Spatio-Temporal Context Modeling for BoW-Based Video Classification

Saehoon Yi and Vladimir Pavlovic
Rutgers, The State University of New Jersey
110 Frelinghuysen Road, Piscataway, NJ 08854, USA
{shyi,vladimir}@cs.rutgers.edu

Abstract

We propose an autocorrelation Cox process that extends the traditional bag-of-words representation to model the spatio-temporal context within a video sequence. Bag-of-words models are effective tools for representing a video by a histogram of visual words that describe local appearance and motion. A major limitation of this model is its inability to encode the spatio-temporal structure of visual words pertaining to the context of the video. Several works have proposed to remedy this by learning the pairwise correlations between words. However, pairwise analysis leads to a quadratic increase in the number of features, making the models prone to overfitting and challenging to learn from data. The proposed autocorrelation Cox process model encodes, in a compact way, the contextual information within a video sequence, leading to improved classification performance. Spatio-temporal autocorrelations of visual words estimated from the Cox process are coupled with the information gain feature selection to discern the essential structure for the classification task. Experiments on crowd activity and human action dataset illustrate that the proposed model achieves state-of-the-art performance while providing intuitive spatio-temporal descriptors of the video context.

1. Introduction

Video classification is a fundamental task in computer vision. As smart phones and hand-held cameras become prevalent, the number of videos people upload and share via websites such as *YouTube* increases tremendously. In order to provide automatic video search by contents, video annotation or classification play a critical role. However, the problems become challenging because traditional video descriptors often display high intra-class variability. For human action or crowd activities, different instances of videos within the same activity class may show significantly different behavior and motions. In addition, the visual appearance and motion changes dramatically if the camera view is changed. Learning commonality within a class for complex and diverse video is the crucial goal of video classification.

Precise detection and modeling of the human pose or detection of objects in video as the means and representation for video classification is typically not possible due to low resolution of the video and varying camera views. Additionally, pose estimation is computationally expensive and often of inconsistent quality, depending on the background clutter or occlusions. To circumvent these problems, Laptev et al. [7] introduced an efficient representation of a video based on spatio-temporal interest points (STIP) [6]. Laptev's STIP detector extends the Harris corner detector to a 3D video volume.

A number of other STIP representations have been introduced as alternatives to [6]. Dollar et al. [5] proposed STIP detector combining 2D Gaussian filter in space with 1D Gabor filter in time. Chen and Hauptmann [3] extended SIFT descriptor by introducing SIFT-like feature describing local motion which has scale invariance property. While these methods detect sparse interest points, Wang et al. [16] proposed extracting dense trajectories to represent video context.

Most STIPs, however, describes appearance and motion only within a local 3D video sub-volume. Laptev [7] describe neighboring video volume by first partition them to small cells and represent with histogram of gradient (HoG) and histogram of optical flow (HoF) in each cell. Wang et al. [16] introduced motion boundary histogram (MBH) which tends to be more robust to noisy camera motion, in addition to HoG and HoF features along the trajectories.

To describe the global video volume, most STIP-based models adopt a bag-of-words (BoW) representation. BoWs are typically non-parametric models, such as histograms, that attempt to encode the distribution of discrete codewords (STIP-types) constructed from STIPs. These representations are simple and enable a surprisingly good empirical video classification tool.

However, the main drawback of the BoW model is that it disregards the spatio-temporal structure i.e., the context within a video. To resolve the problem, Laptev [7] adapted

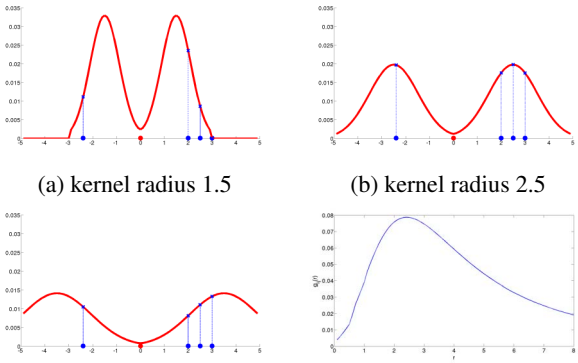
the spatial pyramid matching to a video representation by dividing the 3D video volume. Niebles et al. [12] model the temporal structure of complex actions by decomposing them into simpler action segments, with each segment modeled as a BoW. Bregonzio et al. [2] detect a region of interests(ROI) in each image and model the size and motion of ROIs over subsequent image frames.

