SWARM-BASED MOTION FEATURES FOR ANOMALY DETECTION IN CROWDS

Vagia Kaltsa\textsuperscript{1,2}, Alexia Briassouli \textsuperscript{2}, Ioannis Kompatsiaris\textsuperscript{2}, Michael G. Strintzis\textsuperscript{1}

\textsuperscript{(1)}Aristotle University of Thessaloniki, \textsuperscript{(2)}Informatics and Telematics Institute, CERTH

ABSTRACT

In this work we propose a novel approach to the detection of anomalous events occurring in crowded scenes. Swarm theory is applied for the creation of a motion feature first introduced in this work, the Histograms of Oriented Swarm Accelerations (HOSA), which are shown to effectively capture a scene’s motion dynamics. The HOSA, together with the well known Histograms of Oriented Gradients (HOGs) describing appearance, are combined to provide a final descriptor based on both motion and appearance, to effectively characterize a crowded scene. Appearance and motion features are only extracted within spatiotemporal volumes of moving pixels (regions of interest) to ensure robustness to local noise and allow the detection of anomalies occurring only in a small region of the frame. Experiments and comparisons with the State of the Art (SoA) on a variety of benchmark datasets demonstrate the effectiveness of the proposed method, its flexibility and applicability to different crowd environments, and its superiority over currently existing approaches.

\textit{Index Terms}— crowds, anomaly, swarm intelligence

1. INTRODUCTION

The widespread use of surveillance systems in roads, stations, airports or malls has resulted in a huge amount of videos of crowded scenes that need to be analyzed for security, retrieval or even commercial reasons. Crowds pose challenges for traditional computer vision and image/video processing methods, due to the presence of occlusions, varying crowd densities and the complex stochastic nature of their motion, so numerous alternative approaches have been developed to overcome these barriers.

In this work, we deal with crowd videos of medium density and focus on detecting abnormal patterns which emerge as spatiotemporal changes, not only in velocities, but also in appearance, such as detecting a bicycle passing through a crowd. Even though tracking approaches have been used in the past\textsuperscript{[1], [2]}, the idea has been abandoned due to the nature of the problem, as tracking seems to be inefficient in videos with crowds of high or medium density. As a result, the current SoA focuses mostly on the analysis of entire frames in space, time, or both. Existing methods can be classified in two main categories: those that utilize only motion information in order to detect an abnormality in the scene and those that use both appearance and motion information to describe the scene’s dynamics. In the first category, representative works are the one of Wu et al.\textsuperscript{[3]} who uses chaotic dynamics to model particles’ trajectories and that of Mehran et al.\textsuperscript{[4]} who use the social force model to describe a crowd’s normal behavior. In the same category, other interesting works based exclusively on motion characteristics are\textsuperscript{[5], [3], [6], [7], [8], [9], [10]}.

A common problem encountered by the methods mentioned above, is their inability to successfully detect abnormalities whose motion is similar to that of the normal crowd. The second category of methods tackles this issue by incorporating appearance information as well. One very good approach in this category is that of\textsuperscript{[11]} that uses mixtures of dynamic textures, however, the computational cost of the algorithm proposed, makes it prohibitive for many applications. A joint modeling of appearance and dynamics is also proposed by Ito et al.\textsuperscript{[12]} for detecting interesting events, but the method is only suited to detecting salient events, ignoring any abnormalities happening in a smaller scale. Another work in the same category is that of\textsuperscript{[13]} however, in this approach classification is determined by a pre-defined threshold which makes the method sensitive to input video.

In this work we propose a new method for anomaly detection and localization that incorporates both motion and appearance information. We introduce a new descriptor created from histograms of oriented gradients (HOG) and newly introduced histograms of oriented swarm accelerations (HOSA) to capture frame dynamics. Swarm intelligence has been used in the past only within the framework of Particle Swarm Optimization (PSO), whereas this work introduces an innovative deployment of swarm intelligence, which, together with the HOG descriptor, succeeds in forming a new final feature capable of successfully determining a region’s “normality” in an SVM framework. Our contribution can be summarized as follows:

1. Swarms are used for the first time in videos for anomaly detection and are shown to provide credibly filtered crowd velocity with minimal noisy flow values.

