Unsupervised Discovery of Crowd Activities by Saliency-based Clustering

Tingting Han, Hongxun Yao*, Xiaoshuai Sun, Sicheng Zhao, Yanhao Zhang
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.

Abstract

Along with the rapid development of digital information technology, video surveillance systems have been widely used in numerous public places, such as squares, shopping malls and banks, to monitor crowd in case of anomalous events. Meanwhile, great challenges have been posed to worldwide researchers because the analysis of the exponentially-growing crowd activity data is an arduous task. In this paper, we develop a novel unsupervised crowd activity discovery algorithm aiming to automatically explore latent action patterns among crowd activities and partition them into meaningful clusters. Inspired by the computational model of human vision system, we present a spatio-temporal saliency-based representation to simulate visual attention mechanism and encode human-focused components in an activity stream. Combining with feature pooling, we can obtain a more compact and robust activity representation. Based on affinity matrix of activities, N-cut is performed to generate clusters with meaningful activity patterns. We carry out experiments on our HIT-BJUT dataset and the UMN dataset. The experimental results demonstrate that the proposed unsupervised discovery method is capable of automatically mining meaningful activities from large-scale video data with mixed crowd activities.

Keywords: Spatio-temporal Saliency, Unsupervised Discovery, Crowd Activity Analysis, Spatio-temporal Pooling, Dynamic Time Warping, Normalized Cut

*Corresponding author.

Email addresses: ttinghan@hit.edu.cn (Tingting Han), h.yao@hit.edu.cn (Hongxun Yao)

Preprint submitted to Neurocomputing December 10, 2014
1. Introduction

Unsupervised learning plays an important role in knowledge exploration and discovery. Since the increasing amount of multimedia data brings more and more laborious labeling work, it becomes a hot trend to apply unsupervised methods to handle relevant problems in computer vision research. For the last decades, unsupervised approaches have been extensively studied for object localization [1] and segmentation [2], action categorization [3] [4], human activity analysis [5] [6] and facial image analysis [7], etc.

Meanwhile, crowd activity analysis has attracted significant research interests in recent years for its potential applications in intelligent security monitoring of public places, such as railway stations, shopping malls and crowded sports arenas, etc. Most of existing works on activity analysis [8] [9] [10] are based on supervised learning on labeled data and detect abnormal activities by matching or classification strategies. However, there are some limitations of these methods: 1) the supervised approaches are appropriate only when there are abundant labeled training samples, which is obviously burdensome and impractical; 2) even if adequate samples are given, it is still not flexible to build one single model for accurate inferences on large scale and diverse data stream. Furthermore, it is very challenging to analyze crowd activities in a non-supervised manner. Unlike action analysis of a single person, the understanding of activities performed by multiple or a crowd of individuals has to overcome thorny problems such as diverse semantics, various expressions, complex interaction and occlusions, etc. Therefore, in order to better describe activities, an effective and robust feature representation is required.

In this paper, we attempt to automatically cluster a set of unlabeled crowd activity videos based on similarity affinities and discover semantically meaningful patterns for unsupervised crowd activity analysis. The idea is motivated by a recent work of Faktor et al. [11], which attempts to discover image categories in an unsupervised manner. It takes advantage of the wisdom of crowds of images to obtain a sparse yet meaningful image affinities. With a graph clustering employed to the affinity matrix, the underlying structures are explored to
partition the multimedia collections into clusters of similar visual properties.

We present a spatiotemporal saliency strategy to locate the human-focused dynamic changes within the stream of activities, which can comprehend activities from the perspective of human. Benefiting from the effectiveness of the proposed features, our algorithm performs well on highly unbalanced data. To support our unsupervised crowd activity analysis, we construct a new dataset named HIT-BJUT containing various types of human activities which is recorded in the campus of BJUT and will be released soon.

One preliminary version on unsupervised discovery of crowd activities was first introduced in our previous work [12]. In this paper we have improvements in three aspects: 1) we perform a more comprehensive survey of related works; 2) we enrich the theory of the proposed method; and 3) we conduct more comparative experiments and provide more analysis of the results. The remainder of this paper is organized as follows. Section 2 reviews the related methods and features for crowd activity analysis. The proposed unsupervised crowd activity analysis method based on spatio-temporal representation is described in Section 3. The test datasets and evaluation criteria are presented in Section 4, followed by experimental results and analysis in Section 5. Finally, conclusions are drawn in 6.

2. Related Work

Significant progress has been made for action recognition over the last decade [13] [14]. By far, the most popular representation for RGB video is Bag-of-Visual-Words (BoVW) model based on low-level features, such as local interest points [15] and dense trajectories [16]. More recently, increasing efforts have been focused on exploring mid-level features. Discriminative action parts such as motionlets [17], actons [18] and attributes [19] have been mined to describe actions. With the development of the techniques for capturing depth data, researchers have also been exploring action representation of 3D actions [20] [21].