Savarese et al. [14] introduced modeling of pairwise correlation between visual words to represent contextual dependencies in the spatio-temporal structure. To accomplish this, they modeled the co-occurrence of all possible pairs of STIP words by a family of cubic-shaped kernels centered on each word. Kernels of different shape and size produced spatio-temporal (ST) correlograms as descriptors of the word "interactions" within a video volume. Although this representation provides additional contextual information, it also suffers from high dimensionality of the newly introduced features. [14] defined a concept *correlaton*, a clustering of ST-correlograms into their own discrete groups remedy this problem. However, performance of this representation on complex video data has not been satisfying, in part because of the loss of information due to the second clustering step.

We propose an autocorrelation Cox process model that encodes spatio-temporal context in a video. In contrast to the ST-correlograms that represent cross correlation between a pair of visual words, the autocorrelation Cox process model takes advantage of a compact representation while focusing on spatio-temporal autocorrelations within independent visual words.

Our contribution is two-fold:

- We propose a novel remedy to quadratic explosion in the feature dimension, which originates from pairwise analysis of visual words in Cox models. [11] reduces the feature dimension by clustering pairs of visual words into meta-clusters. A critical limitation of that approach is that meta-clusters do not rely on spatio-temporal information of individual STIPs in the pair, but only considers their correlation patterns. Since individual STIP information is important in activity analysis problems, our approach retains this information while reducing the feature dimension from quadratic to linear in the number of visual words. Moreover, [11]'s approach of parametric fitting the correlation function is not applicable to the spatio-temporal domain where STIPs occur sparsely in space and time unlike the interest points in texture images. We retain non-parametric forms of correlation instead of [11]'s exponential approximation. Finally, we show that in the case of video data it is more beneficial to retain only the auto-correlation patterns (AutoCox) rather than the cross-correlation texture patterns in [11].



(a) kernel radius 1.5 (b) kernel radius 2.5

(c) kernel radius 3.5 (d) correlation function

Figure 1: Example of how Cox process correlation function is estimated in 1D space. (a)-(c) Kernel weights for corresponding radius, which contribute to correlation estimation (d) Cox process correlation function is estimated over different radii

- We proposed an approach that filters out non-discriminative Cox components using information gain. Texture images exhibit repetitive dense patterns and [11] considers all STIPs as contributing to classification. This is not the case in video where, among many possible between-STIP correlation patterns, only few contribute to discrimination of activities. A data-driven pre-filtering of correlation patterns is thus essential for good generalization performance of Cox models in the video domain.

The rest of the paper is organized as follows. Our proposed autocorrelation Cox process model is described in Section 2. Related work are reviewed in Section 3. The model is evaluated and compared to state-of-the-art in Section 4. Finally, discussions are presented in Section 5 followed by the conclusion in Section 6.

2. Spatio-temporal context model

2.1. Univariate Cox process

Cox process X is a point process defined on a locally finite subset $S \subset \mathbb{R}^2$. The intensity Λ of Cox process X follows from stochastic process. If intensity Λ is spatially constant, the Cox process follows homogeneous Poison process. Møller et al. [10] proposed a Log Gaussian Cox Process (LGCP) to model the spatial point process. The intensity process Λ of LGCP follows the log Gaussian process:

$$\Lambda = \{\Lambda(s) : s \in \mathbb{R}^2\}, \tag{1}$$

$$\Lambda(s) = \exp\{Y(s)\}, \tag{2}$$

$$Y \sim \mathcal{N}(\mu, \sigma^2) \tag{3}$$

Summary statistics of the Log-Gaussian Cox process X with intensity Λ are defined by the first and second order moments. Møller et al. [10] suggest efficient nonparametric estimation of the mean intensity ρ and the correlation function $c(r)$ for a univariate Cox process X:

$$\hat{\rho} = n/A(S), \tag{4}$$

$$\hat{c}(r) = \log \hat{g}(r), \tag{5}$$

$$\hat{g}(r) = \frac{1}{2\pi r \hat{\rho}^2 A(S)} \sum_i \sum_{j \neq i} k_h(r - ||x_i - x_j||_2) b_{ij}, \tag{6}$$

$$k_h(a) = \frac{3}{4h}(1 - a^2/h^2)\delta(a),$$

$$\delta(x) = \begin{cases} 1 & \text{if } -h \leq x \leq h \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

(4) estimates the mean intensity, where n is the number of points observed in S and $A(S)$ is a surface area of the plane $S \subset \mathbb{R}^2$. (5) estimates autocorrelation as a function of radius r. In (6), x_i and x_j are the coordinates of points i and j sampled from X and b_{ij} is the proportion of circumference of the circular kernel centered at x_i within S. In (7), h defines the kernel width that controls smoothness of the correlation function. Møller et al. [10] selected Epanecnikov kernel and also listed different possible choices for the kernel function.

