# Sparse Representation based Anomaly Detection using HOMV in H.264 Compressed Videos

Sovan Biswas R. Venkatesh Babu
Video Analytics Lab
Indian Institute of Science, Bangalore, India

***Abstract*—In this paper, we have proposed an anomaly detection algorithm based on Histogram of Oriented Motion Vectors (HOMV) [1] in sparse representation framework. Usual behavior is learned at each location by sparsely representing the HOMVs over learnt normal feature bases obtained using an online dictionary learning algorithm. In the end, anomaly is detected based on the likelihood of the occurrence of sparse coefficients at that location. The proposed approach is found to be robust compared to existing methods as demonstrated in the experiments on UCSD Ped1 and UCSD Ped2 datasets.**

***Index Terms*—Anomaly detection, Histogram of Oriented Motion Vectors, Sparse representation**

## I. Introduction

Security is of utmost importance today. With, variety of security issues ranging from bombings to burglary, security personnel have a challenging task in their hands in form of potentially avoiding the catastrophe from ever occurring. One of the key components helping the security personnel in avoiding major catastrophe is video surveillance. Monitoring a crowded area $24 \times 7$ can probably avoid any major attack from happening or in other unfortunate cases, capturing the culprit post attack through careful analysis of the surveillance videos. Even though, these videos helps in analysis but can be sometime problematic as one needs to scavenge hours of video data to achieve different objectives which can range from abnormal behavior detection, face detection, tracking, etc.

Computer vision, at current stage, provides many elegant solutions for the above objectives through suitably automating the scavenging process and reducing the human workload. In case of anomaly detection, the objective is to detect any behavior that distinctively varies from usual crowd behavior in video. Few of the anomalies are shown in Fig. 1. Anomaly detection needs to achieve two major targets : Accuracy of anomaly detection and processing speed. Majority of the existing algorithms [2], [3], [4], [5] work at pixel level information, after complete decoding, but fail to cater the need of real-time processing. Low processing speed can be attributed to decoding computation to get the pixel level information and handling huge amount of data for extracting feature for high level processing. In the proposed work, we tried to effectively boost up processing speed without affecting the overall accuracy of detection by considering the compression parameters used for encoding.

Multimedia technology has seen a drastic boom in last decade. With recent advances, especially in video compression, highly compressed high quality videos are now available at a fixed bit rate. Majority of the recent advances can be accredited to H.264/AVC compression standard [6]. Furthermore, due to the advancements in hardware, modern surveillance cameras are equipped with H.264 encoder [7]. This has tremendously reduced the cost of the cameras and increased the feasibility of large scale surveillance.

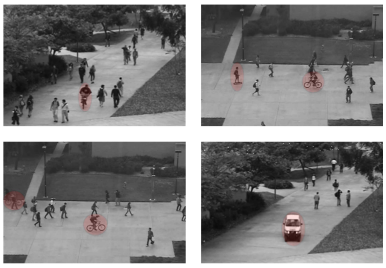

Fig. 1. Few abnormal behavior samples (Marked in red). Anomalies near to camera has different statistics (higher motion magnitude) compared to farther ones.

Aligning to the existing compression standards, the majority of the compression in H.264/AVC is achieved through exploiting temporal redundancy using motion estimation. The motion vectors achieved through motion estimation in H.264/AVC are more accurate than existing standard motion vectors due to use of variable block size motion compensation and quarter-pel motion estimation technique. Variable block-size motion compensation supports motion prediction for $16 \times 16$ to $4 \times 4$ block sizes, enabling better segmentation and motion prediction. Subsequently, half pel and quarter pel motion prediction are used to further improve the accuracy resulting in coarse approximation of optical flow. The proposed algorithm limits itself in using motion vectors for anomaly detection which could be easily extracted through partial decoding that results in huge gain processing speed for large scale video surveillance scenario.

The rest of the paper is divided into five sections. Section II presents few of the related work and discusses their contributions. The extraction of motion feature is explained in section III, which is followed by anomaly detection algorithm in section IV. Later, we presented the experiments and results demonstrating the capability of the algorithm in section V and subsequently, conclude in section VI.

