REAL TIME ANOMALY DETECTION IN H.264 COMPRESSED VIDEOS

Sovan Biswas R Venkatesh Babu
Video Analytics Lab
Supercomputer Education Research Center Indian Institute of Science
Bangalore, India - 560012

Abstract—**Real time anomaly detection is the need of the hour for any security applications. In this paper, we have proposed a real-time anomaly detection algorithm by utilizing cues from the motion vectors in H.264/AVC compressed domain. The discussed work is principally motivated by the observation that motion vectors (MVs) exhibit different characteristics during anomaly. We have observed that H.264 motion vector magnitude contains relevant information which can be used to model the usual behavior (UB) effectively. This is subsequently extended to detect abnormality/anomaly based on the probability of occurrence of a behavior. Additionally, we have suggested a hierarchical approach through Motion Pyramid for High Resolution videos to further increase the detection rate. The proposed algorithm has performed extremely well on UMN and Peds Anomaly Detection Video datasets, with a detection speed of >150 and $65-75$ frames per sec in respective datasets resulting in more than $200\times$ speedup along with comparable accuracy to pixel domain state-of-the-art algorithms.**

Index Terms—**Anomaly Detection, H.264/AVC, Compressed Domain, Real Time Detection, Video Surveillance**

I. Introduction

Analysis of Crowd Behavior has been point of focus for various vision research groups for decades, due to the increasing importance to physical security. The requirement of crowd behaviors analysis spans from anomaly detection in surveillance scenarios to global crowd pattern analysis for safety planning in highly populated regions, while former being of particular interest to security personnel. As a stitch in time saves nine is the main motive for security personnel, the focus is on real time anomaly detection in surveillance videos. With more and more regions being covered with surveillance cameras, the problem of real time analysis for human personnel becomes challenging as well as error prone.

Computer vision with recent advances is able to solve the potential problem to a major extent with sufficient accuracy. New algorithms are being developed to improve the robustness of anomaly detection [1], [2], [3], [4]. Adam et al. [5] used histogram to characterize optical flow in a patch. In [6], Kartz et al. harness spatio-temporal gradient to detect the abnormality. Kim et al. [7] models local optical flow with Mixtures of Probabilistic Principal Component Analyzers (MPPCA) and enforce the consistency by Markov Random Field. On the other hand, Mehran et al. [8] proposes Social force concept. Mahadevan et al. [1] tries mixture of dynamic textures (MDT) [9] to model normal crowd behavior. Wu et al. [10] tries to utilize particle trajectories to model crowd.

But, most of the algorithms lacks in catering utmost urge of real time detection. As majority of the current algorithms work with fully decompressed videos having pixel level information, there is an additional need of decompressing the video. Along with video decoding from any compressed format (MPEG-2, MPEG-4, etc), huge amount of data and more complex feature extraction process involved in pixel domain leads to slower execution. The problem of decompression overhead is trivial for few short videos, but becomes humongous when dealing with thousands of long duration videos. For example, consider a building being surveilled 24×7 by a single CCTV camera recording at 25 frames per second, results in 24×3600×25 = 2160000 frames per day. Assuming decompressing algorithms achieve 100 to 300 frames per second to decode, one would require 7200 to 21600 seconds ≈ 2 to 6 hours only to decode 24 hours of video!

The focus of this paper is to reduce these computational overheads by performing video anomaly detection in compressed domain. Even though the accuracy is compromised to some extent, this can provide fast initial screening for pixel domain analysis on uncompressed videos for better accuracy.

H.264 [11] is the state-of-the-art video compression standard and widely used due to its high compression efficiency for the given video quality. Similar to any video compression standard, majority of compression is achieved through removing temporal redundancy between neighboring frames, by using Motion Vectors (MVs). MVs contains the shift of macro blocks between candidate and reference frames. Only the motion compensated errors along with MVs are coded, thus reducing the amount of bits to be coded. In crude sense, MVs is considered to be a coarse approximation of optical flow. But, in H.264, MVs are more accurate than any previous standards because of use of variable block-size motion compensation and quarter-pel motion estimation technique. Variable block-size motion compensation supports motion prediction for 16 × 16 to 4 × 4 block sizes, enabling precise segmentation and motion prediction. Half pel and quarter pel motion prediction are subsequently used to improve accuracy of block motion, resulting in nearly optical flow like characteristics.