2. The method can be effectively applied when crowds move non-uniformly in space and time, unlike the SoA.
3. Inspired by crowd psychology and the analysis of individuals’ movements in crowds [14], swarms match real world behavior in crowds, resulting in a more accurate representation, as our experiments show.

2. OVERVIEW OF THE METHOD

Our algorithm is only applied in regions of interest (ROIs) instead of the whole frame (Fig. 1), to achieve a lower computational cost, fewer false alarms, greater precision, and to detect and localize anomalies both on a global and local scale. We first apply background subtraction using weighted moving mean [15] and define interest points as a dense grid on the foreground. Tracking these interest points into the next frames using a KLT tracker and updating with new interest points from the extracted foreground, we define ROIs as rectangular areas of specified size around each interest point.

Appearance information is obtained from Histograms of Oriented Gradients (HOGs) [16], using the implementation of [17]. HOG uses grey scale images, which make it color invariant, and it is invariant to illumination and local geometric transformations as a result of normalization. Meaningful motion information is captured by the novel concept of Histograms of Oriented Swarm’s Acceleration descriptor (HOSA), which is introduced for the first time in this work (Sec. 3).

In order to achieve scale invariance in both space and time, for both the HOG and HOSA, each ROI forms a block that is divided into $2 \times 2$ cells. For the HOG, a weighted histogram of 9 bins for gradient orientations is created and normalized. The 4 resulting cell histograms are concatenated, forming a $1 \times 36$ block descriptor, which is normalized for noise elimination. To exploit temporal information, each block is tracked over time and its average is taken over every 3 frames, so the final appearance descriptor for each block is a concatenation of these average triplets. For example, a 15 frame time window will result in 5 concatenated triplets of $1 \times 36$ descriptors, resulting in a $1 \times 180$ spatiotemporal appearance descriptor. It should be noted that this procedure is followed both for appearance and motion descriptors (HOG, HOSA), which are then concatenated to be used in a one-class SVM (OC-SVM). An overview of the method is illustrated in Fig. 1.

3. ESTIMATING HOSA DESCRIPTOR

In our implementation, we adopt physics-based modeling of the crowd, as crowd properties appear to be very similar with those of a naturally occurring swarm. The swarm model that is used is based on the general theory described in [18]. The core idea is the simulation of a natural swarm of “predators” which is hunting a prey. According to this approach, swarms are made up of agents and a prey: The agents “track” the prey, but also interact with each other as they would in a real swarm. Thus, agents (“predators”) are subject to three forces: physical forces like inertia and friction, interaction forces among them, and external forces dependent on the prey, in our case the optical flow magnitude (OF). Interaction forces ensure the cohesion of the swarm of agents, friction forces maintain elementary memory of agents’ velocity, while external forces depend on the prey characteristics being tracked. Swarms are deployed to capture scene’s motion dynamics and detect anomalous events in them by forming Histograms of Swarm’s Acceleration (HOSA) in each region of interest. The main concepts of the swarm descriptor are presented below.

3.1. Prey generation

In this work, the prey followed by a swarm is generated by the OF values of pixels lying inside ROIs, instead of their luminance values as in [18]. Thus, the number of prey in each frame is equal to the number of ROIs in that frame. In this section we describe how prey data is extracted, namely how it is mapped to be tracked by agents (“predators”). As mentioned previously, ROIs correspond to rectangle areas around each interest point containing a constant number of $n$ pixels. In order to form the prey, we consider the pixels of the ROI rectangle sequentially, so that the pixel at position $(i, j)$ is assigned its cardinality $t$ with $1 \leq t \leq n$. We then use the magnitudes of their OF $O(t)$ as prey positions, so the swarm will track the OF magnitudes of all pixels $1, \ldots, n$ in the particular ROI: $\{x_p(1), x_p(2), \ldots, x_p(n)\} = \{O(1), O(2), \ldots, O(n)\}$ and its dynamics will be used to create HOSA.