An example of the Cox process correlation model is depicted in Figure 1. The univariate point process is shown on the horizontal line. For each radius, different kernel weights are calculated as a function of the distance from the red point x_i to the blue points x_j. Figure 1 (d) shows the correlation function of this univariate Cox process. We can see that the function peaks at radius 2.5 which is close to the average pairwise distance between all points.

2.2. Autocorrelation Cox process

In this section, we propose the autocorrelation Cox process (AutoCox) that encodes spatio-temporal context in a video.

Each video $n \in N$ is represented by K visual word point processes. Point $\mathbf{x}$ consists of (l, v, h, t) : the visual word label l, vertical location v, horizontal location h, and the frame number t. For each point process X_k, the univariate Cox process $\hat{g}_k$ is estimated as described in section 2.1. $\hat{g}_k$ represents the spatio-temporal distribution of X_k. Our goal is to learn a common structure of each visual word within a video class.

Correlation estimates of the univariate Cox process for all visual words are taken as the input feature of each video. In order to infer the autocorrelation structure relevant for classification, we adopt the information gain feature selection principle. From an initial pool of features D with C classes, information gain of each feature $\mathbf{d_f}$ is calculated as (8). Feature value of $\mathbf{d}_f$ is partitioned into V bins. A set $\{\mathbf{d}_f = i\}$ is the subset of $\mathbf{d_f}$ which falls into the ith bin. In (9), $p_j(D)$ is the class prior probability for class $j \in C$.

Algorithm 1 Autocorrelation Cox process

Require: Set of point processes $X = \{\mathbf{x}_i : (l_i, v_i, h_i, t_i)\}$

1. Estimate univariate Cox process
for all video $n \in N$ **do**
  for all point process $X_k, k \in K$ **do**
    for all radii $r \in R$ **do**
      Estimate $\hat{g}_k(r)$ using (6)
      $\hat{g}_k \leftarrow \hat{g}_k / \sum \hat{g}_k$
    end for
  end for
  Feature D = $\{\mathbf{d}_n : (\hat{g}_{n1}, ..., \hat{g}_{nK})\}$
end for

2. Feature selection
for all feature $f \in |\mathbf{D}|$ **do**
  Calculate information gain using (8)
  Feature $D_{IG} = \{\mathbf{d}_f | \text{IG}_f > \text{Thres}\}$
end for

3. SVM classification
$\min \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \right\}$
subject to $y_i(\mathbf{w} \cdot \mathbf{d_f^i} - b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; i \in N$

$$IG(D, \mathbf{d_f}) = H(D) - \sum_{i=1}^{V} \frac{|\mathbf{d}_f = i|}{|D|} H(\mathbf{d}_f = i), \tag{8}$$

$$H(D) = -\sum_{j=1}^{C} p_j(D) \log p_j(D) \tag{9}$$

Features with information gain higher than the threshold are selected as input features to an SVM classifier. Algorithm 1 summarizes the AutoCox modeling approach for video classification.

3. Related work

Nguyen et al. [11] proposed a new representation of texture images modeled by multivariate log-Gaussian Cox processes, with the goal of classifying different types of textures. First, interest point detectors are applied to texture images and each point is labeled by a visual word from vocabularies constructed using the K-means clustering algorithm. Then, they used the non-parametric estimation of a multivariate Cox process proposed by [10] to model the 2D spatial correlation between pairs of visual words.

Nguyen et al. [11] represent a texture image with the Cox process intensity $\hat{\rho}$ and the correlation function $\hat{c}(r)$. In K-

variate log Gaussian Cox process, $\hat{\rho}$ is $1 \times K$ vector and $\hat{c}(r)$ is $K \times K \times R$ tensor, where R is the number of sampled radii. Rather than use the nonparametric model of the correlation, they fit an exponential profile $exp(-(r/\beta)^{\alpha})$ with two parameters α and β.

In order to resolve the quadratic increase of the number of features induced by pairwise relationships, [11] proposed hierarchical clustering to group pairs of visual words into a new codebook of size $K^* \ll K^2$ and represent the Cox process by a $3K^*$ dimensional vector including K^* density and $2K^*$ parameters of the exponential profile. This representation is subsequently coupled with an SVM classifier to detect different types of natural textures.

