Detecting Abnormal Behaviors in Crowded Scenes

1Oluwatoyin P. Popoola and 2Hui Ma
1Systems Engineering Department, Faculty of Engineering, University of Lagos, Nigeria
2College of Electronic Engineering, Heilongjiang University, Harbin 150080, PR China

Abstract: Situational awareness is a basic function of the human visual system, which is attracting a lot of research attention in machine vision and related research communities. There is an increasing demand for smarter video surveillance of public and private space using intelligent vision systems which can distinguish what is semantically meaningful to the human observer as ‘normal’ and ‘abnormal’ behaviors. In this study we propose a novel robust behavior descriptor for encoding the intrinsic local and global behavior signatures in crowded scenes. Crowd scenes transitioning from normal to abnormal behaviors such as “rush”, “scatter” and “herding” were modeled and detected. The descriptor uses features that encode both local and global signatures of crowd interactions. Bayesian topic modeling is used to capture the intrinsic structure of atomic activity in the video frames and used to detect the transition from normal to abnormal behavior. Experimental results and analysis of the proposed framework on two publicly available crowd behavior datasets show the effectiveness of this method compared to other methods for anomaly detection in crowds with a very good detection accuracy rates.

Keywords: Abnormal behavior detection, intelligent video surveillance, situation awareness

INTRODUCTION

Situational awareness is a very robust function of the human visual system. This visual system can quickly and efficiently scan through large quantity of time-varying low-level imagery data and pick out information that ‘stand-out’, semantically meaningful and useful for situational awareness. Such ability of the human visual system has inspired a great deal of research in the use of computational modeling techniques for visual analysis of human behavior in the computer vision and machine learning communities. Intelligent visual surveillance systems should be able to provide answers to queries such as ‘what is going on in the scene?’ or ‘what type of event or activity is taking place?’, or ‘where is this action taking place in the scene?’. The visual phenomena although very difficult to model analytically, generally are more effectively approached through statistical machine learning.

The ever increasing demand from governments and commercial business for tools to effectively monitor public and private spaces towards crime detection and prevention, efficient public transport management, as well as personalized healthcare delivery, has fuelled interest and research funding for machine-based visual analysis of behavior. Some of the recent advances in applications include:

- Healthcare delivery using social robots perform automated visual analysis of human behavior in elderly peoples’ home care facilities. These personalized assistants learn from visual information, to mimic and interact with the elderly. They provide routine assessment of the persons’ behaviors and provide real-time assistance under conditions of sickness, accident, or fall.
- Efficient video content search, summarization and indexing is applied to search for certain behaviors in video database which would be tedious and sometimes impossible without meta-data like names, time of event etc. However with automated visual search by object behavior categories such tasks can be done be very rapidly.
- Human Computer Interaction (HCI) systems are developed for more and more user-friendly human computer interaction systems like computer game console applications produced by the gaming industry and teleconferencing systems where visual sensors capture and perform visual analysis on users’ behaviors for intelligent and meaningful communication and interaction.

- Intelligent visual surveillance systems: These electronically perceive and understanding the behaviors of people (or objects moved by humans such as vehicles) in image sequences (frames of videos clips). It is by far the most popular motivation for situational awareness. ‘Behavior’ is a generic term often referring to the observable...
actions of agents such as persons, or other moving objects in the scene. Behaviors become salient when they are different from the regular patterns in that context and do not conform either temporally or spatially; that is they ‘stand-out’ as different relative to the context of their surrounding in space and time.

Such unusual activities are by nature rare, unexpected out-of-the-ordinary. These characteristics make modeling behavior for purpose of detecting abnormal behavior a non-trivial challenge. The objective of this study is towards early detection of contextually abnormal crowd behaviors using a robust behavior representation.