While action recognition focuses on motion patterns of single person or pair-wise persons, crowd activity analysis takes the interaction of the crowd into account and treat it as a whole
to represent, which is more complicated and challenging due to the problems of diverse semantics, various expressions, complex interactions and occlusions, etc. Recent works on crowd activities mainly focus on the topic of anomaly detection, which aims to discover the abnormal activities from a set of normal activities. Conventional methods tend to address the challenging task in a supervised manner, and can be coarsely divided into two categories: model-based methods and particle advection-based methods.

In the framework of model-based method, it is a convention to train a model for normal crowd activities, and those activities which cannot be covered by the trained model are identified as the anomalies. In literature [22], a dynamic texture model jointly modeling appearance and dynamic information in a crowd scene was employed to detect both temporal and spatial anomalies. L. Kratz et al. [23] analyzed the underlying structure formed by the spatial and temporal variations in the motion to exploit the steady state motion of crowd activities and modeled the motion patterns with a Hidden Markov Model (HMM). In this way, abnormal activities could be detected as the motion patterns with low likelihood. In [24], J. Kim et al. tackled the abnormal activity detection by spatial-temporal Markov Random Field (MRF). Mohamed R. Amer et al. [25] presented a new deep model, called Hierarchical Random Field (HiRF), for representing and recognizing collective activities in videos. B. Antic et al. [26] proposed a technique based on video parsing to achieve abnormality detection. Cong et al. [27] proposed to detect the presence of anomalies in crowded scenes by a sparse reconstruction cost. While more popular methods in recent years are based on particle advection schemes. In these methods, a grid of particles are considered in each frame which are then advected using the underlying motion data [8] [10] [28] [29]. In [8], Social Force Model (SFM) was employed to detect abnormalities by estimating the interaction force, which in turn, was used to describe crowd behavior. Although these methods have obtained the state-of-the-art performances, the difficulties of labeling huge amount of training samples and modeling a single model for activities with complex semantics and diverse expressions still remain to be overcome.
In addition to the learning methods, another key point is to extract effective features from the spatio-temporal video data. There are usually two types of widely used features: 1) high-level features abstracted from the detected objects by tracking and recognizing process; 2) low-level features directly derived from the image pixels. Most commonly used object-based features are the position and trajectory of the object’s centroid, which have been proved to be effective in detection of various types of anomalies, such as running and falling [30] and traffic anomaly [31]. In addition, features such as limb angles [32] [33] can be used to identify the poses or actions performed by the objects when the resolution of the videos is high enough. However, to abstract these features, the separation and tracking of multiple targets have to be solved as a preprocessing stage. Unfortunately, these two problems themselves remain challenging because of the severe occlusion and other distractions existing in the videos. On the other hand, low-level features become more and more popular when dealing with abnormal activity detection benefiting from their robustness against the aspects, such as occlusion, which would affect the tracking accuracy negatively. What’s more, the abstraction of low-level features does not rely on the segmentation or tracking of objects, so it makes them effective when there are large numbers of targets in view. As the most basic and successful feature to describe motion information in videos, optical flow is widely used in the task of anomaly detection [8] [34] [35]. Features such as color, texture, gradient and shape have also been explored for this purpose [22] [36] [37] [38].

Different from these methods, we develop an unsupervised activity analysis method to explore the underlying patterns existing in the data without involving manually labeling and complex modeling. Moreover, we propose a saliency-based feature map which can abstract information inconsistently capturing more important components that humans focused on and ignoring those of little significance which would cause unnecessary computational burden. And then we build a global representation to depict multi-person activities as a whole. The proposed method is able to automatically discover meaningful activities from the mixed data fast and effectively.
3. Method

In this section, we introduce our proposed unsupervised activity discovery method in detail. Firstly, we compute the spatio-temporal saliency as the feature map to capture human-focused components in activities. In this way, the feature representation can be abstracted to describe activities more effectively by pooling from spatio-temporal saliency. To obtain the similarity affinities between activities of various lengths, Dynamic Time Warping (DTW) matching is leveraged. Finally, we apply a graph theory-based clustering method, named Normalized Cut, to the affinity matrix and partition the crowd activity dataset into multiple clusters. The framework is shown in Fig. 1.

Fig. 1. The framework of unsupervised crowd activity discovery method.

3.1. Spatio-Temporal Saliency Map

In the problem of crowd activity analysis, to abstract the most important and reasonable information for activities is the most critical step. In this paper, we attempt to handle it inspired by the computational model of human vision system. We employ a saliency-
based method to capture meaningful information from the raw data, which contributes to locate what humans really focus on during an activity. We consider saliency spatially and temporally, where spatial saliency describes the spatial distribution of human-focused objects while temporal saliency captures the dynamic changes that humans pay attention to.