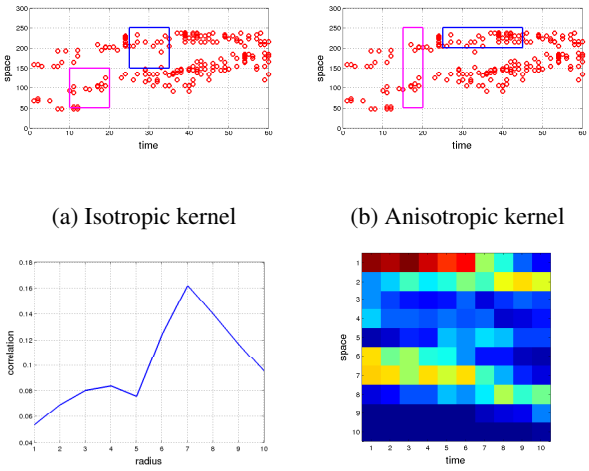
The same approach can be extended to video data. Bivariate Cox process can be considered to model the spatio-temporal point process of STIPs. x_i and x_j in (6) will now represent the spatial location and the frame number as described in 2.2. Isotropic or anisotropic kernel on the spatio-temporal dimension can be used. $||x_i - x_j||_2$ is the Euclidean distance in the spatio-temporal 3D space. However, the dimensionality reduction technique which essentially reduces the number of visual words used in [11] will typically lack descriptive power in video.

3.1. Spatio-temporal Correlaton

A model related to the proposed Cox process is presented in [14]. Savarese et al. [14] proposed a spatio-temporal correlation model that counts the visual word co-occurrence using cubic shape kernels. This model can be seen as a special case of the Cox process where the correlation function estimation in eq (6) is replaced by (10), where r are radii and h is the width of the kernel.

$$g_h(r) = \sum_i \sum_j \delta(||x_i - x_j||_2 - r),$$
$$\delta(x) = \begin{cases} 1 & \text{if } -h \leq x \leq h \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

In [14], in order to effectively handle the quadratic dimensional problem, the authors propose to cluster the estimated pairwise correlation profiles. They call such cluster centers the spatio-temporal correlatons(ST-correlatons) and represent a video by a histogram of ST-correlatons, in addition to the unigram histogram of visual words. This reduces the video representation from the $K \times K \times R$ size correlation tensor to a $K + K^*$ histogram vector. However, this representation of correlation structure in the video disregards the identity information of the two visual words that originally produced the ST-correlaton. The authors claim that such clustering works despite of information loss because it is robust to geometric transformations induced by the pose change. However, our experiments indicate that the proposed clustering approach fails to achieve strong classification performance.



(a) Isotropic kernel (b) Anisotropic kernel

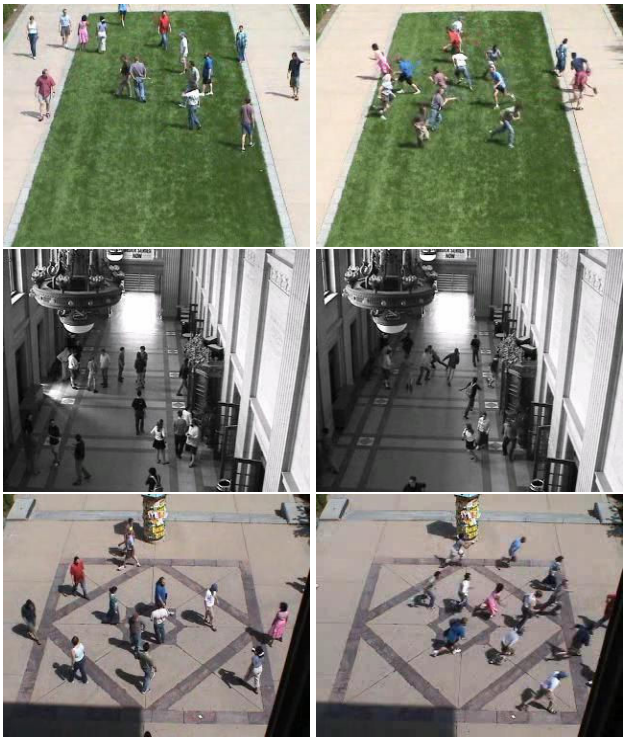
(c) Isotropic correlation (d) Anisotropic correlation

Figure 2: Example of the AutoCox using isotropic and anisotropic kernels. Anisotropic kernel has capability to describe spatio-temporal correlation independently. (a), (b) Point process from top-view. The vertical and horizontal axis indicate space and time dimension. (c) Kernel in blue can capture more points. As a result, higher correlation is expected to temporal dimension. (d) confirms high correlation from short to mid temporal radii on the first row. High correlation is represented by red color.