LITERATURE REVIEW

Approaches vary from the ‘holistic’ methods which try to obtain coarser-level (global) information from the main crowd flows for use in detecting anomalies, to bottom-up approaches i.e., from the pixel level. Davies et al. (1995) measure motion features at pixel or pixel-neighborhood levels which are then aggregated to obtain motion properties for larger regions in an image. The aggregated results establish the overall dominant crowd direction and magnitude (velocities). Boghossian and Velastin (1999) worked on how to detect emergency including circular flow around the exits detected in the Hough space which could indicate traffic jams; multi-directional divergence of crowd flow indicating potential danger such as fire outbreak; and obstacles in the flow paths that might correspond to injured pedestrians. Andrade et al. (2006) analyzed crowd motion using optical flow and Hidden Markov Models (HMMs) to characterize “usual behavior” in the crowd. Ali and Shah (2007) proposed a method of segmenting crowd flow, which revealed the crowd’s instabilities as observed from the coherent structures which appear by applying Lagrangian particle dynamics on the optical flow fields. Instability is detected when the number of segments of some analyzed sequence changes. Mehran et al. (2009) used the social force model (Helbing Dirk, 1995) to model crowds for abnormal crowd behavior detection. Particle advection according to the computed optical flow is done to estimate the social forces. Normal patterns of forces over time are modeled by mapping the magnitude of the interaction force vectors to the image plane. The likelihood force flow is estimated to classify each frame as normal or abnormal.

In this study, we adopt an approach that uses the probability of occurrence of feature vectors as indicators of the presence or absence of an anomaly.

The features used to encode behavior can be “global” or “local”, either spatial or temporal or both. Spatial-temporal features have shown particular promise in motion understanding due to its rich descriptive power (Niebles and Li, 2007) and therefore widely used as feature descriptor. Shape dynamics has also been used in Vaswani et al. (2003) and Vaswani et al. (2005) to detect abnormal activity in a group of moving and interacting objects by modeling the changing configuration as a moving and deforming “shape” and used continuous Hidden Markov Models (HMMs) to capture the landmark shape dynamics. Tracking however requires effective background subtraction and is limited by factors like occlusion and shadows; thus restricting it’s utility to encode complex behavior in the real world occurrence of anomalies. It is also sensitive to tracking errors even if they occur in only a few frames. This approach also fails when modeling crowded and complicated scenes.

PROPOSED METHODOLOGY

In this study we propose a new motion descriptor for modeling crowd flows based on local pixel motion from frame to frame. This descriptor is made up of the optical flow, potentials and the Lyaponuv exponent computations on the sequence of moving frames. Low-level local and global spatio-temporal features are used to describe crowd behaviors instead of using the tracking based approaches. The feature descriptor is a vector of optical flow, potential fields and Lyaponuv exponent components in the flow sequences.

Optical flow: This approximates two-dimensional flow field from the image intensities. The vectors are difference in position of the same pixel content between two images and is called flow vector. Details can be found in Brox et al. (2004).

Potentials functions: Drawing from flow concepts in fluid dynamics, we posit that incompressible and irrational flow properties of the optical flow field are constituents of the behavior signature that characterizes the motion types. A simplified mathematical model of fluids assumes that the fluid is incompressible and irrational, based on conservation properties of fluids. These assumptions lead to potential functions, which are scalar functions that characterize the flow in a unique way. Optical flow \(O_F(u,v)\) denotes a planar vector field.

According to Helmholtz decomposition theorem: ‘Let \( F(r) \) be any continuous vector field with continuous first partial derivatives. Then \( F(r) \) can be uniquely expressed in terms of the negative gradient of a scalar potential \( \phi(r) \) the curl of a vector potential \( a(r) \):

\[
F(r) = -\nabla \phi(r) + \nabla \times a(r) \quad \text{or} \quad \quad (1) \\
F(r) = F_1(r) + F_2(r)
\]

where, \( F_1(r) \) is the scalar potential and \( F_2(r) \) is the vector potential, which denote the incompressible and irrational parts of the vector field. An incompressible vector field is divergence free \( \nabla \cdot F(r) = 0 \) and an
irrotational vector field is curl free $\nabla \times F(r) = 0$. These two functions are known as the stream function $\psi$ and the velocity potential $\phi$, respectively. Following (Shandong et al., 2010), we use Fourier transforms to decompose incompressible and irrotational parts of the vector field and estimate the potential functions using:

$$\phi(x, y) = \phi_0 + 0.5 \int_0^X (u_r(s, y) + u_r(s, 0))ds + 0.5 \int_0^Y (v_r(x, s) + v_r(0, s))ds$$