**Spatial saliency** : For each frame of a given activity video, a spatial saliency map is constructed using the method proposed in [39]. There are many off-the-shelf approaches to compute saliency [40] [41] [42]. The Spectral Residual (SR) approach proposed by Hou et al. [41] is based on Fourier Transform, which is an unsupervised and rapid saliency detection method. After that, Guo et al. [39] proposed to use Phase preserving inverse Fourier Transform (PFT) to detect saliency and further simplified the original SR approach. Taking both computational efficiency and biological plausibility into consideration, we utilize the PFT model for our spatial saliency, which is defined as

\[
S(x, y) = \|F^{-1}(e^{i P(F(I(x,y)))})\|, \quad (1)
\]

where \(S(x, y)\) is the spatial saliency, of a given frame \(I(x, y)\), \(F\) and \(F^{-1}\) represent the Fourier Transform and its inverse respectively, and \(P(\cdot)\) denotes the phase spectrum.

**Temporal saliency** : For the successive frames in an activity video, we compute temporal saliency to capture the human-focused dynamic changes. Firstly, we compute the difference image between two adjacent frames to obtain global changes. Then, we execute saliency detection on the difference image to capture the local changes of human attention. The process can be formulated as follows:

\[
d(x, y) = F(I_i(x, y) - I_{i-2}(x, y)), \quad (2)
\]

\[
T(x, y) = \|F^{-1}(e^{i P(F(d(x,y)))})\|, \quad (3)
\]

where \(d(x, y)\) is a difference image, the subscripts \(i\) and \(i - 2\) denote the temporal locations of the computed frames. Given the difference image \(d(x, y)\), the computation process of
temporal saliency $T(x, y)$ is the same as Equ. (1). The reason of computing difference image of every other frames is to avoid the disturbance caused by the coding and decoding process of videos. The computation of spatio-temporal saliency in this stage acts as a mapping from spatio-temporal space to feature space, which abstracts only important and distinctive components from an activity. In order to obtain an effective feature representation for activities, spatio-temporal pooling is employed to form more compact and robust features in the second stage.

3.2. *Spatio-Temporal Pooling*

Pooling based feature extraction contributes to reduce dimensions of features and thus improves the computability. Besides, features obtained by pooling have the “stationarity” property which makes them more robust against local feature noise. In this paper, we divide the spatio-temporal saliency sequences into multiple overlapping spatio-temporal cells and perform pooling operation, as shown in Fig. 2.

Fig. 2. The process of “Spatio-Temporal Pooling” operation.
The raw frames and corresponding pooled features of different activities. The activities showed are fighting, escaping, escaping and two normal activities from top to bottom.

There are many types of pooling techniques. Specifically, two pooling methods have been investigated in this paper, Average Pooling [43] & Max Pooling [44] (AveP & MaxP):

\[
S_{AveP}(x, y, z) = \frac{1}{\| \Psi \|} \sum_{(i,j,t) \in \Psi} S(i, j, t),
\]

(4)

\[
S_{MaxP}(x, y, z) = \max_{(i,j,t) \in \Psi} S(i, j, t),
\]

(5)

where \( \Psi \) is the spatio-temporal cells centered at \((x, y, t)\). Combining pooled values from
spatial cells in current temporal cell $S(\cdot,\cdot,t^{(n)})$ into a vector, and hence we obtain a feature representation $F$ in $R^{M \times N}$, where $M$ depends on the size of spatio-temporal cells, $N$ is related to the duration of the activity video. Fig. 3 shows raw frames and visualized pooled feature sequences of five activities. It can be seen from the figure that the salient points keep in one area during the whole period of fighting activities, while in escaping activities, the saliency changes from the centralized to the decentralized. Different from them, the saliency points distribute chaotically all the time in normal activities. The characteristic differences among different crowd activities demonstrate that the proposed spatio-temporal feature mapping can effectively capture the distinctive components from different type of crowd activities.

3.3. Dynamic Time Warping (DTW)

The variety of time durations of different activity videos results in features of various lengths. Therefore, we adopt Dynamic Time Warping (DTW) distance [45] to calculate reasonable similarity between inconsistent features. Here, we do not employ complicated behavior modeling strategy, such as variable length Markov model [46], because they usually involve complicated training phases and hence are not applicable in our unsupervised activity discovery.