4. Experiments

We analyze the autocorrelation structure from the Cox process on two different video classification tasks. The two tasks are abnormal activity detection and human action classification.

For all datasets, local motion and visual appearance feature are extracted using the Dense Trajectories Video Descriptor [16]. This descriptor extracts dense trajectories of frame length 15 from optical flow and provides 4 different channels of features: normalized trajectory, histogram of gradient (HoG), histogram of optical flow (HoF) and motion binary histogram (MBH). We chose the dense trajectories descriptor over sparse spatio-temporal interest points because dense sampling typically provides more meaningful and versatile features for low resolution videos. Thus, it is well matched for analyzing point process distributions. Once the descriptor is extracted, K-means clustering is applied on each channel to construct the visual words sets and assign all descriptor to its closest center. Each descriptor in a video has four independent labels each corresponding to four feature channels.



(a) Normal behavior (b) Abnormal behavior

Figure 3: Example of UMN dataset

In the next phase, autocorrelation of each visual word is estimated using the Cox process with eq (10). We treat each trajectory as an interest point and we use the first frame location of the trajectory as the point coordinates in the 3D video volume. The correlation function can be estimated for an isotropic kernel in 3D or an anisotropic kernel with independent space and time profiles. Radii settings and the width of the kernel are decided empirically. Illustration of isotropic and anisotropic kernels is shown in Figure 2.

Finally, the discriminative autocorrelation element is selected using the information gain feature selection. From the training set, we calculate the information gain with respect to the class label on each correlation element (K, s, t) triplet where K is visual word index and s and t is spatio-temporal radii. Correlation elements which exceed an information gain threshold are selected as input feature $\mathbf{d}_f^i$ for video i. A Gram matrix between video $\mathbf{d}^i$ and $\mathbf{d}^j$ is calculated from the normalized dot product, $k(\mathbf{d}^i, \mathbf{d}^j) = \frac{\mathbf{d}^{iT}\mathbf{d}^j}{||\mathbf{d}^i||||\mathbf{d}^j||}$ and an SVM model is learned for classification.

For each dataset, we compared performance to state-of-the-art published results as well as the performance of the BoW model. For BoW, the kernel is calculated using a χ^2 distance of histogram.

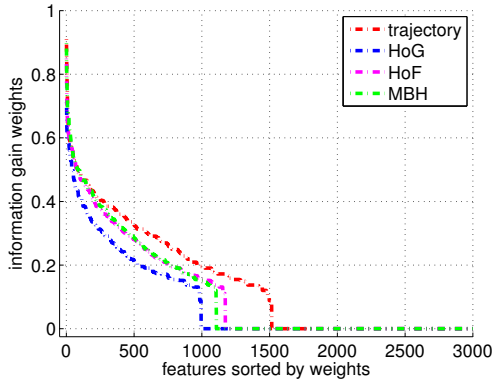


Figure 4: Information gain of autocorrelation features from UMN dataset. Features in each descriptor channel are sorted in descending order of information gain.

4.1. Abnormal group activity detection

The UMN dataset [1] consist of group activities in two different states, normal and abnormal. The video is recorded in three different scenes including one indoor and two outdoor scenes, with a total of eleven video sequences that start as normal state of natural walk and end in abnormal "disperse" motion of the group of people. Examples are illustrated in Figure 3.

As in [9], we also used $K = 30$ visual words on dense trajectories[16]. From eleven video sequences, we take a sliding window of 60 frames as a video clip to model the AutoCox. Each window slides in steps of 10 frames. We chose 60 frames to ensure a sufficient number of points to estimate the autocorrelation for each visual word. We used the first 100, 190 and 130 frames from the three scenes respectively for normal and abnormal videos for training, which corresponds to 8.26% of whole dataset. The rest of video segments are used for testing.