3.2. Forces extraction

As mentioned above, each agent of the swarm (Eq. (6)) moves according to the result of three forces: neighborhood force, friction force and external force. The force for prey $t$ between agent $i$ and all agents $j$ in its neighborhood $V_i$ is:

$$F_{\text{neigh}}(i,t) = \sum_{j \in V_i} F_{\text{int}}(i,j,t)$$  \hspace{1cm} (1)
where the interaction force $F_{int}$ between each agent $i$ and all other agents of the swarm in its neighborhood (agents whose distance from $i$ is smaller than $\rho$) is given by:

$$F_{int}(i,j,t) = \begin{cases} 
\frac{\beta \cdot (r_i - r_j)}{d(i,j)}, & |r_i - r_j| \leq d_{min} \\
-\alpha \cdot \frac{(r_i - r_j)}{d(i,j)}, & d_{min} < |r_i - r_j| \leq \rho 
\end{cases}$$

Here $d(i,j)$ denotes the distance between agents $i$ and $j$, $r_i$ is the previous position of agent $i$ and $\alpha$, $\beta$ are weighting parameters set equal to 1, as agent–prey distances are found to be a sufficient measure for internal force strength, and do not need to be amplified or compressed. $F_{neigh}$ can be attractive or repulsive, as its role is to prevent collisions of agents and ensure swarm cohesion. The friction force $F_{fric}$ that acts on each agent $i$ offers elementary memory to the swarm by being a function of agent velocity, corresponding to the previous prey location $t - 1$:

$$F_{fric}(i,t) = -\mu \cdot \dot{x}_i(t - 1)$$

where $0 \leq \mu \leq 1$ is the friction coefficient. After experimentation, the value $\mu = 1$ is found to provide the best tradeoff between tracking speed, smoothness and accuracy. Finally, the swarm is driven across the frame mainly by the external prey–agent force $F_{ext}$ of Eq. (4) that guides the swarm “over” the prey.

$$F_{ext}(i,p,t) = \lambda \cdot (x_p(t - 1) - x_i(t - 1))$$

From Eq. (4) it is clear that the external force between the swarm agents and the prey pixels is directly dependent on their relative values (in practice OF magnitude), with the force becoming weaker as the swarm agent diverges from its prey. This force is similar to the restoration force of a harmonic oscillator, so $\lambda$ represents the positive spring constant, whose best value is found experimentally to be $\lambda = 1$.

### 3.3. HOSA descriptor

In order to form the HOSA, we examine the evolution of agents’ accelerations according to prey motion patterns and the forces affecting the agents. The velocity of agent $i$ at location $x_i$ is given by:

$$\dot{x}_i(t) = \gamma \cdot \dot{x}_i(t - 1) + \delta \cdot \ddot{x}_i(t),$$

while the acceleration is given by:

$$\ddot{x}_i(t) = \gamma \cdot \ddot{x}_i(t - 1) + F_{neigh,x}(i,t) + F_{fric,x}(i,t) + F_{ext,x}(i,p,t).$$

Here, $\gamma$ is a memory parameter relating past values of the acceleration with the current one, essentially forming an autoregressive process where flow values change slowly over space and time. If a pixel’s flow undergoes a sudden change, it will be captured by the forces acting upon it in Eq. (6), so the influence of its previous value will be mitigated.

Swarm agents positions are randomly generated for the first pixel $t = 0$, and their speeds and accelerations are initially set to zero. Their values change over time depending on prey locations $O(t)$, and the forces affecting the agents. During training, ROIs are extracted and the pixel OF in them is tracked by the agents. We then compute the average of the swarm agents’ accelerations and follow a process similar to HOG extraction (Sec. 2) to create weighted histograms of agents’ accelerations according to OF’s orientation (HOSA) (Fig. 2).