MVs are often used for different video analysis tasks including video object segmentation, action recognition, etc. In this work, we have explored MVs for detecting anomalous event. MV characteristics, corresponding to anomalous regions, vary significantly. The proposed algorithm tries to capture the mentioned property in a probabilistic manner to detect anomalies.

The rest of the article is divided in 3 sections. Section II begins with explanation of the proposed algorithm in detail followed by experiments and results in section III. The paper is concluded with the conclusion and future work in section IV.

II. Algorithm

Anomaly is defined by the departure from usual characteristics. Mathematically, let $y = [x_1, \ldots x_n]$ be the complete set of features at a particular region, where $x_1,\ldots x_n$ are different features of event y. Then, anomaly is defined to be the event whose probability of occurrence is less than a certain threshold. ie $\mathrm{P}(y) \leq \tau_1$ where τ_1 denotes decision threshold. In other terms, probability of non occurrence of the event is close to 1.

$$P(y) = \prod_{i=1}^{n} P(x_i)$$

$$P\left(\bar{y}\right) = 1 - P\left(y\right)$$
$$P\left(\bar{y}\right) \geq \tau_2 \approx 1 \tag{1}$$

Now, the problem is reduced to extracting relevant features that could discriminate the usual from unusual behaviors. One such feature is motion magnitude, which can effectively distinguish an abnormal event. For example, a person riding a cycle on footpath has different magnitude than that of a person walking. Also, in case of sudden suspicious activity, there is change in crowd movement magnitude than usual.

Motion Vectors (MVs) for H.264/AVC, are defined for a minimum of 4×4 block, which reduces the computation by a maximum of one-sixteenth times than pixel level. But, with increase in video resolutions, computation increase proportionally. Though handling high resolutions video are computationally expensive, it can be processed effectively by utilizing pyramid structure. So, one can perform processing at coarser level (reduced frame size) and in case of ambiguity, move to finer level (actual frame size) to resolve it. This hierarchical processing creates different levels of motion magnitude, which we denote as Motion Pyramids (Fig. 1).

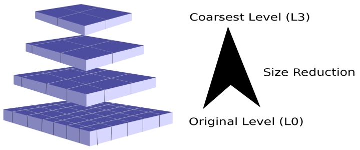


Fig. 1. Motion Pyramid with 4 Levels from L0 to L3

In this article, we tried to tap the local variation in magnitude of MVs (feature) to classify normal and abnormal events. The major modules of the proposed algorithm includes : a) Training Usual Behavior b) Detecting Abnormal Pattern.

A. Training Usual Behavior

The proposed algorithm is trained for usual observations from training videos or from the initial frames of each videos containing usual behavior pattern. Training is performed for each pyramidal level to learn the usual behavior pattern at each level. Subsequently, training is divided into two stages : i) Pre-processing and ii) Usual Behavior Modeling.

1) Pre-processing : Original Level : Since MVs are aimed at reducing the amount of bits required for encoding specific macroblock, it does not always reflect the true object motion. In order to minimize the effect of noisy MVs, a spatio-temporal median filter is applied on the actual motion magnitudes (Eq. 2). For effective filtering, temporal range is divided into equal amount of past and future frames.

$$x\left[m, n, t\right] = median\{\tilde{x}\left[p, q, r\right], \left(p, q, r\right) \epsilon w\} \tag{2}$$

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

Coarser Level (Reduced Size) : Coarse Level is obtained from previous level by spatial bi-linear interpolation of higher resolution raw MV magnitudes. Though, MVs are noisy, interpolation reduces the noise drastically and further filtering is futile. In a bid to have temporal consistency in coarser level, MV magnitude is averaged out temporally at each location.

2) Usual Behavior (UB) Modeling: As behavior is characterized by motion magnitude, UB modeling is done by forming histogram of motion magnitudes. Generally, MVs for the regions which are away from the camera exhibit lower magnitude than those near to camera. Thus, we prefer dense statistic modeling at each location for all the pyramidal levels. Histograms are formed from temporal observation of motion magnitude. But, due to insufficient and inconsistent information at each location in training videos, histograms are noisy and needs to be corrected. Recent research in Approximate nearest neighbor fields (ANNF mapping [12]) based on coherency have demonstrated that two neighboring location tends to exhibit same property with high probability. Therefore, accumulated statistics are further interpolated based on neighbor characteristics. A Kernel Density Estimator (KDE) is applied for interpolation. Since, $P\left(x_{i,j,k\pm a}\right)$ and $P\left(x_{i\pm b,j\pm b,k}\right)$ have effect on $P\left(x_{i,j,k}\right)$ where, $x_{i,j,k}$ denotes motion magnitude at $[i, j]$ spatial location falling in k^{th} magnitude bin. We apply Gaussian KDE (Eq.3) both spatially and across histogram bins.