For estimation of the Cox process within each sliding video, we used a cubic kernel in (10) with spatio-temporal radii incremented by 16 pixels from 16 to 160 pixels and the kernel width is also fixed at 16 pixels. Each video is represented by autocorrelation of $g(K, s, t)$, where $K = 1, ..., 30, s = 16, 32, ..., 160, t = 16, 32, ..., 160$ forming a tensor of size $30 \times 10 \times 10$. Autocorrelation for a video is then normalized to 1. Example of autocorrelation with the corresponding point process is illustrated in Figure 7

Information gain threshold is set as 0.21 during the training phase. Correlation features with the gain higher than the threshold are selected. Information gain of autocorrelation values from four different feature channel is shown in Figure 4. On average over four channels, 700 features are selected from the initial pool of 3000 features in each channel. Finally, an SVM model is trained for classification of normal and abnormal sliding windows.

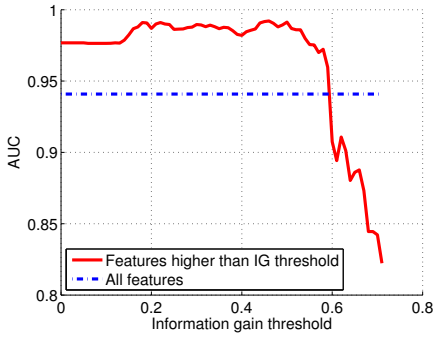


Figure 5: AUC results over different Information gain threshold

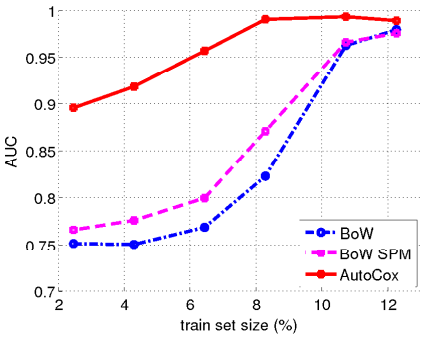


Figure 6: AUC comparison using a different number of training set among AutoCox, BoW, and SPM models

Impact of the information gain feature selection is shown in Figure 5. The blue dashed line is the AUC score of the classifier that uses all 3000 features of the AutoCox in each channel. The red solid line shows the AUC scores as a function of the information gain threshold. Selecting the higher information gain threshold results in selecting the fewer features. It shows that accuracy dropped after choosing the gain threshold over 0.6 because too few features were selected. The figure shows that selecting subset of features actually improves AUC over using all features. This justifies our claim that not all correlation structure is useful for classification.

The abnormality detection results are reported in Table 1. The AutoCox improves the result of the BoW model and achieves the best area under ROC curve(AUC) among state-of-the-art. Our proposed model achieved the same AUC as Wu et al.[17]. However, our model was able to accomplish this in the training set using fewer frames. Wu et al.[17] used 75% of normal clips from six sequences, whereas we used only 8.26% including normal and abnormal clips from all eleven sequences. Achieving high accuracy from smaller training set is critical in the video analysis where extracting and storing large sets of features can be prohibitive and labeling large datasets may be expensive.

AUC comparison among the AutoCox, the BoW, and the BoWSPM models as a function of the training set size is reported in Figure 6. The AutoCox achieves higher AUC than the BoW or BoWSPM models across different sizes of the training set. This result reveals that the structure of the point process distribution of each visual word conveys higher discriminative information compared to the histogram of occurrence frequency that the BoW model uses. In fact, BoWSPM model also incorporates spatio-temporal context in coarse level, which outperforms the BoW model. However, the partition grid of BoWSPM suggested by [7] too coarse to capture the precise spatio-temporal context in videos.

Figure 7 illustrates examples of the point process and the autocorrelation on two videos. The top row shows a video in normal state and the bottom displays abnormal video. In each cell, left figure is showing the point process from the top-view. The vertical axis is space while the horizontal axis indicates time. Right figure displays the autocorrelation of space(row) and time(column). Red inidcates high correlation with blue describing the opposite.

For Visual Word 1, the point process of normal and abnormal video shows similar patterns. From the autocorrelation of visual word 1 from the normal video, we can infer that point processes have strong correlation in short spatial radius and short to medium temporal radius. Nevertheless, the point process concentrated spatially, thus lead to high correlation only on short spatial radius. In fact, points form two clusters where one resides on the top and the other one on the bottom. However, the distance between the two cluster is farther apart than the maximum spatial radius, resulting in no cotribution to the correlation element on the far range spatial radius.

In contrast, for visual word 2, the two videos show very different point process patterns, as reflected in different autocorrelation profiles. Limited to this example, we can infer that visual word 2 will be have higher discriminative power compared to visual word 1.