978-1-4799-4665-5/14/$31.00 ©2014 IEEE

## II. Related Work

Many algorithms have been proposed recently in anomaly detection. All the algorithm can be coarsely divided into two major categories, as suggested in [8]: Trajectory based analysis [9], [8], [10] and feature based anomaly detection [2], [3], [4], [5]. Trajectory analysis involves learning the usual behavior pattern through tracking of normal objects/persons and interaction of those tracked objects/person, whereas video feature based analysis involves anomaly detection using the features extracted from a space-time cube. The proposed algorithm computes motion feature using the motion vector to detect anomaly.

Using video features for anomaly detection started with Itti and Baldi [11], who proposed Poisson modeling of feature descriptor computed at every location and detect surprise events. Adam et al. [4] used histograms of optical flows as local monitors to detect anomaly. Kim et al. [3] modeled normal pattern using Mixture of Probabilistic Principal Component Analyzers and later proposed a space-time Markov Random Field to detect anomaly. Spatio-temporal gradient was used by Kartz et al. [12] as feature and later model the usual behavior using 3D gaussian distribution in heavily crowed scene. Mahadevan et al. [2] used Mixture of dynamic textures (MDT) to model normal crowd behavior successfully. More recently, Saligrama et al. [5] assumed anomaly has significant local spatio-temporal signatures that occur for a very small interval and developed a probabilistic framework to detect them.

In case of H.264 compressed video anomaly detection, Biswas et al. [13] exploited motion magnitude using Gaussian modeling alongwith motion pyramid to detect real-time anomaly in high resolution videos. The existing algorithm is capable of detecting unusual behaviors differing only in motion magnitude but fails for anomalies due to motion direction change. In contrast, the proposed approach uses Histogram of Oriented Motion Vector (HOMV) [1] to detect anomaly that captures both motion magnitude and motion direction that makes it robust than the existing methods in compressed video anomaly detection. Additionally, sparse coefficients for candidate HOMVs are computed through a single global dictionary learned over all usual HOMV features, which further improves the detection of anomaly compared to existing methods.

## III. Extraction of Motion Features

H.264 compression standard, like its predecessors, achieves majority of compression due to motion estimation. Motion vectors remove redundancy across consecutive frames to achieve compression.

Biswas et al. [1] defined Histogram of Oriented Motion Vector (HOMV), which captured the motion characteristic in various direction based on MVs. The effectiveness of the HOMV was demonstrated for action recognition through experiments on large scale action datasets. To generate effective HOMV, authors additionally described region of interest (ROI). In this work, we have adapted the HOMV feature to suit the anomaly detection in sparse representation framework.

### A. Pre-processing

MVs extracted from H.264 surveillance videos are usually available for $4 \times 4$ to $16 \times 16$ macroblock. Thus, we first replicate the motion for each macro-blocks to its constituent $4 \times 4$ blocks. Additionally, MVs are noisy as motion estimation is aimed at data compression. Thus, the motion vectors are subjected to space-time median filtering using Eq. (1) that removes the noise. Furthermore, median filtering guarantees motion smoothing for I-frames that lacks motion information. For effective filtering, temporal range is divided into equal amount of past and future frames.

$$d\left[m,n,t\right] = median\{\tilde{d}\left[p,q,r\right], \left(p,q,r\right) \epsilon w\} \qquad (1)$$

where, $d$ and $\tilde{d}$ are the filtered $x$ & $y$ component of MVs and raw $x$ & $y$ component of MV, respectively. $w$ represents a neighborhood centered around location $\left(m,n,t\right)$ in the spatio-temporal cube. As median filtering is a computationally expensive step, we have used a small cube of size $3 \times 3 \times 5$.

### B. Region Of Interest (ROI)

Region of Interest (ROI) correspond to locations which are possible candidates for moving object. The extraction of ROI, proposed in [1], was based on gradient of orientation and magnitude of MVs. Instead, we considered $x$ and $y$ component of MVs, a counterpart to optical flow in [14], to generate ROI. ROI is obtained as:

$$ROI = \left[\frac{1}{k}\sum_{i-k/2}^{i+k/2}\sqrt{\bigtriangledown(d_x)^2 + \bigtriangledown(d_y)^2}\right] > Th \qquad (2)$$

where, $d_x$ and $d_y$ are $x$ and $y$ component of the median filtered MV. $\bigtriangledown$ denotes gradient and $k$ denotes number of frames used as temporal support. ROI computed above capture the boundary of the moving object. Subsequently, holes are filled to obtain final ROI.