$$f_h(z) = \frac{1}{n}\sum_{l=1}^{n} K_h(z - z_l) \tag{3}$$

where,
$K_h(z) = (2\pi)^{-k/2}\left|\Sigma\right|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{z} - \mu)^T\Sigma^{-1}(\mathbf{z} - \mu)\right)$,
k= $3, \mu = \begin{pmatrix}0\\0\\0\end{pmatrix}, \Sigma = \begin{pmatrix}\sigma_s^2 & 0 & 0\\ 0 & \sigma_s^2 & 0\\ 0 & 0 & \sigma_b^2\end{pmatrix}$

K_h is 3D Gaussian kernel with σ_s spatial standard deviation and σ_b across histogram range. z being the 3D data.

Histograms are subsequently normalized to obtain UB probability density function. Probability densities are computed densely which consume significant memory for storage and handling. Instead, we propose fitting Gaussian mixture model on the density function, characterized by its parameters. In comparison to original functions, storing these parameters occupies less memory space. Additionally, we propose Gaussian mixture models to counter scenarios where multimodal motion magnitude variation tend to appear. For example, analyzing a dense crowded region, where high motion magnitude is encountered during day followed by no movement at night.

B. Detecting Abnormal Pattern

The use of motion pyramid plays a major role in reducing the computation without affecting the quality. In nutshell, detection is started at coarsest level and moves to finer level only if anomaly is suspected at current level (Fig. 2).

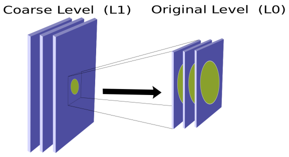


Fig. 2. Detection from Level L1 to L0. Anomaly (Circle) is suspected at Coarse Level L1 is further processed at Finer Level L0

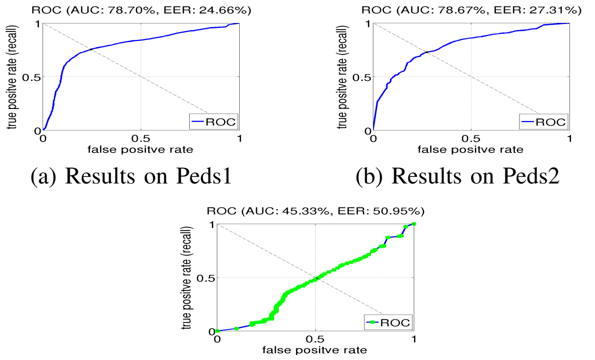
In a bid to increase detection rate in test videos, raw MVs are pre-processed only at coarsest level (similar to training phase) and detection is performed based on probability of occurrence of anomaly at every location, frame-by-frame. Pre-processing and detection at other pyramidal level are delayed till a probable anomalous candidate region is detected at coarser levels. Anomaly is indicated if

$P(\bar{y}_{coarse}) = P(\bar{x}_{coarse}) > \tau_{coarse}$ where, $\bar{y}_{coarse}$ represents a probable abnormal event at coarse level, x_{coarse} denotes the motion magnitude and τ_{coarse} act as decision threshold. In case a candidate is found at a particular location at previous level, pre-processing and detection is performed in and around that location at current level (Fig. 2). Finer level processing is avoided otherwise, resulting in reduction of large amount of redundant computation. At original level, $P(\bar{y}_{original}) = P(\bar{x}_{original}) > \tau_{original}$ indicates abnormality (Ref. eq1) where, $\bar{y}_{original}$, $x_{original}$ and $\tau_{original}$ represents event, motion magnitude and decision thresholds respectively.

This provides a tremendous boost up for high resolution videos as majority of the computation is bypassed through processing at coarser levels. In low resolution videos, processing is only performed at original level as coarser levels fail to capture the motion magnitude variation satisfactorily leading to many mis-detection.