Table 1: Abnormality detection results on UMN dataset

	AUC
Mehran et al. [9]	0.96
Wu et al. [17]	0.99
Cong et al. [4]	0.97
Saligrama and Chen [13]	0.98
BoW on dense trajectory	0.96
BoW w/ spatial pyramid match	0.97
AutoCox	**0.99**

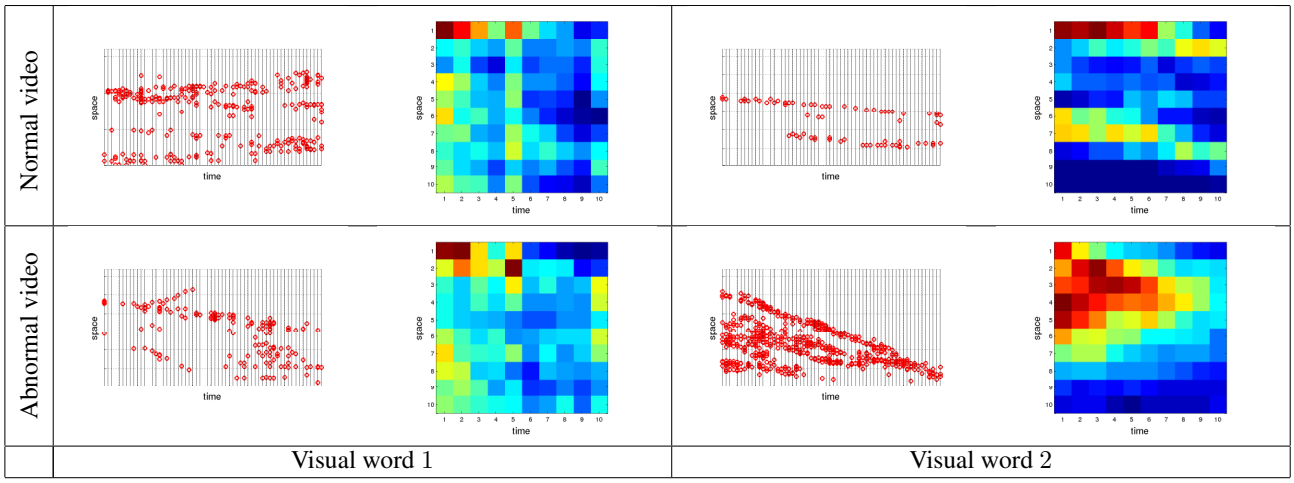


Figure 7: Cox process and correlation function for two instances from UMN group activity dataset

4.2. Human action classification

In this section, the experiment is performed on the YouTube dataset [8]. The dataset consists of 1168 videos from 11 actions of *basketball, biking, diving, golf swing, horse riding, soccer juggling, swing, tennis swing, trampoline jumping, volleyball spiking* and *walking*. Following [8], a 25 fold cross validation was used for classification.

In this experiment, we first calculate a BoW histogram of $M = 2000$ visual words for each video. However, the visual words for $M = 2000$ are oversegmented, making it very challenging to extract meaningful correlation profiles. Instead, we estimate the AutoCox model for $M = 10$ visual words and calculate the Gram matrix as proposed. For spatial and temporal radii of Cox autocorrelation, the same settings are used as in section 4.1. Information gain for each descriptor channel is shown in Figure 8. Information gain threshold is selected as 0.19. Finally, we combine the Gram matrix from the AutoCox model with the that of the BoW with $M = 2000$ model.

We compare the classification result of the AutoCox model with $M = 10$ to the AutoCox with $M = 2000$ visual words and the ST-Correlatons [14]. The ST-Correlaton model[14] estimates the cross correlation of every pair of $M = 10$ visual words. $K^* = 2000$ examplars of cross correlation pattern are clustered. ST-Correlatons histogram of $K^* = 2000$ was concatenated to that of the BoW of $M = 2000$.

Our experiment result of the BoW model achieved 82.88%, which is slightly below what [15] reported. We believe that it is due to the number of visual words they used ($M = 4000$) and the different setting of Dense trajectory extractor. We increased the sampling stride setting of the dense trajectory extractor from 5 to 10, which results fewer trajectories.