DTW is a dynamic programming algorithm of finding the optimal aligning between two given sequences that are warped non-linearly in the time dimension. It has been applied to automatic speech recognition [45], speaker recognition [47], online signature recognition [48] and shape matching [49]. In our application, the inputs of DTW are two pooled spatio-temporal saliency features sequences, denoted by $R_n$ and $T_m$. The subscripts $n$ and $m$ are the lengths of the two feature sequences, which depend on the videos’ lengths and the size of the temporal cells.

\[ R_n = r_1, r_2, \cdots, r_i, \cdots, r_n, \]  
\[ T_m = t_1, t_2, \cdots, t_j, \cdots, t_m, \]
where element \( r_i \) is a combined feature vector in the \( i \)th temporal cell. And the feature sequences \( R_n \) and \( T_m \) can be arranged to form a \( n \)-by-\( m \) grid, where each grid point \((i, j)\) corresponds to the distance between elements \( r_i \) and \( t_j \), which is calculated by Euclidean distance metric,

\[
d(i, j) = \sqrt{\sum (r_{ik} - t_{ik})^2}, k = 1, 2, \cdots K,
\]

where \( K \) is the length of the vector in a certain temporal cell, and its value relies on the size of spatial cells,

\[
K = \frac{(2W - \Delta W)(2H - \Delta H)}{\Delta W \Delta H}.
\]

Then, the DTW distance between \( R_n \) and \( T_m \) can be computed by the following dynamic programming process:

\[
DTW(R_n, T_m) = d(n, m) + \min \left\{ \begin{array}{l}
DTW(R_{n-1}, T_{m-1}), \\
DTW(R_{n-1}, T_m), \\
DTW(R_n, T_{m-1})
\end{array} \right\},
\]

\[
DTW(\langle \rangle, \langle \rangle) = 0,
\]

\[
DTW(R_n, \langle \rangle) = DTW(\langle \rangle, T_m) = \infty,
\]

where \( \langle \rangle \) denotes an empty sequence, and the subscripts denote the length of sequences. By applying DTW matching to any two activities in the dataset, we can get a nonnegative and symmetric activity affinity matrix \( W \), recording the similarities among all the behaviors. After that, we utilize a graph cut method, called “Normalized Cut”, to implement the partition of crowd activities based on the affinity matrix \( W \).

3.4. N-cut Clustering

Given the similarity affinity matrix, the set of activities can be regarded as a weighted undirected graph \( G = (V, E, W) \), where the vertices set \( V \) represents activities, and the edge set \( (E) \) is formed between every pair of nodes. The weights on the edges are given by the similarity matric \( W \) we have obtained in Section 3.3. Therefore, we can solve the activity
grouping by graph-based spectral clustering algorithms benefiting from the global optimal clustering. We focus on the $K$-way normalized cut (N-cut) spectral clustering algorithm [50], which adds a normalization term to traditional minimum cut criteria and successfully avoids unnatural bias for partitioning out small sets of points.

The goal of $K$-way normalized cuts is to partition the vertices set $V$ into $K$ disjoint sets, i.e., $V = \bigcup_{i=1}^{K} V_i$ and $V_k \cap V_l = \emptyset, \forall k \neq l$. Denote the K-way partitioning by $\Gamma^K_V = \{V_1, \cdots, V_K\}$.

If $A, B \subset V$, define $\text{links}(A, B)$ to be the total weighted connections from $A$ to $B$:

$$\text{links}(A, B) = \sum_{i \in A, j \in B} W(i, j),$$

the normalized link ratio of $(A, B)$ is defined as:

$$\text{linkratio}(A, B) = \frac{\text{links}(A, B)}{\text{links}(A, V)}$$

where the normalization term $\text{links}(A, V) = \sum_{i \in A, j \in V} W(i, j)$ is the total connection from vertices in $A$ to all vertices in the graph.

Specially, $\text{linkratio}(A, A)$ measures how many links stay within $A$ itself, and its complement $\text{linkratio}(A, V \setminus A)$ measures how many links escape from $A$. In $K$-way partitioning, these two indicators can be measured by the following normalized associations and normalized cuts criteria:

$$\text{knassoc}(\Gamma^K_V) = \frac{1}{K} \sum_{i=1}^{K} \text{linkratio}(V_i, V_i),$$

$$\text{kncuts}(\Gamma^K_V) = \frac{1}{K} \sum_{i=1}^{K} \text{linkratio}(V_i, V \setminus V_i).$$

To obtain a good clustering, both tight connections within partitions and loose connections between partitions is desired. Therefore, we hope to maximize the associations, i.e. Equ. (15) and minimize the cuts, i.e. Equ. (17). Since $\text{knassoc}(\Gamma^K_V) + \text{kncuts}(\Gamma^K_V) = 1$, optimizing only one of them can simultaneously achieve two goals. Finally, for a $K$-way
partitioning of the vertices, we need to solve the problem \( kncuts(\Gamma^K) \),

\[
\text{minimize} = \frac{1}{K} \sum_{l=1}^{K} \text{linkratio}(V_l, V \setminus V_l). \tag{17}
\]

After relaxing this NP-complete problem into an eigenvalue problem and solving it, we can find near-global optima for the \( K \)-way normalized cuts. By applying \( K \)-way normalized cut clustering, we have automatically partitioned a set of activities into multiple clusters. We wonder if the unsupervised process has discovered meaningful activity patterns. In Section 4, we will set several experiments to test the effectiveness of our proposed unsupervised activity discovery method.

