions (Fig. 2(c))	180 (0.2000)
4	A fire engine entered the junction from the right entrance and caused interruptions to the left-right turn traffic (Fig. 2(d))	132 (0.1467)
5	Strange driving behaviour of a left-turning car (Fig. 2(e))	93 (0.1033)
6	A police vehicle entered the junction using an improper lane of traffic (Fig. 2(f))	150 (0.1667)
7-14	Jaywalking (Fig. 2(g))	average: 105 (0.1167)
15-28	Illegal u-turn (Fig. 2(h))	average: 104 (0.1156)

Table 1: Ground truth.

Anomaly Detection using GP models : We experimented with two types of covariance functions (ARD-enabled and neural network) and two anomaly scoring strategies (squared residual and predictive log-likelihood). Receiving Operating Characteristic (ROC) curves were generated by varying the detection threshold Th. The area under ROC (AUROC) was used as performance measure.

The AUROC obtained using different covariance functions and scoring strategies are summarised in Table 2. It shows that predictive log-likelihood performed better than squared residual because the former takes the predictive uncertainty into account. The neural-network covariance function yielded the best result among all the covariance functions with an AUROC of **0.7643**, whilst squared exponential + ARD outperformed the squared exponential covariance function without ARD, indicating the effectiveness of modelling the relevancy of input features with ARD.

The characteristic length-scales $\boldsymbol{l}$ learned when adopting the ARD-enabled covariance function (see Sec. 2.2.1) provides useful insights on the strength of correlations among re-



(a) Case 1: ambulance using an improper lane

(b) Case 2: police vehicle using an improper lane

(c) Case 3: fire engine causing interruptions to vertical traffic flow

(d) Case 4: fire engine causing interruptions to left-right turn traffic flow

(e) Case 5: Strange driving behaviour of a left-turning car

(f) Case 6: police vehicle using an improper lane

(g) Case 13: jaywalking

(h) Case 18: illegal u-turn

Figure 2: Examples of detected anomaly using the GP models with neural network covariance function (abnormal regions are highlighted in red colour whilst the key object is highlighted with a box). The threshold was set to a value to keep the FPR at 0.05. Case 5 in (e) and Case 13 in (g) were missed at this threshold setting (the objects that caused the anomalies are highlighted in green boxes).

Covariance function	Squared residual	Predictive log-likelihood
Squared exponential	0.7351	0.7385
Squared exponential + ARD	0.7509	0.7556
Neural network	0.7464	**0.7643**

Table 2: Comparison of AUROC yielded by different covariance functions and scoring strategies in anomaly detection.

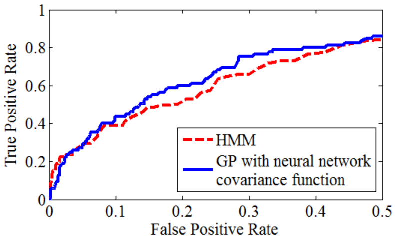
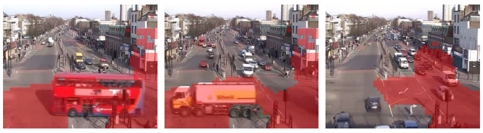


Figure 3: The ROC curves.

gions. For instance, examining the hyper-parameters of the GP model for feature $\hat{\mathbf{v}}$ at region 4 shows that the length-scale of regions 1, 2, 3, 5, 6, 7 and 8 are [4.4509, 3.9672, 0.3717, 5.1920, 4.3222, 1023.5, 3.7875] respectively. The region which has the lowest length-scale is region 3, implying that it has the highest influence on the activity patterns in region 4, whilst region 7 is effectively irrelevant to the prediction. This understanding inferred by our models agrees with the human understanding of the traffic activity patterns in this particular scene. Specifically, vehicles in region 4 typically need to pass region 3 first; the activities in the two regions are thus closely correlated. The regional activities in region 7 are mainly from pedestrians walking on the pavement, which has little relevance on vehicle activities in region 3.

Some examples of detected anomaly using the GP models with neural network covariance function are depicted in Fig. 2. It is observed that most of the anomalies caused by emergency vehicles were successfully detected by the GP models. For instance, the GP models detected the ambulance in Fig. 2(a) that entered the junction using an improper lane, as its activity patterns in region 2 were contrary to the predictive distribution computed using the past observations from other regions. Some of the anomalies such as jaywalking cases are too subtle and difficult to detect due to severe occlusion and small object size.

Figure 4 shows some false alarms. Figures 4(a) and 4(b) are examples of false alarms caused by large objects, *e.g.*, buses moving across the scene. False alarm in Fig. 4(c) refers to activities that are insufficiently captured in the training set. The fact that region 6 was empty deviated from the model's prediction as the vehicles in region 1 were expected to proceed into region 6 when the vertical traffic flow started as captured in the training set. This type of mis-detections is caused by statistical infrequency and can be solved by including more data during the training stage.



(a) (b) (c)

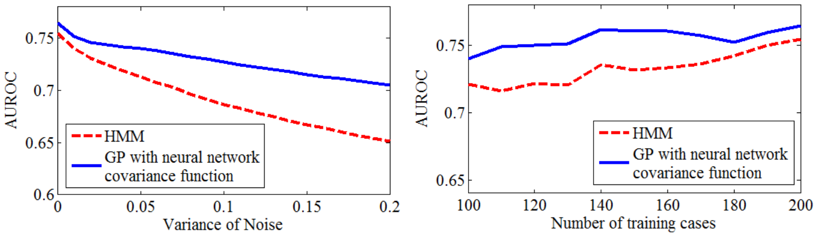
Figure 4: False alarms by the GP models.