III. Experiments and Results

In this section, we introduce the datasets used for the evaluation as well as describe the evaluation procedure. We conducted experiments on three video databases to demonstrate the capability of the algorithm to handle wide range of variations with accuracy comparable to state-of-the-art techniques. Since, these datasets were not encoded in H.264 format, we encoded the same in H.264 format using Baseline profile with 1 reference frame. Group of Pictures (GOP) length is set to 30 and videos are encoded at a rate of 25 frames per sec. Typically, H.264/AVC encoding is performed on variable block size. So, we replicated the MVs for higher size macroblocks to its 4×4 constituents, resulting in MVs for every 4×4 blocks. All the experimentation were performed using MATLAB on single core 3.4 GHz processor.



(a) Results on Peds1

(b) Results on Peds2

(c) Anomaly Localization on Peds1 with Pixel GroundTruth

Fig. 3. ROC curve for the two datasets

A. Evaluation Procedure

Algorithm is evaluated on two aspects based on global anomaly detection and localized anomaly detection. Two datasets Peds1 and Peds2 (HD videos) are used to test global and localized anomaly detection results, whereas UMN crowd dataset is used for global anomaly detection only.

Peds1 contains training set of 34 clips, where as Peds2 have 16 sets of clips. The testing set consist of 36 clips for Peds1 and 12 clips for Peds2. Anomalies are divided into two categories a) Non-pedestrians among the pedestrians and b) Pedestrians moving into unsuited regions. The aim is to detect abnormality in a frame and localize the regions. Evaluation is performed along the aspects as mentioned in [1] for comparison purposes.

Approaches	Peds 1	Peds 2
SF [8], [1]	31%	42%
MPPCA [7], [1]	40%	30%
SF-MPPCA	32%	36%
MDT [1]	25%	25%
Sparse [13]	19%	-
Ours	**24.66%**	**27.31%**

TABLE I
Equal Error Rate (EER) on Peds Datasets

Approaches	RD	AUC	Detection Rate
SF [8], [1]	21%	17.9%	-
MPPCA [7], [1]	18%	20.5%	-
SF-MPPCA	18	21.3%	-
MDT [1]	45%	44%	0.04 fps
Sparse [13]	46%	46.1%	0.25 fps
Ours	**49.05%**	**45.33%**	**70 fps**

TABLE II
Rate of Detection (RD), Area Under the curve (AUC) and Detection Rate for Peds 1

UMN dataset consist of single video with 11 sequences where an abrupt crowd anomaly occurs at the end of each sequence. The basic intention is to detect abnormal frames. The abnormal frames are marked through tags at the upper left corner.

B. Quantitative Performance Analysis

The proposed algorithm is capable of achieving real-time prediction by compromising some accuracy. An interesting observation is the fact that anomaly detection is erroneous if the reduced size becomes too small. Keeping the same in mind, we used two level pyramid for Peds Dataset, where size is reduced by one-fourth at each level. As UMN dataset contains low resolution videos, we restricted to original level processing. τ_{coarse} set to probability at 3.5σ from μ. The datasets used contains short clips that do not have multimodal magnitude distributions. Thus, we have used single Gaussian for the model.

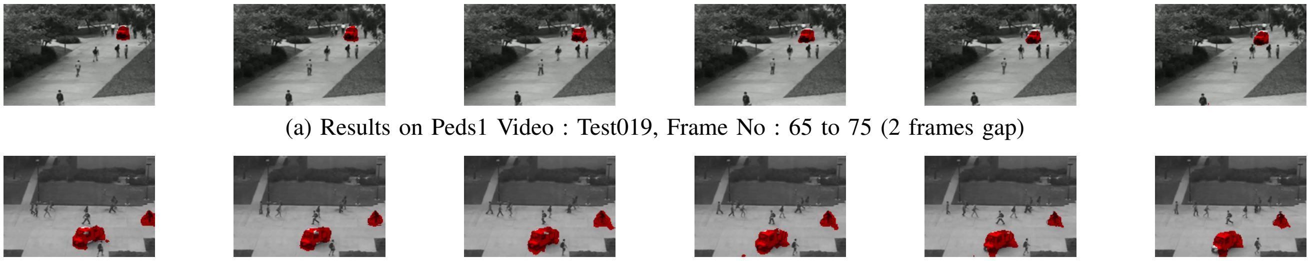
Peds1 and Peds2 : The proposed algorithm has achieved frame-level anomaly detection of equal error rate (EER) 24.66% on Peds1 (Fig.3(a)) and 27.31% on Peds2 (Fig.3(b)), which is comparable to Cong et al. [13] and MDT [1](Table I), with better localization of anomaly than any existing methods. This is demonstrated by detection rate of 49.05% on Peds1 (Fig.3(c)). Refer Tab.II for comparisons. Additionally, it is able to reduce computational complexity resulting in real time detections. We have achieved around 70 frames per sec for 720×480 resolution video using pyramidal approach compared to 32 frames per sec when processing at original size directly.