4. Experimental Setup

In this section, we introduce the datasets on which the proposed automatic unsupervised grouping method is verified. And to measure the performances, we utilize multiple evaluation metrics, including Purity, Entropy and Precision Recall and F scores.

4.1. Dataset Selection

Two activity databases are adopted in performance evaluation, and one of them is first proposed in this paper which is named HIT-BJUT Dataset and will be published afterwards. The other one is a traditional baseline database for abnormal activity detection, called UMN dataset. The detailed introduction about these two datasets are as follows:

**HIT-BJUT Dataset** is taken in crowded indoor scenes with about 15 people in the shot and it consists of two parts which are captured by two different cameras from two views. We denote the two subsets as Camera1 and Camera2 for short. Camera1 contains 151 videos and among them, 5 for fighting, 1 for shooting, 6 for escaping and 139 for normal activities. Camera2 contains 130 videos, in which, 6 for escaping and 124 for normal behaviors. The resolution of videos is 240×400. The normal and abnormal videos are imbalanced and it conforms to the actual situations where abnormal events occur rarely. Some sample images
from activity videos in our proposed dataset have been shown in Fig. 4 and Fig. 5 which
displays videos in Camera1 and Camera2, respectively.

Fig. 4. Samples from Camera1 of HIT-BJUT dataset.

Fig. 5. Samples from Camera2 of HIT-BJUT dataset.
UMN Dataset is collected from University of Minnesota [8] and consists of 11 clips of crowded escape events shot in 3 different scenes, including both indoor and outdoor. Each video begins with normal behaviors and ends with panic escaping. In our experiment, we further divide these 11 videos into 22 clips, in which, 11 for normal and 11 for panic escaping. Figure 6 shows some sample frames of the UMN dataset.

![Sample frames of UMN dataset.](image)

Fig. 6. Sample frames of UMN dataset.

4.2. Evaluation criteria

The unsupervised discovery method automatically groups the videos in the dataset into several clusters. Each video has one class label. Thus, we can employ common evaluation methods for clustering methods to measure the effectiveness of our proposed unsupervised activity discovery method.

4.2.1. Clustering Purity Criterion

Suppose the unsupervised process has clustered the activities into $K$ clusters, $D = \{D_1, D_2, \cdots, D_K\}$. The purity of each cluster $D_i$ can be calculated by

$$\text{purity}(D_i) = \max_j (Pr(c_j)),$$

where $Pr(c_j)$ is the probability that a randomly selected video from cluster $D_i$ belongs to class $c_j$. The formula above is used to measure the purity of each cluster:

$$\text{purity}(D_i) = \max_j (Pr(c_j)).$$
and the total purity of the clustering result can be obtained by

\[ purity_{total}(D) = \sum_{i=1}^{K} \frac{|D_i|}{|D|} \times purity(D_i), \]  

(19)

where \( Pr(c_j) \) is the percentage of data belongs to class \( c_j \) in cluster \( D_i \), and \( |D_i| \) refers to the total number of videos in cluster \( D_i \).

4.2.2. Entropy of Clusters

Entropy is a measure of randomness, and in this paper, we use it to measure the performance of clustering results. The smaller the entropy, the better the clustering results. Given \( K \) clusters with \( L \) class labels, \( D = \{ D_1, D_2, \ldots, D_K \} \), the entropy of each cluster can be calculated by the following formula:

\[ entropy(D_i) = -\sum_{j=1}^{L} Pr_i(c_j) \log_2 Pr_i(c_j), \]  

(20)

then the total entropy of the clustering result can be obtained by

\[ entropy_{total}(D) = \sum_{i=1}^{K} \frac{|D_i|}{|D|} \times entropy(D_i), \]  

(21)

4.2.3. Precision Recall and F measure

Beside unsupervised clustering evaluation metrics, we also regard abnormal activities as objects to be detected and calculate Precision and Recall scores, and then use F-measure to acquire comprehensive evaluation results. If we use \( tp \) and \( fp \) to represent true positives and false positives in retrieved items, and \( fn \) and \( tn \) is used to denote false negatives and true negatives in records not retrieved. Then, the Precision and Recall are defined by

\[ P = \frac{tp}{tp + fp}, \]  

(22)

\[ R = \frac{tp}{tp + fn}, \]  

(23)

\[ F = \frac{2PR}{P + R}. \]  

(24)
5. Results and Discussions

The experiments are performed separately on the two datasets described in Section 4.1: HIT-BJUT Dataset and UMN Dataset.

5.1. Experiments on HIT-BJUT Dataset

When $K$ is set to 2, the purity and entropy values on two subsets of HIT-BJUT dataset Camera1 and Camera2, along with amounts of normal and abnormal activities in each cluster, are listed in Table 1 to Table 4 respectively. Here, we compare the results of two pooling methods combined with four feature mapping methods. Among them, Table 1 and Table 2 display the results obtained by sum-pooling, Table 2 and Table 4 are derived by max-pooling. The $T$ denotes the method which adopts only temporal saliency as the feature map; Similarly, $S$ denotes the method adopts only spatial saliency map; $T+S$ represents the method combines both temporal and spatial saliency map, and $\text{Diff}$ is the method use frame differences as the feature map.

