# Video Anomaly Detection and Localization Using Hierarchical Feature Representation and Gaussian Process Regression

Kai-Wen Cheng, Yie-Tarng Chen, Wen-Hsien Fang
Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.
Email: {D10102101, ytchen, whf}@mail.ntust.edu.tw

Visual analysis of suspicious events is a topic of great importance in video surveillance. Traditional approaches use either object trajectory or local statistics of low-level observations for video anomaly analysis, which is known to be unreliable in unconstrained scenes, and has limited applications. Similar to [2], video event anomalies can be classified as local and global anomalies. A local anomaly is defined as an event that is different from its spatio-temporal neighboring events; whereas, a global anomaly is defined as multiple events that globally interact in an unusual manner, even if any individual local event can be normal. Most research on anomaly detection have focused more on detecting local anomalies such as objects with strange appearance or speed, but less on global anomaly. Global anomalies are common phenomenon in many scenarios like traffic surveillance. The methods in [1, 4] were devised to model the spatio-temporal relationships of dense features with heavy load in space and time, and did not work that well for modeling sparse features.

This paper presents a hierarchical framework, shown in Fig. 1, for detecting local and global anomalies via hierarchical feature representation and Gaussian process regression. As video events can be discriminated from their geometric relations of spatio-temporal interest points (STIPs) in Fig. 2, we propose a unified framework to detect both local and global anomalies using a *sparse* set of STIPs. We identify local anomalies as those STIP features with low-likelihood visual patterns. To deal with inter-event interactions, we further collect an ensemble of the nearby STIP features and consider that an observed ensemble is regular if its semantic (appearance) and structural (position) relations of the nearby STIP features occur frequently. Global anomalies are identified as interactions that have either dissimilar semantics or misaligned structures with respect to the probabilistic normal models.

More specifically, the proposed approach has two main stages to deal with global anomaly. As recognizing global anomaly requires a set of normal interaction templates, we first pose the extraction of normal interactions from training videos as the problem of finding the frequent geometric relations of the nearby interest points. As shown in Fig. 2(b), the proposed extraction method builds a high-level codebook of interaction templates, each of which has an ensemble of STIPs arranged in a non-rigid deformable configuration. Moreover, it can efficiently deal with large training data by utilizing an optimal computation of high-dimensional integral images [5]. We next model the geometric relations of STIP features and propose a novel inference method using Gaussian process regression (GPR) [3]. Given a new sparse set of STIPs, each of which has a relative location $\mathbf{v}_*$ and pattern descriptor $y_*$, the likelihood of being normal with respect to a normal GPR model $\theta$ for each STIP can be defined as follows:

$$p(y_*|\mathbf{v}_*,\theta) = \int p(f_*|\mathbf{v}_*,\theta)p(y_*|f_*)\mathrm{d}f_* \qquad (1)$$

where $p(f_*|\mathbf{v}_*,\theta)$ accounts for the positional distribution which is modeled by GPR, and $p(y_*|f_*)$ captures the appearance similarity which is modeled by i.i.d Gaussian distribution to allow for slight topological deformation. Note that GPR is more suitable on the topic of anomaly detection since it is fully non-parametric and robust to the noisy data and it also supports missing input values like sparse STIPs.

Compared to previous works, our method possesses several advantages: 1) it provides a novel hierarchical event representation to simultaneously deal with local and global anomalies; 2) it employs an efficient clustering method to extract deformable templates of inter-event interactions from training videos; 3) it constructs a GPR model based on a sparse set of STIPs,

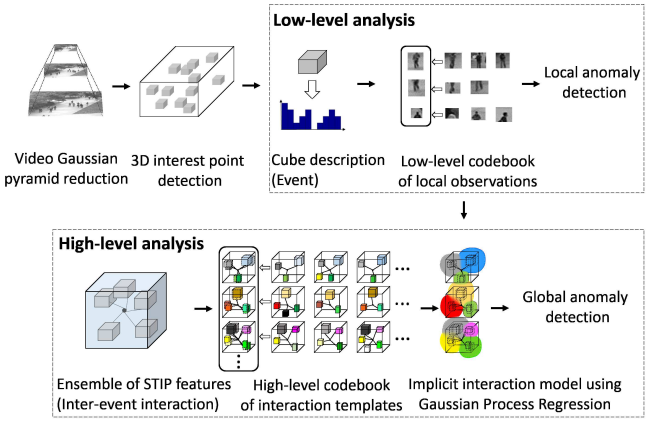


Figure 1: **Overview of the proposed method**

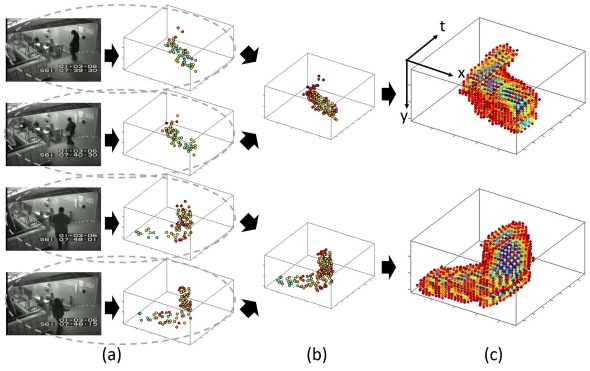


Figure 2: **Complex interaction modeling**: (a) Input videos are represented by a sparse set of interest points. (b) Incidents with similar spatio-temporal relationships of interest points are merged altogether to form deformable interaction templates. (c) Gaussian process regression is used to model each templates. The likelihood of being part of a specific interaction is indicated from low (red) to high (blue). Unlikely locations are invisible for better visualization.

which is not only adaptive based on the available data, it can also learn interactions in a large context while individually locate abnormal events instead of taking an entire interaction as an atomic unit. Note that since our model is built upon sparse STIPs rather than densely-sampled patches [1, 4], the space and time complexity of the event modeling can be greatly reduced (*e.g.*, 150 STIPs v.s. 35301 dense patches in a $41 \times 41 \times 21$ volume), and our method achieves even higher performance. Experiments on three public datasets are conducted and the comparisons with the main state-of-the-art methods verify the superiority of our method.

[1] Oren Boiman and Michal Irani. Detecting irregularities in images and in video. *Int. J. Comput. Vis.*, 74:17 – 31, August 2007.

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *Conf. Comput. Vis. Pattern Recognit.*, pages 3449 – 3456, June 2011.

[3] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[4] Mehrsan Javan Rohstkhari and Martin D. Levine. Online dominant and anomalous behavior detection in videos. In *Conf. Comput. Vis. Pattern Recognit.*, pages 2611 – 2618, June 2013.

[5] Ernesto Tapia. A note on the computation of high-dimensional integral images. *Pattern Recognit. Lett.*, 32:197 – 201, January 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.