In comparison to 0.04 frames per sec by MDT [1], and comparison to 0.25 frames per sec by Sparse approach [13], we have achieved around $1750\times$ speedup and $280\times$ speedup respectively. Few of the sample results are shown in Fig.4.

(a) Frame No 525

(b) Frame No 615

Fig. 5. Frames of UMN dataset wrongly marked a) Abnormal Frame marked as Normal Frame b) Normal Frame marked as Abnormal Frame



(a) Results on Peds1 Video : Test019, Frame No : 65 to 75 (2 frames gap)

(b) Results on Peds2 Video : Test004, Frame No : 130 to 150 (4 frames gap)

Fig. 4. Results on Peds1 and Peds2

UMN dataset : Before divulging into experiments on UMN dataset, we observed some interesting points about UMN dataset sequences and its abnormal frame marking. In Fig.5a, though abnormality has already begun but the ground truth indicate otherwise. Additionally, in Fig.5b, the scene is almost empty whereas ground truth indicates the opposite. On evaluating the algorithm to detect global anomaly detection on UMN dataset, we observed the proposed algorithm is able to detect all the abnormality but by a shift, which was expected because of rationale discussed earlier. Since the basic feature extraction of the proposed algorithm depends on the MVs, the algorithm fails to detect anomaly if there is no motion. For training we have used first 200 frames of each sequence and the remaining frames in each sequence for testing. The evaluation is done on each sequence independently. On comparing our results with original ground truth we have achieved ROC curve with area under the curve (AUC) 73.69%, but with corrected ground truth we achieved around AUC of 95.41%. (Refer Fig7). Since, resolution of this dataset is less, we have processed only at finer level. Even then, computationally, we achieved around 150 frames per sec for this dataset (a speedup of 120× compared to Sparse approach [13]). The huge difference in speed between Peds and UMN dataset, is due to latter's low video resolution.

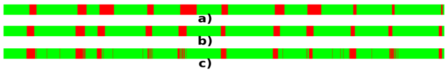
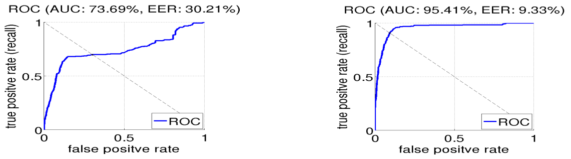


Fig. 6. Result comparison on UMN dataset : Labels of each test frame a) Original GT bar b) Modified GT bar c) Actual Detection bar; Green : Normal frame, Red : Abnormal frame



(a) ROC with Original GT

(b) ROC with modified GT

Fig. 7. ROC curve for UMN datasets a) Based on Original Ground Truth b) Based on Modified Ground Truth

IV. Conclusion and Future Work

We have proposed a compressed domain approach in H.264/AVC framework to detect anomalies in surveillance videos. MV magnitude contains enough information for anomaly detection, which was tapped by profiling of MVs through Motion Pyramid. The method is able to get state-of-the-art results with tremendous reduction in computation time resulting in real time performance. Even though initial results are encouraging, the effect of other compression parameters can be studied further to detect anomalies.

V. Acknowledgement

This work was supported by CARS (CARS-25) project from CAIR, DRDO, Govt. of India and VADS project from DST, Govt. of India.

References

[1] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

[2] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Proceedings of the 2005 IEEE International Conference on Computer Vision*. IEEE, 2005.

[3] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2005.

[4] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[5] 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, pp. 555–560, 2008.

[6] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

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

[8] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[9] A. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 909–926, 2008.

[10] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories," in *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011.

[11] 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, pp. 560–576, 2003.

[12] S. A. Ramakanth and R. V. Babu, "Feature match: an efficient low dimensional patchmatch technique," in *Proceedings of the 2012 Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, 2012.

[13] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.