From the tables, we notice that our results show abnormal with the purity metric. In Table 2, the best method is $\text{Diff}$ whose purity is 0.98, but in fact it performs badly because it only discovers 1 abnormal activity out from the whole dataset. However, purity of the real best method $T+S$, which separates 5 anomalies despite involving 3 normal activities, is 0.808. In addition, in other tables, though the best methods own the highest purities, the differences of various methods are non-significant. We can find that the above problems are caused by the high unbalance of data. Considering purity metric is appropriate only when amounts of normal and abnormal activities are comparable, the highly unbalanced data results in high purities though two classes of videos are not distinguished effectively. And as expected, the smallest entropies all correspond to the best discovery results as shown in the tables. And the differences between different methods have also been widened.
Table 1. Purity and entropy for 2 clusters of Camera1 (abnormal 12, normal 139) by AveP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>3</td>
<td>139</td>
<td>0.979</td>
<td>0.148</td>
<td>0.989</td>
</tr>
<tr>
<td>S</td>
<td>7</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>5</td>
<td>139</td>
<td>0.965</td>
<td>0.218</td>
<td>0.982</td>
</tr>
<tr>
<td>T+S</td>
<td>7</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>5</td>
<td>139</td>
<td>0.965</td>
<td>0.218</td>
<td>0.982</td>
</tr>
<tr>
<td>Diff</td>
<td>2</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>10</td>
<td>139</td>
<td>0.932</td>
<td>0.355</td>
<td>0.966</td>
</tr>
</tbody>
</table>

Table 2. Purity and entropy for 2 clusters of Camera2 (abnormal 6, normal 124) by AveP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>3</td>
<td>0.571</td>
<td>0.985</td>
<td>2</td>
<td>124</td>
<td>0.983</td>
<td>0.120</td>
<td>0.778</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>49</td>
<td>0.980</td>
<td>0.141</td>
<td>5</td>
<td>75</td>
<td>0.938</td>
<td>0.337</td>
<td>0.959</td>
</tr>
<tr>
<td>T+S</td>
<td>5</td>
<td>3</td>
<td>0.625</td>
<td>0.954</td>
<td>1</td>
<td>121</td>
<td>0.992</td>
<td>0.069</td>
<td>0.808</td>
</tr>
<tr>
<td>Diff</td>
<td>1</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>5</td>
<td>124</td>
<td>0.961</td>
<td>0.236</td>
<td>0.980</td>
</tr>
</tbody>
</table>

Table 3. Purity and entropy for 2 clusters of Camera1 (abnormal 12, normal 139) by MaxP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>5</td>
<td>139</td>
<td>0.983</td>
<td>0.207</td>
<td>0.989</td>
</tr>
<tr>
<td>S</td>
<td>7</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>5</td>
<td>139</td>
<td>0.965</td>
<td>0.218</td>
<td>0.982</td>
</tr>
<tr>
<td>T+S</td>
<td>7</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>5</td>
<td>139</td>
<td>0.965</td>
<td>0.218</td>
<td>0.982</td>
</tr>
<tr>
<td>Diff</td>
<td>2</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>10</td>
<td>139</td>
<td>0.932</td>
<td>0.355</td>
<td>0.966</td>
</tr>
</tbody>
</table>
Table 4. Purity and entropy for 2 clusters of Camera2 (abnormal 6, normal 124) by MaxP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th></th>
<th>Cluster2</th>
<th></th>
<th></th>
<th>Total</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>anomaly</td>
<td>normal</td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>3</td>
<td>124</td>
<td>0.976</td>
<td>0.161</td>
<td>0.988</td>
</tr>
<tr>
<td>S</td>
<td>5</td>
<td>71</td>
<td>0.934</td>
<td>0.350</td>
<td>1</td>
<td>53</td>
<td>0.981</td>
<td>0.133</td>
<td>0.959</td>
</tr>
<tr>
<td>T+S</td>
<td>6</td>
<td>123</td>
<td>0.953</td>
<td>0.271</td>
<td>0</td>
<td>1</td>
<td>1.000</td>
<td>0.000</td>
<td>0.977</td>
</tr>
<tr>
<td>Diff</td>
<td>6</td>
<td>123</td>
<td>0.953</td>
<td>0.271</td>
<td>0</td>
<td>1</td>
<td>1.000</td>
<td>0.000</td>
<td>0.977</td>
</tr>
</tbody>
</table>