(2)

$$\psi(x, y) = \psi_0 + 0.5 \int_0^X (u_c(s, y) + u_c(s, 0))ds + 0.5 \int_0^Y (v_c(x, s) + v_c(0, s))ds$$

(3)

The stream function provides the information regarding the steady and non-divergent parts of the flow, while the velocity potential encodes information regarding the local changes in the non-curling motions.

**Finite Time Lyapunov Exponents (FTLE):** FTLE is a scalar field that measures how much particles separate after a given interval of time. It describes mass transport in a system based on the vector fields. Any two particles starting in the same region of the exponent field will tend to flow together in time as the dynamical system evolves, whereas two particles straddling a ridge in the field will tend to diverge exponentially in time Shadden et al. (2005). For example, suppose a point located at: $x \in D$ at time $t_0$. When this point is advected by the flow, it moves to $\psi_{t_0}^T(x)$ after time interval $T$. The amount of stretching about this trajectory can be seen by considering a perturbed point $+\delta x(0)$, where, $\delta x(0)$ is infinitesimally small and oriented arbitrarily. After the time interval $T$, the perturbation becomes:

$$\delta x(T) = \psi_{t_0}^T(y) - \psi_{t_0}^T(x) = \frac{d\psi_{t_0}^T(x)}{dx} \delta x(0) + O \left( ||\delta x(0)||^2 \right)$$

for a positive function $f$ means that $f(x)/g(x)$ remains bounded for all $x \in \mathbb{R}$. By dropping the $O(||\delta x(0)||^2)$ terms, the growth of linearized perturbations are obtained. Using Euclidean norm, the magnitude of the perturbation is given by:

$$||\delta x(T)|| = \sqrt{(\delta x(0), \frac{d\psi_{t_0}^T(x)}{dx}, \frac{d\psi_{t_0}^T(x)}{dx}) \delta x(0)}$$

(4)

where, $N^*$ means $N$ transpose. The symmetric matrix:

$$\Delta = \frac{d\psi_{t_0}^T(x)}{dx} \cdot \frac{d\psi_{t_0}^T(x)}{dx}$$

is a finite-time version of the Cauchy-Green deformation tensor. Although $\Delta$ is a function of $x$, $t_0$ and $T$. Maximum stretching occurs between two very close particles when $\delta x(0)$ is aligned with the eigenvector associated with the maximum eigenvalue of $\Delta$. Thus if $\lambda_{max}(\Delta)$ is the maximum eigenvalue of $\Delta$, then:

$$\max_{\delta x(0)} || \delta x(T) || = \sqrt{\lambda_{max}(\Delta)} || \delta x(0) ||$$

(5)

where, $\delta x(0)$ is aligned with the eigenvector associated with $\lambda_{max}(\Delta)$. This can be re-written as:

$$\max_{\delta x(0)} || \delta x(T) || = e^{\sigma_{T}(\delta x(x))} || \delta x(0) ||$$

(6)

where, Eq. (2.20) represents the largest finite-time Lyapunov exponent with a finite integration time $T$, corresponding to point $x \in D$ at time $t_0$.