### C. Histogram of Oriented Motion Vectors (HOMV)

HOMVs are histograms of MV orientation for a space-time cube, binned on primary angle and weighted according to its magnitude. Unlike [1], space-time cube is defined for a every location by generating a cube of $m \times m \times n$ with current location as center (where $m$ and $n$ defines the spatial and temporal support of the cube respectively). HOMV is generated for each such overlapping spatio-temporal cube (same as in [1]). Figure 2 further illustrates the orientation bins. The feature extraction is summarized in Algorithm 1. Note that the raw HOMVs are normalized by dividing it by number of non zeros motion vectors in the space time cube.

Since, ROI captures moving objects of interest, we limit HOMV extraction only to ROI.

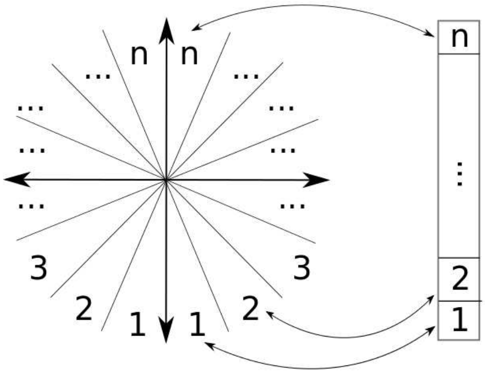

Fig. 2. Orientation Bins : Histograms are binned on primary angle and weighted according to its magnitude

**Algorithm 1** ***HOMV Feature***

```
Input: motion vectors for each Space-Time cube. $n$ = number of orientations, $k$ = number of non zero magnitude MVs in a space-time cube.
Output: HOMV.

  MV = motion vector for Space-Time cubes.
  orientation = $\lfloor tan^{-1}(MV_y/MV_x) * n/\pi \rfloor$.
  magnitude = $\sqrt{(MV_x^2 + MV_y^2)}$.
  initialize : feature $= \mathbf{0}_{1\times n}$

  for all orientation at location $(x, y)$ in $MV$ do
    feature(orientation$(x, y)$) = feature(orientation$(x, y)$) + magnitude$(x, y)$
  end for

  HOMV $= feature/k$
```

## IV. Anomaly Detection

Anomaly is defined as the departure from usual. Mathematically, an event $e$ is defined as $e = \left[f_1, f_2, ..., f_n\right]$ where $f_i$ is a feature for the event $e$. In case of anomaly, $P(e) \leq \tau$, where $P(e)$ is the probability of occurrence of event $e$ and $\tau$ is the decision threshold. Since, an event is conglomeration of different features of the event, $P(e)$ can be defined as $P(f_1 \cap f_2 \cap ... \cap f_n)$. Making an assumption of independence among the different features, it can be rewritten as,

$$P(e) = P(f_1 \cap f_2 \cap ... \cap f_n) \approx P(f_1) \times P(f_2) \times ... \times P(f_n) \quad (3)$$

Thus,

$$P(e) = \prod_{i=1}^{n} P(e_i) \leq \tau \quad (4)$$

Various features have been proposed in the literature ranging from textures of the moving object to spatio-temporal values of a location. But finding effective feature remains the key to detect anomaly.

Recently, sparse reconstruction error was used to detect anomaly by Cong et al. [15]. Instead of using reconstruction error, we explored the coefficient space based on HOMV reconstruction to detect anomaly. Subsequent experiments illustrate that coefficients provide better insight about the underlying motion structure which can effectively be used for solving the problem.

The proposed algorithm models usual behavior and the measure of deviation from this model indicates anomaly. Usual behavior modeling can be divided into two broad stages a) global dictionary training and b) modeling usual behavior through sparse coefficients.

### A. Usual Behavior(UB) modeling

Mathematically, the $l_1$ sparse reconstruction can be defined as

$$\underset{x}{min} \|x\|_1 s.t \|y - Dx\|_2 \leq \epsilon \quad (5)$$

where $y \in \Re^n$ is the input HOMV feature, $D \in \Re^{n \times k}$ is dictionary and $x \in \Re^k$ is the $l_1$ sparse coefficient vector: $\epsilon$ is set to $0.1$. The coefficients are ensured to have positivity constraint.

If two input vectors are similar, then $l_1$ minimization ensure corresponding sparse coefficients to be similar. On the contrary the sparse coefficients in case of dissimilar vectors would depict different coefficient distribution. This forms the basis of anomaly detection. As in case of anomalies the HOMV feature would be very different from usual HOMV features, the model based on coefficient properties can detect anomaly accurately.