Anomaly Detection – GP vs HMM : In this experiment, we compared the anomaly detection performance of our approach against an HMM with continuous observation densities and full covariance matrix. Observation node of each hidden state has 16-dimensional single Gaussian (8 regions with 2 features each) to model the activity patterns from the decomposed regions. The number of hidden states were varied from 2 to 10 with trials on different state connectivities, *i.e.* Bakis (left-right) model and ergodic (fully-connected) model [14], as well as different initialisation strategies (random and uniform initialisation) on the state transitional distributions. The parameters of the observation nodes were initialised using the *k*-means clustering results and estimated using the Baum-Welch algorithm [1]. The log-likelihood of observation computed using the model was employed as a measure of abnormality at each interval.

Through exhaustive testing, a four-state ergodic HMM with random initialisation on the transitional distribution was found to yield the best result among different settings of the HMM. Its performance on anomaly detection is compared with our GP models in Fig. 3, which shows that our models outperform the best HMMs. It is noted that our GP models are more sensitive to anomalies caused by multiple regional activities (*e.g.*, cases 3 and 4). However, the HMM performed better in the cases of illegal u-turning due to the HMM's capability in modelling temporal dynamics. It is worth mentioning that our GP models can be extended to model temporal dynamics explicitly based on the Gaussian Process Dynamical Models proposed in [18].

Sensitivity to Noise – GP vs HMM : We also compared the capability of our GP models with the HMM on handling noisy observations. In this experiment, additive Gaussian noise was introduced and we gradually increased the noise level in the test cases by varying the variance of Gaussian noise from 0 to 0.2. As can be seen from Fig. 5(a), the performance of both methods generally decreased along with the increase of noise variance. However, it is clear that the GP models outperformed the HMM in dealing with noisy observations. The superior robustness of the GP models to noise is mainly due to its capability in modelling noise explicitly.

Sensitivity to Training Sample Size – GP vs HMM : The objective of this experiment is to compare the performance of the GP models and the HMM with different training sample sizes. We varied the number of training samples from 100 to the full 200 cases (each case corresponds to an interval of 50 frames) and observed the trend of the AUROC. As can be seen in Fig. 5(b), the GP models consistently outperformed the HMM with different training sample sizes. Importantly, although the performances of both models decreased, our GP models dropped more graceful, resulting in bigger difference given smaller training sample size. As only 48 free parameters are required for 16 GP models (3 each) used in our approach, compared with 623 parameters needed by the HMM, this result is expected because less parameters means less prone to overfitting given sparse data. It is also found that when the number of cases was less than 140, an HMM with full covariance matrix failed to learn from the data and its covariance matrix had to be switched to diagonal, which essentially assumed that different regional activities were independent. Again, this demonstrates the inflexibility of using a parametric model for activity modelling due to the difficulties in determining optimal model complexity given insufficient data.



(a) Sensitivity to noise (b) Sensitivity to small training sample size

Figure 5: Comparison of the GP models with the HMM in terms of (a) sensitivity to noise and (b) sensitivity to training sample size.

4 Conclusion

We have presented a novel approach for activity modelling and anomaly detection using Gaussian Process (GP) models. In particular, GP regression models are developed to learn the relationship among multi-object activity patterns observed from semantic regions in complex scenes. In addition, a novel one-step ahead prediction strategy was formulated to detect subtle anomalies, with uncertainty being modelled explicitly for a more reliable detection. From our extensive experiments, we demonstrated that the proposed approach outperformed an optimised Hidden Markov Model on both sensitivity to anomaly and robustness to noise. More importantly, the proposed approach avoids the difficult task of adjusting model complexity especially when the training data is scarce.

References

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41(1):164–171, 1970.

[2] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *TPAMI*, 22(7):719–725, 2000.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 74(1): 17–31, 2007.

[4] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *TPAMI*, 22(8):844–851, August 2000.

[5] Y. Du, F. Chen, W. Xu, and Y. Li. Recognizing interaction activities using dynamic bayesian network. *ICPR*, pages 618–621, December 2006.

[6] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-Markov model. In *CVPR*, pages 838–845, 2005.

[7] Agathe Girard, Carl Edward Rasmussen, Joaquin Qui nonero Candela, and Roderick Murray-Smith. Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting. In *NIPS*, pages 529–536, 2003.

[8] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, pages 742–749, 2003.

[9] Nocedal J and S. J. Wright. *Numerical Optimization*. Springer, 1999.

[10] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *ECCV*, pages 383–395, 2008.

[11] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, pages 1988–1995, 2009.

[12] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of Imaging Understanding Workshop*, pages 121–130, 1981.

[13] M. R. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer, 1996.

[14] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, February 1989.

[15] C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning*. MIT Press, 2006.

[16] Carl Edward Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *NIPS*, pages 881–888, 2002.

[17] E. Snelson. *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2007.

[18] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *TPAMI*, 30(2):283–298, 2008.

[19] C. K. I. Williams. Computation with infinite neural networks. *Neural Computation*, 10 (5):1203–1216, 1998.

[20] C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *TPAMI*, 20(12):1342–1351, 1998.

[21] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In *NIPS*, pages 514–520, 1996.

[22] T. Xiang and S. Gong. Optimising dynamic graphical models for video content analysis. *CVIU*, 112(3):310–323, 2008.

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

[24] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.

[25] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted HMMs for unusual event detection. In *CVPR*, pages 611–618, June 2005.

[26] H. Zhou, L. Wang, and D. Suter. Human motion recognition using Gaussian processes classification. In *ICPR*, pages 1–4, 2008.