The semantic descriptive label assigned to each of the crowd scenes as it evolves in the global (frame-level) is considered an ensemble of behavior dynamics aggregated across the local spatio-temporal patches that make up the frames. This implies that the singular behavior of one or a minority of the moving targets does not suffice to infer the global behavior until similar abnormal behavior can be observed to be populated across the scene. Thus the feature vector computed across the patches for all the frames consists of the statistics of low-level features of the flow vectors such as magnitude, potentials (incompressible and irrotational) and the Lyapunov exponent values:

$$f_{patch} = [u, v, \theta, psi, \epsilon]$$

(7)

where, $u, v$ capture motion magnitude, in the velocity field from optical flow; $\theta, psi, \epsilon$ are the incompressible and irrotational flow components from the FTLE potentials and $\epsilon$ is the Lyapunov exponent values indicating the amount of local divergence or convergence.

The feature vectors are clustered using the bio-inspired clustering method proposed in Wang and Popoola (2010) to form a codebook of visual words. With this representation, the original video frames can be discarded and each video clip represented by the visual words from the codebook.

**Modeling with bayesian topic models:** Variants of the basic topic model shown in Fig. 1 namely the Latent Dirichlet Allocation (LDA) (Blei et al., 2010; Blei et al., 2003) are used in machine learning. They provide a simple way to analyze large volumes of unlabeled text such as a huge database of written documents. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.
Data is assumed to be observed from a generative probabilistic process that has hidden variables namely, the thematic structure. This generative model for documents is based on probabilistic sampling rules that describe how words in documents might be generated on the basis of latent (random) variables. The generative process operates on bag-of-words assumption i.e., it does not make any assumptions about the order of words as they appear in documents.

The LDA model for each frame has the following parameters:

- **α**: A K-dimensional vector governing the Dirichlet distributions of the topics in a training video dataset:
  \[ \alpha = [\alpha_1, ..., \alpha_K] \]

- **β**: A \( K \times W \)-dimensional matrix representing the multinomial distributions of codebook words which form the vocabulary for all learned topics, where:
  \[ \beta_k = [\beta_{k1}, ..., \beta_{kw}] \]
  \[ \beta = [\beta_k] \]
  \[ \theta_j = [\theta_{j1}, ..., \theta_{jk}] \]  

For any given document, the probability of the distribution over topics is given by:

\[ p(\theta_j | \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_{j1}^{\alpha_1-1} \cdots \theta_{jk}^{\alpha_k-1} \]
\[ p(\theta_j | \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{i=1}^{N_j} \theta_{ji}^{\alpha_i-1} \]

where, \( N_j \) is the number of words in document \( j \). The marginal likelihood of the words given the hyperparameters \( p(w_j | \alpha, \beta) \) is intractable and therefore the marginal likelihood of the hidden variables given the hyperparameters: \( p(\theta_j, z_j | \alpha, \beta) \) is also intractable. Variational inference (Blei et al., 2003) is used to free parameters \( \gamma_j \) and \( \phi_j \):

\[ q(\theta_j, z_j | \gamma_j, \phi_j) = q(\theta_j | \gamma_j) \prod_{i=1}^{N_j} q(z_{ji} | \phi_{ji}) \]

An optimal \( (\gamma_j, \phi_j) \) is found by calculating the lower bound on the log likelihood given by \( \log p(w_j | \alpha, \beta) \).

Applying this to visual data, document words correspond to feature vectors while ‘topics’ correspond to atomic activities that frequently occur together in the video clips. The log likelihood of observing a previously unseen clip \( d_{doc} \) is approximated by its maximized lower bound \( \log p(d_{doc}^*) \approx L(\gamma^*, \phi^*; \alpha, \beta) \). \( \gamma^* \) is the Dirichlet parameters that represent the video sequence \( d_{doc}^* \) in a topic simplex. \( \phi^* \) are the prior probabilities of topics that show up in \( d_{doc}^* \), while \( \alpha \) and \( \beta \) are conditional probabilities given the words given topics as model parameters, computed in the M-step of the expectation-maximization algorithm.

