In this paper, we present a novel method based on sparsely coded motion attention for detecting abnormal events in crowded scenes. Unlike existing sparse coding based approaches, our model does not need to learn a dictionary and directly sparsely codes the motion features of the center patches with features of its surrounding patches. The sparse coding error is used to measure the motion attention intensity of the center patch. To reflect the crowd abnormal intensity, an online updated weighting scheme is designed to obtain the global activity intensity map. Two publicly available datasets—UMN dataset and UCSD Ped1 dataset are utilized to evaluate our approach in detecting global abnormal event and local abnormal event, respectively. The experiments show our method achieves the promising performance and is competitive with the state-of-the-art approaches.

**Index Terms**— sparse coding, activity intensity, crowd behavior, abnormal detection

1. INTRODUCTION

Due to the wide range of applications based on video surveillance systems, crowd behavior analysis, especially abnormal behavior detection, has been attracting more and more attention in recent years. Unlike the analysis of individuals, it is a much more challenging task to understand crowd behavior because of both some “external” problems (occlusions frequently happen) and “internal” problems (complex interaction between individuals).

According to the scale, abnormal events in video surveillance can be divided into two categories [1]: 1) global abnormal event (GAE) and 2) local abnormal event (LAE). GAE happens in the entire scene e.g. group fight and panic. To detect GAE, Mehran et al. [2] propose a method based on social force model. They initially utilize particle advection to approximatively represent crowd activities, and then use normal local social force patterns to train a Latent Dirichlet Allocation (LDA) model. Finally, frames are labeled normal/abnormal based on that LDA. Cui et al. [3] propose an approach based on interaction energy potentials, which explores the relationships between the current state of a person and his/her reactions. Meanwhile, several approaches perform very well in detecting LAE, namely, the local behavior is different from their neighborhoods. Kim et al. [4] utilize MP-PCA to model the crowd activities and localize the anomaly by Markov Random Field. Mahadevan et al. [5] propose a method which adopts dynamic textures model to describe the crowd.

Different from the works above, we present a novel method based on sparsely coded motion attention for detecting both GAE and LAE in crowded scenes. In this work, we apply the properties of sparse coding in measuring the uniqueness of events from the perspective of motion attention. Fig. 1 illustrates the overview of the proposed method.
In order to obtain the motion attention map $M(x, t)$ at any pixel position $x$ of $t$-th frame, the motion features of a patch centered at position $x$ is sparsely coded by the motion features of the patches surrounding the center patch. The sparse coding error is used to measure the motion attention intensity of the center patch. To detect the crowd abnormal intensity, an online updated weighting scheme is designed to obtain the global activity intensity map. The resulting global activity intensity map is used to train a SVM classifier for detecting GAE. For LAE detection, a threshold of peak maximum value in abnormal activity map is trained to estimate if an input frame is normal or abnormal. In addition, the abnormal events can be localized by this method.

The contributions of this paper include the following aspects: 1) We propose a novel model to detect abnormal events using sparse coding. Different from existing sparse coding based approaches, no dictionary is learned in our method. Therefore, our method can adapt to any new scene immediately without any pre-processing. 2) We design an unsupervised online update scheme to measure the global activity intensity. This method reflects the abnormal events in the real world reasonably and enhance the robustness of detection. 3) Besides detecting both GAE and LAE, we can localize the LAE by our threshold based detection method.

The rest of this paper is organized as follows. In section 2, we introduce the computation of the motion attention map based on sparse coding. Section 3 introduces how to compute the global activity intensity map. We evaluate the performance of the proposed approach in section 4. The paper is concluded in section 5.

2. MOTION ATTENTION BASED ON SPARSE CODING

Our approach based on sparse coding is motivated by the nature of abnormal event patterns—they deviate from the ordinary types. In other words, they are always motion patterns which are different from their neighbors locally (LAE), or different from any known normal events globally (GAE). According to the nature of abnormal, we assume that for LAE, normal motion patterns can be sparsely represented by their neighboring regions with small error. On the contrary, abnormal motion patterns are represented with a large error. Similarly, for GAE, abnormal motion patterns are represented by any known normal ones with larger error. To detect the abnormal motion pattern, we propose to compute a motion attention map for each frame of a test sequence.

2.1. Motion feature extraction

Let $I(x, t)$ be the $t$-th frame of a test sequence, where $x \in \mathcal{X}$ is the 2D coordinates of a pixel in the image plane. Accordingly, let $V(x, t)$ be the motion features extracted from the pixel at position $x$ in the $t$-th frame. Although the proposed method is not related to which kind of motion features are used, in this work, we adopted the optical flow based motion features. Therefore, $V(x, t) = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \in \mathbb{R}^2$ consists of motion vectors along horizontal and vertical directions.

2.2. Local sparse coding

Let $y \in \mathbb{R}^D$ be the feature vector obtained by concatenating motion features of all pixels of the square region with size $d \times d$ centered at position $x$, where $D = d \times d \times 2$. To obtain the bases for sparse coding, we sample some neighboring regions around the position $x$. The neighboring regions are sampled by moving the square region centered at the position $x$ along $K$ equally spaced orientations with a step $l$. To adapt to different scales, we sample the neighboring regions in $R$ rounds, that is $l \in \{s, 2 \times s, \ldots, R \times s\}$. Therefore, we can obtain a total of $N = R \times K$ neighboring regions. Let $Y = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^{D \times N}$ be the matrix consisting of feature vectors of all neighboring regions. In order to deal with any un-predicted situations [6], we extend the basis functions $Y$ to $\tilde{Y} = [y_1, y_2, \ldots, y_N, \Phi_1, \Phi_2, \ldots, \Phi_D] \in \mathbb{R}^{(D+N)\times N}$, where $\Phi_j$ is the $j$-th column vector of the $R \times D$ identity matrix $\Phi$. We linearly represent the feature vector of the square region centered at $x$ using the extended basis functions as