This leads to two major issues. First, the HOMV features captured over time for one location in a video frame shall be very different from HOMV features at another location, due to camera position and topology of the field of view. (Refer Figure 1)

Secondly, as the features of one location is different from another, the sparse reconstructing dictionary needs to be large enough to capture all the possible variations.

We tackle both these problems by generating a global dictionary and then modeling the sparse coefficient pattern for each location individually.

*1) Global Dictionary creation:* The global dictionary is required to capture all the possible variation of the HOMVs in a video. We obtain the dictionary using 'Online Dictionary Technique', proposed by Mairal et al [16], that solves Eq. (6).

$$D, x = \underset{D,x}{min} \|y - Dx\|_2^2 + \lambda \|x\|_1 \quad (6)$$

where $x$ is the sparse coefficient vector, $y$ is the set of all HOMV features extracted during training, $D$ is the normalized optimized dictionary and $\lambda$ is the sparsity constraint set to $0.5$. The dictionary is also set to have positivity constraint.

*2) Modeling the sparse coefficient:* The HOMV feature at each location is sparsely represented in the space spanned by the trained dictionary. The strength of the corresponding sparse coefficients for anomaly is very different from usual behavior pattern. Additionally, the distribution of sparse coefficient changes for each location. Typically, the $l_1$ norm value of the coefficients near to the camera are higher than that to the far from the camera due to varying motion magnitude with respect to depth. Hence, we propose modeling the usual behavior densely for each location.

The modeling involves forming histogram of the $l_1$ norm of sparse coefficient for each location. Ideally, just forming histogram of the $l_1$ norm of sparse coefficient should generate the true statistics. But, in reality movements do not occur at all the spatial locations in the video leading to inconsistent and missing statistics at various locations. So, we rectify the histograms by smoothing based on the neighbor statistics using Kernel Density Estimator (KDE) as in [13].

Subsequently, we compute UB probability density function from the modified histograms. In a bid to reduce memory consumption, we propose parametric modeling with a single gaussian density function, characterized by its parameters mean and standard deviation. This tremendously reduce the memory required to store the statistics. Now for each location, the probability distribution is represented by only two parameters.

### B. *Detection of Anomaly*

We already know, the probability of occurrence of an event $e$ depends on its features (Eq. (3)). Since, we are only relying on a single dimensional feature, ie. $l_1$ norm of the sparse coefficients, Eq. (3) is modified as

$$P(e) = P(f) \leq \tau \tag{7}$$

where $e$ is the event and $f$ is the $l_1$ norm of the sparse coefficients. Further, the anomaly is detected based on Eq. (4). The decision threshold $\tau$ is varied to obtain the corresponding ROC curve.

## V. Experiments and Results

In this section, we first introduce the datasets used for the evaluation as well as describe the evaluation procedure. We have conducted experiments on two video databases UCSD Ped1 and UCSD Ped2 to demonstrate the capability of the proposed algorithm. Experiments show on-par accuracy with state-of-the art pixel level techniques but better accuracy compared to existing H.264 compressed video anomaly detection. Since, these data sets were not originally encoded in H.264 format, we have encoded the same in H.264 format using Baseline profile (only I and B frames) with 1 reference frame and Group of Pictures (GOP) length is set to 30. Baseline profile is ideal for network cameras and video encoders since low latency is achieved due to absence of B-frames [7]. All the experiments were performed using MATLAB on single core 3.4 GHz processor.

### A. *Evaluation Procedure*

The anomaly in the UCSD Ped1 and UCSD Ped2 videos can be divided into two general categories a) Non-pedestrians among the pedestrians and b) Pedestrians moving into unusual regions. Detected anomaly can be evaluated in two aspects as mentioned in [2], global anomaly detection and anomaly localization. We have used UCSD Ped1 for both global and localized anomaly detection. On the other side, UCSD Ped2 is only used for global anomaly detection.

| Approaches | Ped 1 | Ped 2 |
|---|---|---|
| SF[17], [2] | 31% | 42% |
| MPPCA[3], [2] | 40% | 30% |
| SF-MPPCA | 32% | 36% |
| MDT[2] | 25% | 25% |
| Sparse[15] | 19% | - |
| LSA[5] | 16% | - |
| **Ours** | **23.43%** | **19.15%** |

TABLE I
Equal Error Rate (EER) of ROC curve on Ped Datasets compared to Pixel Level Processing

Ped1 contains training set of 34 clips compared to 16 clips in Ped2. The testing set consist of 36 clips for Ped1 and 12 clips for Ped2. Ped1 videos are of frame size $238 \times 158$ whereas Ped2 is of size $360 \times 240$.