Table 5 and Table 6 show the Precision, Recall and F scores on two subsets Camera1 and Camera2 of HIT-BJUT Dataset. From these two tables, we can observe that in most cases, the method taking temporal saliency as feature map T achieves the best results. In some rare cases, T+S, that combines both spatial and temporal saliency map, improves the performance slightly. Diff which does not take visual attention model gets very poor results in all case. It is reasonable to get such results. Compared with Diff, temporal saliency map (T) captures the human-focused dynamic changes, which is the essence of activities. And by this way, it restrains the distractions of background information effectively. Therefore it gains the best results in most cases. By contrary, spatial saliency S abstracts humans attention from static video frames. Apart from information abstracted from humans, masses of information from background objects are also involved. So spatial saliency only improves the performance in rare cases but causes negative effects in most cases. In all, it is proved that it is temporal saliency that captures the most crucial and distinctive elements of the crowd activities, and because of that, the proposed unsupervised activity analysis framework has obtained a excellent performance. Note that, to the best of our knowledge, this is the first work that performs unsupervised discovery on crowded activities from a global perspective, so we did not compare our method with any other comparative method.
Table 5. Precision Recall scores and F scores on Camera1.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-Measure</td>
<td>Precision</td>
</tr>
<tr>
<td>T</td>
<td>1.000</td>
<td>0.750</td>
<td><strong>0.857</strong></td>
<td>1.000</td>
</tr>
<tr>
<td>S</td>
<td>1.000</td>
<td>0.583</td>
<td>0.737</td>
<td>1.000</td>
</tr>
<tr>
<td>T+S</td>
<td>1.000</td>
<td>0.583</td>
<td>0.737</td>
<td>1.000</td>
</tr>
<tr>
<td>Diff</td>
<td>1.000</td>
<td>0.167</td>
<td>0.286</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 6. Precision Recall scores and F scores on Camera2.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-Measure</td>
<td>Precision</td>
</tr>
<tr>
<td>T</td>
<td>0.571</td>
<td>0.667</td>
<td><strong>0.615</strong></td>
<td>1.000</td>
</tr>
<tr>
<td>S</td>
<td>0.020</td>
<td>0.167</td>
<td>0.036</td>
<td>0.066</td>
</tr>
<tr>
<td>T+S</td>
<td>0.625</td>
<td>0.833</td>
<td><strong>0.714</strong></td>
<td>0.047</td>
</tr>
<tr>
<td>Diff</td>
<td>1.000</td>
<td>0.167</td>
<td>0.286</td>
<td>0.047</td>
</tr>
</tbody>
</table>

We also explore the case when Camera1 is grouped into 4 clusters. The amounts of various activities in every cluster by four different methods are listed in Table 7. And the purity and entropy results are listed in Table 8. It is indicated that the best result is produced by T+S. It can be explained by that there are a few abnormal videos taken from slight different perspectives from others, which results in the difference of scenes. At this situation, when a more fined clustering is required, spatial saliency (S) would produce positive improvement of the performance. Moreover, temporal saliency (T) would not be expected to cause negative effects to the performance. Therefore, T+S obtains consistently good results as S.
Table 7. Amounts result for 4 clusters of Camera1 (fight 5, gun 1, escaping 6, normal 139) by MaxP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fight</td>
<td>shoot</td>
<td>escape</td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T+S</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Diff</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 8. Evaluation result on Camera1 (fight 5, gun 1, escaping 6, normal 139) for 4 clusters by MaxP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
<th>Cluster3</th>
<th></th>
<th>Cluster4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
<td>Entropy</td>
<td>Purity</td>
</tr>
<tr>
<td>T</td>
<td>0.965</td>
<td>0.220</td>
<td>0.833</td>
<td>0.650</td>
<td>1.000</td>
<td>0.000</td>
<td>0.950</td>
</tr>
<tr>
<td>S</td>
<td>0.965</td>
<td>0.217</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.991</td>
</tr>
<tr>
<td>T+S</td>
<td><strong>0.965</strong></td>
<td><strong>0.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>0.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>0.000</strong></td>
<td><strong>0.991</strong></td>
</tr>
<tr>
<td>Diff</td>
<td>0.965</td>
<td>0.251</td>
<td>0.500</td>
<td>1.000</td>
<td>0.667</td>
<td>0.918</td>
<td>0.658</td>
</tr>
</tbody>
</table>

Besides, we analyze the influences of the size of overlapping cells on the entropy values in Figure 7. The horizontal ordinate values 1 to 9 correspond to size $2 \times 2$, $4 \times 4$, $5 \times 5$, $8 \times 8$, $10 \times 10$, $16 \times 16$, $20 \times 20$, $24 \times 25$, $30 \times 40$, respectively. It can be observed that the
entropy stands at a high value at the first coordinate point, then tends to drop until reaches the lowest point, after which, the entropy raises up with the increase of the cells’ size and eventually reaches a steady state. From the figures, we find that this trend is more likely to be with the best method $T$ and the best performance occurs about the fourth coordinate value, i.e. when the size of overlapping cells is $8 \times 8$.