$$y = \alpha_1 y_1 + \cdots + \alpha_N y_N + \beta_1 \Phi_1 + \cdots + \beta_D \Phi_D$$

$$= Y \alpha + \Phi \beta$$

$$= \tilde{Y} \gamma \quad \quad \quad (1)$$

where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T \in \mathbb{R}^N$, $\beta = [\beta_1, \beta_2, \ldots, \beta_D]^T \in \mathbb{R}^D$, $\gamma = [\alpha]^T \in \mathbb{R}^{N+D}$. The coefficient vector $\gamma$ can be solved using $\ell_1$ norm minimization

$$\gamma = \arg \min_{\gamma} \frac{1}{2} \|y - \tilde{Y} \gamma\|_2^2 + \lambda_1 \|\gamma\|_1 \quad \quad \quad (4)$$

The attention degree of the pixel at $x$ is measured as the error of representing $y$ using basis functions $\tilde{Y}$ and the associated coefficients $\alpha$ as

$$M_{local}(x, t) = \|y - Y \alpha\|_2^2 \quad \quad \quad (5)$$

We called $M_{local}(x, t)$ as the local motion attention map.

3. MEASUREMENT OF ACTIVITY INTENSITY

In real world scenes, abnormal events always happen with a higher intensity of activity. For instance, a group of people discussing (lower activity intensity) is normal, however, its abnormal when they fight (higher activity intensity). With this natural idea, we design a weighting scheme on the motion attention map to reflect the crowd activity intensity.
Due to the various crowd densities and speeds of crowd movement, the normal activity intensities are always not in the identical level for different scenes. Therefore, we compute an intensity threshold $\tau$ which indicates the average normal activity intensity up to time $t$ by

$$\tau_t = \frac{1}{|\mathcal{L}| \times |\mathcal{X}|} \sum_{l \in \mathcal{L}} \sum_{x \in \mathcal{X}} V(x, l)$$  \hspace{1cm} \text{(6)}$$

where $\mathcal{L}$ is the index set of all detected normal frames up to time $t$. The weight matrix $W(x, t)$ can be computed as

$$W(x, t) = \exp(V(x, t) - \tau_t)$$  \hspace{1cm} \text{(7)}$$

The global activity intensity map $M_{\text{global}}(x, t)$ can be obtained as

$$M_{\text{global}}(x, t) = W(x, t) \times M_{\text{local}}(x, t)$$  \hspace{1cm} \text{(8)}$$

We need no prior knowledge to compute $\tau_t$. It is constantly updated using the detected normal frames up to time $t$ via Eq. 6. Then the updated $\tau_{t+1}$ is used to generate the global activity intensity map for the test frame at $t + 1$. Fig. 2 illustrates this online update process for $\tau_t$.

4. EXPERIMENTS

4.1. Global abnormal event detection

We evaluate our approach for GAE detection on the UMN dataset [7] which contains the videos of 11 crowd panic escaping events in 3 different scenes. Fig. 3 shows samples of this dataset.

For construction of visual words, we randomly extract 30 spatial-temporal volumes in size of $5 \times 5 \times 10$ as setup in [3]. A codebook is built which comprises 30 cluster centers. In the experiment, we use 10 video clips to train our model with SVM. Then, we compute the true positive rate (TPR) and false positive rate (FPR) on the rest video clip. The ROC curve is shown in Fig. 4. To evaluate our approach, we use the conventional Optical Flow and Social Force model [2] as baseline.

In addition, the novel method of Streakline Potentials [8] and Interaction Energy Potentials [3] joins the comparison. The AUCs are shown in table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Optical Flow</th>
<th>SF</th>
<th>Streakline Potentials</th>
<th>Interaction Energy Potentials</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.84</td>
<td>0.96</td>
<td>0.9</td>
<td>0.985</td>
<td>0.9887</td>
</tr>
</tbody>
</table>

4.2. Local abnormal event detection

We evaluate our approach for LAE detection on UCSD Ped1 dataset [9]. The UCSD Ped 1 dataset contains 34 training clips of normal events and 36 testing clips consisting of local abnormal events. Each clip has 200 frames with a resolution of $158 \times 238$. As our approach needs no training for LAE detection, we directly test our approach on the 36 testing clips. In order to robustly evaluate the abnormal degree of this map, we compute the average of 5 largest values in the global activity intensity map for each testing frame. Finally, we use a
For LAE detection, our approach has ability to reveal where abnormal event happens based on the computed global activity intensity map. Fig. 6 shows original video frames (up row) and the global activity intensity map (bottom row) in the No.35 test clip of UCSD Ped1. The people on skateboard has an obvious higher motion attention degree than the other areas and is detected as LAE. And the location of this LAE is localized by the red area shown in the global activity intensity map.

5. CONCLUSION

In this paper, we present a novel method based on sparsely coded motion attention for detecting both GAE and LAE in crowded scenes. Given an input video frame and its optical flow field, we apply the theory of sparse coding to generate a pixel-level motion attention map. Unlike existing sparse coding based approaches, we do not need any pre-learnt dictionary. Instead, to represent a patch, a collection of basis is obtained from its neighboring patches. Moreover, we design a weighting scheme to compute the global activity intensity. To evaluate the effectiveness, we compare the proposed method with several state-of-the-art methods on two publicly available UMN and UCSD datasets. The experimental results indicate that our method performs well in detecting both GAE and LAE in crowded scenes.

6. REFERENCES