### B. *Parameters*

The proposed algorithm uses three major parameters, namely, the size of the space-time cubes for which the HOMV feature is extracted, the dictionary size and the decision threshold $\tau$. We have set a space-time cube of $3 \times 3 \times 5$ for HOMV feature extraction. The size of the dictionary plays an critical role in trade-off between detection accuracy and computation complexity. As, the sparse solving is a time consuming process, the dictionary size empirically is set to $10 \times 30$ to optimize the detection accuracy with respect to the time taken in sparse solving. Here, 10 is the length of HOMV feature and 30 is the number of dictionary atoms. $\tau$ is an decision threshold that is varied to generate the ROC curve.

### C. *Quantitative Performance Analysis*

The proposed algorithm results in Equal Error Rate (EER) of 23.43% & 19.15% and AUC in ROC curve of 81.05% & 87.71% on Ped1 and Ped2 respectively with respect to frame-level anomaly.

The detection accuracy are comparable to existing pixel level processing algorithms shown in Tab. I and Fig. 4 respectively. But, the proposed approach performs better with EER of 23.43% compared to 24.66% of Biswas et al. [13] in terms anomaly detection (See Tab. II). The detection speed of the proposed approach is still real-time, even though it is lesser than existing compressed video anomaly approach of Biswas et al. [13]. Additionally, we also compared the proposed algorithm with non-parametric modeling of usual behavior based on orientation and magnitude at individual location.

Localization of anomaly is another important aspect and the proposed approach achieves AUC of 53.07% faring drastically better than most of the existing algorithms as shown in Tab. III only failing to recently proposed uncompressed anomaly detection [18]. Some of the sample outputs are displayed in

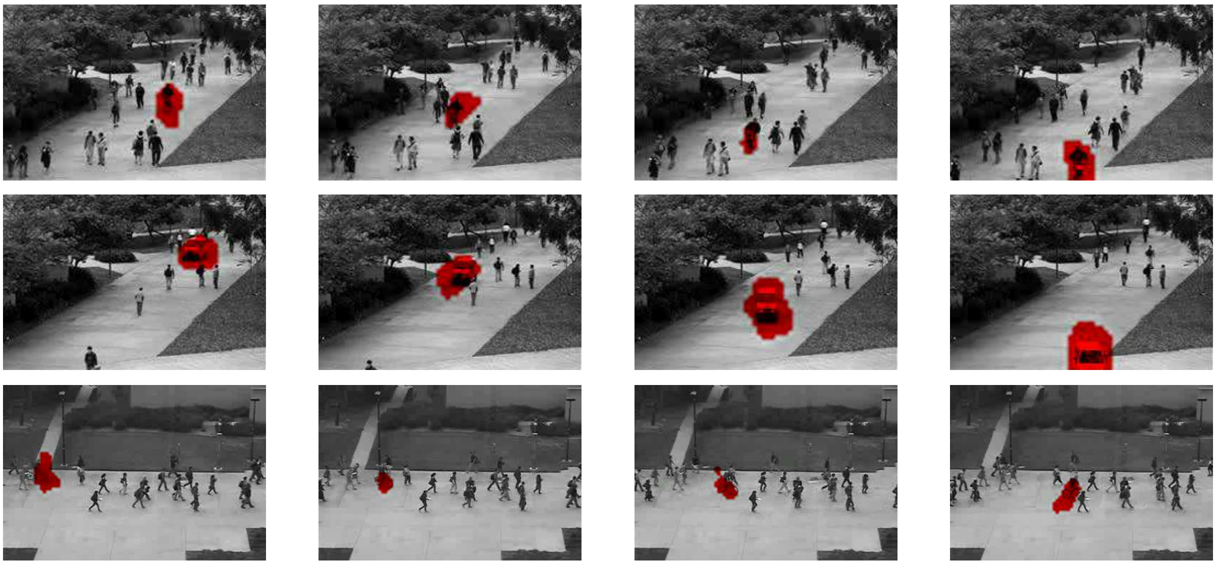

Fig. 3. Few Results on Ped1 and Ped2 dataset. Each Row contains frames of a single video

| Approaches | EER | AUC | Video Resolution |
|---|---|---|---|
| Biswas et. al [13] | 24.66% | 78.70% | 720 × 480 |
| Non-Parameteric Model | 35.99% | 72.39% | 238 × 158 |
| **Ours** | **23.43%** | **81.05%** | **238 × 158** |