Fig. 7. The influence of size of overlapping cells on the measure of entropy. Results in group (a) are for Camera1 and results in group (b) are for Camera2. In each group, the left one is for AveP and the right one is for MaxP.
Fig. 8. The discovery results on our HIT-BJUT dataset. (a) is the result of subset Camera1. It can be seen that all videos grouped in cluster 2 are abnormal activities and in cluster 1, only 2 abnormal activities failed to be separated from all 139 normal activities. (b) is the result of subset Camera2. All abnormal activities except one are clustered in cluster 2 with 3 normal activities.

5.2. Experiments on UMN Dataset

The same with HIT-BJUT dataset, for UMN dataset, we employ purity, entropy and precision recall and F scores to measure the discovery performance. The purity and entropy values of multiple methods combining with MaxP are shown in Table 9. Table 10 illustrates
the precision and recall scores and their comprehensive F scores. From the tables, we can find that for the balanced data, both purity and entropy metrics give effective evaluation results. Furthermore, it can be seen that T and T+S get the best results, whereas S gets worse performance. It once again proves that it is temporal saliency map that plays a critical role in capturing information from activity videos. On the other hand, after analyzing, we find that the poor bad result of S probably could be ascribed to that videos in the same activity class are taken in various scenes. Because of that, the background information captured by spatial saliency disturbs the clustering of crowd activities and eventually result in bad performance.

Table 9. Purity and entropy on UMN (abnormal 11, normal 11) for 2 clusters by MaxP.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1 Anomaly</th>
<th>Cluster1 Normal</th>
<th>Cluster2 Anomaly</th>
<th>Cluster2 Normal</th>
<th>Total Purity</th>
<th>Total Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>10</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>0.958</td>
<td>0.226</td>
</tr>
<tr>
<td>S</td>
<td>11</td>
<td>6</td>
<td>0.647</td>
<td>0.936</td>
<td>0.824</td>
<td>0.724</td>
</tr>
<tr>
<td>T+S</td>
<td>10</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
<td>0.958</td>
<td>0.226</td>
</tr>
</tbody>
</table>

Table 10. Precision Recall scores and F scores on UMN dataset.

<table>
<thead>
<tr>
<th></th>
<th>Cluster1</th>
<th></th>
<th>Cluster2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-Measure</td>
<td>Precision</td>
<td>Recall</td>
<td>F-Measure</td>
</tr>
<tr>
<td>T</td>
<td>1.000</td>
<td>0.909</td>
<td><strong>0.953</strong></td>
<td>1.000</td>
<td>0.909</td>
<td><strong>0.953</strong></td>
</tr>
<tr>
<td>S</td>
<td>0.647</td>
<td>1.000</td>
<td>0.786</td>
<td>0.647</td>
<td>1.000</td>
<td>0.786</td>
</tr>
<tr>
<td>T+S</td>
<td>1.000</td>
<td>0.909</td>
<td><strong>0.953</strong></td>
<td>1.000</td>
<td>0.909</td>
<td><strong>0.953</strong></td>
</tr>
</tbody>
</table>

5.3. Time Efficiency Analysis

Time consumption of each stage of our proposed unsupervised crowd activity analysis framework is illustrated in Table 11. It can be seen that the most time-consuming step is the abstraction of features, including saliency-based feature mapping and spatio-temporal
pooling, which took about 447 seconds for 151 videos with 97440 frames in total, while the computation of affinity matrix took 6.438 seconds. Moreover, the process of clustering with N-Cut algorithm only took 0.03 seconds. It is worth noting that although the most largest part of time was cost by feature extraction phase, the average speed reached 218 frames per second, which proves that in spite of effectiveness, the proposed method also achieved high time efficiency.

### Table 11. Time consumption of each stage.

<table>
<thead>
<tr>
<th>Stages</th>
<th>Feature Extraction</th>
<th>Affinity Computation</th>
<th>N-Cut Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Consumption (s)</td>
<td>446.866</td>
<td>6.438</td>
<td>0.03</td>
</tr>
</tbody>
</table>

### 6. CONCLUSION

In this paper, we proposed an unsupervised discovery method for crowd activities. Unlike previous activity analysis methods, we discovered useful information automatically without involving laborious labeling or complex modeling. The proposed spatio-temporal saliency strategy can effectively capture human-focused components in activities. And the experimental results have proved that our proposed saliency based crowd activity discovery method to be fast, robust and effective. There are several aspects to be further studied in the future. For example, a new saliency detection algorithm specific to crowd activities remain to be investigate to extract the unique characteristics of the motion of crowd. We are also studying other applications of the proposed unsupervised discovery framework.

### References


26
[32] Z. Wang, J. Zhang, Detecting pedestrian abnormal behavior based on fuzzy associative memory, in:


[45] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series, in: KDD workshop,