### 4. ANOMALY DETECTION AND LOCALIZATION

Appearance and motion descriptors are combined to form the descriptor we deploy for crowd anomaly detection. In a temporal window of $n$ frames, HOG and HOSA features are averaged over three consecutive frames at a time, and their averages are then concatenated, resulting in a final feature vector. The overall process is depicted in Fig. 1. An One-Class SVM (OC-SVM) framework is then used for anomaly detection. In comparison to other machine learning methods, SVMs generally lead to good performance at a low computational cost, allowing for reliable real time detection. Furthermore, they are able to easily handle large data sets, which generally appear in practical situations. Because of the innumerable anomalies that can appear, they cannot be all provided as training samples, so a one class classifier is chosen for detecting deviations from a normal crowd situation. After training our model, localization of anomalies is easy, as descriptors have already been estimated in specific spatial locations, namely in ROIs around interest points. Thus, our algorithm checks each ROI independently to determine if it can be characterised as normal, and subsequently notify the system accordingly.

As descriptors in each ROI are examined and characterized separately, our method has the advantage of being capable to deal with non-uniformly moving and evolving motion.
crowds. Indeed the experiments that follow demonstrate that our method accurately localizes different anomalies in space and time, in realistic benchmark videos.

5. EXPERIMENTAL EVALUATION

To evaluate the performance of our algorithm, two SoA datasets were used. The first one is UCSD Ped2 [11], which consists of 16 training clips of normal crowd situations and 14 testing clips in which some frames have one or more abnormal situations. The ground truth used is described in [11] and includes both frame level temporal anomaly detection (frames with anomalous events) and pixel level spatial anomaly localization. Table 1 depicts the Equal Error Rate (EER) which is the point where the proportion of false positives is equal to the proportion of false negatives. The lower the equal error rate is, the higher the accuracy of the system. As it can be seen our method outperforms the SoA in the challenging problem of spatial anomaly localization, while it is very close to the SoA [11] in global (frame - level) anomaly detection, despite only processing data locally. Frames where anomalies are detected are shown in Fig.3, where it is obvious that we can detect different types of anomalies (in red) even in challenging situations where two different anomalous events or entities are present in the same frame (e.g. a bicycle and a skateboard, Fig. 3(d)).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame level</td>
<td>42%</td>
<td>30%</td>
<td>42%</td>
<td>25%</td>
<td>27.87%</td>
</tr>
<tr>
<td>Pixel level</td>
<td>79%</td>
<td>82%</td>
<td>72%</td>
<td>55%</td>
<td>36.17%</td>
</tr>
</tbody>
</table>

Table 1. Comparison of performance of the proposed method with other SoA works in Ped2 dataset. The first row depicts EER for frame anomaly detection, while the second row depicts EER for pixel level anomaly localization.

Fig. 3. Ped2 anomalies shown in red blocks. Red contours depict foreground areas where the algorithm is applied.

We also examined the widely used benchmark UMN dataset [19], which includes 3 different scenes of crowds of people walking and suddenly dispersing in an evacuation. About 350 normal frames were used for OC-SVM training, while the rest were used for testing. A frame is marked as abnormal if more than two anomalies are detected in it, as fewer anomalies correspond to false alarms. This post processing step was added as our method is based on spatial information in order to detect smaller anomalies and thus more frames were needed for training reasons in order to omit this step. Fig.4 compares the temporal localization that our algorithm achieves with manually extracted ground truth, where it can be seen that our method successfully localizes abnormalities with great accuracy both on a pixel level and on a frame level. More example videos can be found in the demo at http://youtu.be/u2v4LaGvVSQ.

6. CONCLUSIONS

In this work, an innovative application of swarm-based processing is presented by introducing a novel motion feature based on histograms of swarm accelerations (HOSA). HOSA combined with HOG descriptor form a strong representative frame descriptor able to discern spatiotemporal anomalies. This is supported by comparisons with the SoA, which show that our method achieves good performance in challenging datasets such as the UCSD videos. Future work includes generalizing our method to completely different datasets with other kinds of anomalies and not crowd events in particular.
7. REFERENCES


“Unusual crowd activity dataset made available by the university of minnesota at: http://mha.cs.umn.edu/movies/crowdactivity-all.avi/,”.