Given the model, \( L(\gamma^*, \phi^*; \alpha, \beta) \) is evaluated by setting \( \alpha, \beta \) as constants and updating \( \gamma^*, \phi^* \) until \( L(\gamma^*, \phi^*; \alpha, \beta) \) is maximized. The normality score for a test clip is defined as:

\[ S(d_{doc}^*) = \frac{L(\gamma^*, \phi^*; \alpha, \beta)}{N_w} \]

Scores below a defined threshold is considered an indicator that an anomaly is present in that frame.

**EXPERIMENTS AND RESULTS**

The crowd behavior data used in our experiments are extracted from the UMN (Mehran and Shah, 2011) and PETS (University of Reading, 2010) surveillance video database for monitoring crowd activity. Clips with the following behavior types were used for training:

- **Type 1**: Sudden accelerated motion “rush”. Crowd motion that gradually changes into a panic situation but motion is towards one direction.
- **Type 2**: A crowd moving at normal pace suddenly disperses in different directions i.e., “scatter”.
- **Type 3**: Instantaneous movement towards an exit: “herding”.

4174
Two of the scenes are outdoors while one is in an enclosed environment. There are typical outdoor issues like lightening variation, occlusion and noisy image capture. After a random sample of clips, we select portions that contain transitions for use as validation or test-set. These were not used in the training stage. Atomic behaviors from patches were modeled using low-level visual pixel features in the flow fields. These form the building blocks for modeling the global behaviors as described in preceding sections. Long video sequences of three scenes were first divided into 2-3 sec clips and features were extracted (Fig. 2). Feature quantization was done using ant clustering and the visual codebook was built through which the clips
Table 1: Detection accuracy for two pixel patch sizes of feature descriptors

<table>
<thead>
<tr>
<th>Patch size</th>
<th>Rush</th>
<th>Scatter</th>
<th>Herding</th>
</tr>
</thead>
<tbody>
<tr>
<td>7x7</td>
<td>0.935</td>
<td>0.915</td>
<td>0.945</td>
</tr>
<tr>
<td>9x9</td>
<td>0.861</td>
<td>0.970</td>
<td>0.901</td>
</tr>
</tbody>
</table>

Fig. 6: ROC curve for classifier performance on the three behavior types using 7×7 patch size

are represented as counts of codebook vector types that appear in the clips.

A 20-topic Bayesian LDA model was fitted to build the models for the three different crowd scenes using only normal frames. The likelihood of a clip containing frames with abnormal behavior is determined given the model parameters of the corresponding normal behavior model. Frames whose negative log-likelihood falls below a defined threshold are classified as abnormal. As shown in Fig. 3, 4 and 5. The figures show sample scene of normal behavior and an abnormal one. Each figure also shows the ground truth (as viewed by a human observer) about the frames being classified by the classifier (green indicates normal behavior while red indicates an anomaly is present in the frame.

The effect of feature patch size on detection accuracy (Table 1) shows that a 7×7 pixel patch seems to give consistently high detection rates across the three behavior types (Fig. 6). Of the three scenes, the best performance is achieved for the ‘herding’ scene.

CONCLUSION

This study proposes a novel behavior descriptor for compact representation of crowd behavior, towards abnormal behavior detection. It uses an aggregation of features from optical flow, rotational and irrotational components of flow fields potentials and Lyapunov exponent to capture both local and global characteristics for compact representation of crowd behavior. This descriptor is not based on motion tracking. Also, it uses bio-inspired ant-clustering algorithm in the codebook development process instead of the commonly used k-means algorithm. The effectiveness of this descriptor in modeling crowd behavior was demonstrated using a Bayesian topic model based on the bag-of-features framework to detect transitions from normal to abnormal crowd behaviors in three different crowd scenes for crowd behaviors transiting contextually normal to abnormal behavior such as running, scattering and herding. Such early detection of the transition between normal phase and abnormal phase is of crucial importance in surveillance applications. Further study will be done in automatic recognition of emerging anomalies in crowds and indicating the local regions in the scenes where such anomalies occur.

ACKNOWLEDGMENT

The authors thank Prof. Wang Kejun (Harbin Engineering University) for his helpful comments.

REFERENCES