Table 2: Classification results on Youtube dataset

Accuracy(%)	
Liu et al.[8]	71.2
Wang et al.[15]	84.1
BoW w/ M2000	82.88
BoW + ST-Correlaton K*2000	79.91
BoW + AutoCox w/ M2000	81.44
BoW + AutoCox w/ M10	**84.23**

In Table 2, the AutoCox model with $M = 10$ achieved the best result. ST-Correlaton degrades the classification accuracy of BoW when it combines histogram of ST-Correlaton with that of BoW. The autoCox model with $M = 2000$ visual words also deteriorates the accuracy of BoW. This result suggests that the meaningful correlation structure can be extracted from using a number of visual words significantly smaller from that of traditional, BoW histograms.

Two remarks can be made from the result of this experiment. First, for complex video datasets which require high number of visual words for BoW model, ST-structure analysis is effective when it is applied on small number of visual words independent to BoW visual words. Second, autocorrelation of ST-structure (AutoCox) still outperforms cross correlation of pairwise visual words (ST-Correlaton), even when small number of visual words are used. With small number of visual words, quadratic feature size does not impose a huge burden. Nevertheless, we showed that autocorrelation of each visual word is capable of modeling discriminative video context than cross correlation.

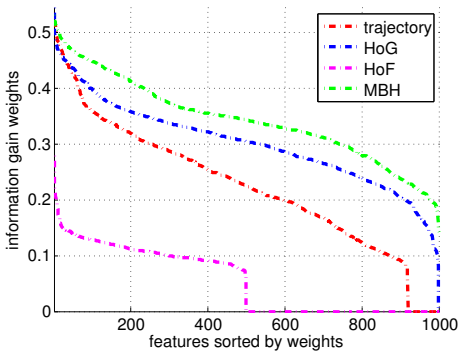


Figure 8: Info gain of features from YouTube dataset

5. Discussion

A major challenge of the pairwise correlation representation for image and video classification is that it suffers from a quadratic explosion in the number of pairwise relationships. If the size of the training set is not sufficiently large, a complex representation of the data is bound to overfit.

The proposed model instead focuses on the autocorrelation analysis which will reduce the number of features from quadratic to linear with respect to the number of visual words in a video. The impact of this representation is two-fold. First, it requires fewer training samples to learn a meaningful structure. Nevertheless, the proposed model is still able to encode the spatio-temporal contextual information important for interpretation of visual activities, which is not present in traditional "unigram" BoW models or is too coarsely represented in pyramid type of representations. Finally, the autocorr representation is computationally efficient, resulting in reduced time as well as space complexity.

In addition, we empirically showed that not all correlation elements are equally important for classification. If a specific pattern of a visual word occurs in the videos of different classes, correlation structure correspond to such pattern will deteriorates classification accuracy. Figure 7 illustrates that this in effect happens.

6. Conclusion

We present a novel method that enables learning the spatio-temporal context in videos without suffering quadratic increase in the number of features. The proposed AutoCox model is used to generate contextual autocorrelation spatio-temporal features, one per each visual word, to describe longer range co-occurrence patterns in space and time. Information gain is then applied to extract meaningful features that are subsequently used to classify visual events and activities. Our proposed model outperforms the BoW and achieved state-of-the-art performance for anomaly crowd activity detection and human action classification problem.

References

[1] Unusual crowd activity dataset: http://mha.cs.umn.edu/movies/crowd-activity-all.avi. 5

[2] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *Computer Vision and Pattern Recognition*, pages 1948–1955, 2009. 2

[3] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. *CMU-CS-09-161, Carnegie Mellon University*, 2009. 1

[4] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *Computer Vision and Pattern Recognition*, pages 3449–3456, 2011. 6

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005. 1

[6] I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, pages 432–439, 2003. 1

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition*, 2008. 1, 6

[8] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *Computer Vision and Pattern Recognition*, 2009. 7

[9] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition*, pages 935–942, 2009. 5, 6

[10] J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox Processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998. 2, 3

[11] H.-G. Nguyen, R. Fablet, and J.-M. Boucher. Visual textures as realizations of multivariate log-gaussian cox processes. In *Computer Vision and Pattern Recognition*, pages 2945–2952, 2011. 2, 3, 4

[12] J. C. Niebles, C.-W. Chen, , and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision*, September 2010. 2

[13] V. Saligrama and Z. Chen. Video anomaly detection based on local statistical aggregates. In *Computer Vision and Pattern Recognition*, pages 2112–2119, 2012. 6

[14] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei. Spatial-Temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion and video Computing, 2008. WMVC 2008*, pages 1–8, 2008. 2, 4, 7

[15] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011. 7

[16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *Computer Vision and Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011. 1, 4, 5

[17] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition*, pages 2054–2060, 2010. 6