TABLE II
EQUAL ERROR RATE (EER), AREA UNDER THE CURVE (AUC) AND VIDEO RESOLUTIONS OF ROC ON PED1 DATASET COMPARED TO EXISTING COMPRESSED VIDEO PROCESSING

| Approaches | RD | AUC | Detection Rate |
|---|---|---|---|
| MDT[2] | 45% | 44% | 0.04 fps |
| Sparse[15] | 46% | 46.1% | 0.25 fps |
| Video Parsing [18] | 68% | 76% | - |
| Biswas et al.[13] | 49.05% | 45.33% | 70 fps |
| **Ours** | **57.16%** | **53.07%** | **26 fps** |

TABLE III
RATE OF DETECTION (RD) AND AREA UNDER THE CURVE (AUC) OF ROC CURVE FOR LOCALIZATION OF ANOMALY ON PED1

Fig. 3. More results are available at: http://val.serc.iisc.ernet.in/sparse_anomaly_results/

## VI. CONCLUSION

In this paper, we have proposed an robust anomaly detection algorithm using H.264 motion vector. The approach uses Histogram of Oriented Motion Vectors (HOMV) [1] as underlying low level feature, that captures both orientation and magnitude of a moving object in a space-time cube effectively. During training, normal variation is learned at each location by modeling the usual behavior of $l_1$ norm of the sparse coefficients represented over a global HOMV feature dictionary. The robustness of the proposed approach was demonstrated through experiments on UCSD Ped1 and UCSD Ped2 datasets.

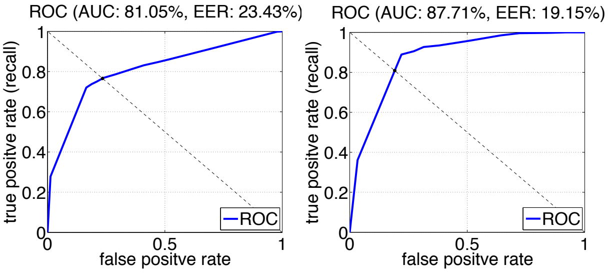

Fig. 4. Frame Level Anomaly on Ped1 and Ped2 respectively

## REFERENCES

[1] S. Biswas and R. V. Babu, “H.264 compressed video classification using histogram of oriented motion vectors (HOMV),” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 2040–2044.

[2] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[3] J. Kim and K. Grauman, “Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2921–2928.

[4] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust real-time unusual event detection using multiple fixed-location monitors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[5] V. Saligrama and Z. Chen, “Video anomaly detection based on local statistical aggregates,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2112–2119.

[6] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H. 264/AVC video coding standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[7] http://www.axis.com/products/video/about_networkvideo/compression_formats.htm.

[8] C. Li, Z. Han, Q. Ye, and J. Jiao, “Visual abnormal behavior detection based on trajectory sparse reconstruction analysis,” *Neurocomputing*, vol. 119, no. 0, pp. 94–100, 2013.

[9] C. Stauffer and W. E. L. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[10] C. Piciarelli, C. Micheloni, and G. L. Foresti, “Trajectory-based anomalous event detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[11] L. Itti and P. Baldi, “A principled approach to detecting surprising events in video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 631–637.

[12] L. Kratz and K. Nishino, “Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1446–1453.

[13] S. Biswas and R. V. Babu, “Real-time anomaly detection in H.264 compressed videos,” in *Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2013.

[14] M. K. Reddy, S. Arora, and R. V. Babu, “Spatio-temporal feature based VLAD for efficient video retrieval,” in *Proceeding of the National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2013.

[15] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3449–3456.

[16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *Proceedings of the Annual International Conference on Machine Learning*, 2009, pp. 689–696.

[17] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 935–942.

[18] B. Antic and B. Ommer, “Video parsing for abnormality detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2415–2422.