Interactivity Proposals for Surveillance Videos
Shuo Chen, Pascal Mettes, Tao Hu and Cees G. M. Snoek
University of Amsterdam

ABSTRACT
This paper introduces spatio-temporal interactivity proposals for video surveillance. Rather than focusing solely on actions performed by subjects, we explicitly include the objects that the subjects interact with. To enable interactivity proposals, we introduce the notion of interactivityness, a score that reflects the likelihood that a subject and object have an interplay. For its estimation, we propose a network containing an interactivity block and geometric encoding between subjects and objects. The network computes local interactivity likelihoods from subject and object trajectories, which we use to link intervals of high scores into spatio-temporal proposals. Experiments on an interactivity dataset with new evaluation metrics show the general benefit of interactivity proposals as well as its favorable performance compared to traditional temporal and spatio-temporal action proposals.

KEYWORDS
interactivity detection; action detection; proposal generation

1 INTRODUCTION
The goal of this paper is to generate spatio-temporal proposals that capture the interaction between subjects and objects in surveillance videos. Spatio-temporal proposals in videos are generally focused on actions [15, 17, 21, 41, 49], i.e., centered around subjects only. The objects with which actions might interact are generally ignored or only used implicitly. In surveillance settings, interactions between subjects and objects are key, because they denote important events to analyze. Think about a person entering a car or loading gear into a trunk. Since surveillance videos may contain several events that happen simultaneously, localizing the temporal extent of an interactivity is insufficient; spatial localization is mandatory. We aim to explicitly capture subjects performing actions, and the objects with which they interact, in space and time. We focus on the proposal generation step, where a video is split into spatio-temporal segments, upon which detection algorithms can be applied.

To arrive at spatio-temporal interactivity proposals, we take inspiration from objectness [1] and actionness [8, 44]. These approaches estimate the likelihood of object presence in a spatial region or action presence in a spatio-temporal region. Based on the likelihood, object or action proposals can be generated. Subsequently, such proposals are scored by classifiers to obtain object or action detections. Here, we take this line of work further and introduce interactivityness. Rather than estimating the individual likelihoods of objects or subjects performing an action, we estimate when and where subjects and objects are jointly occurring and are also in interaction. Akin to objectness and actionness, we use interactivityness to obtain interactivity proposals, which we define as pairs of subject and object trajectories with the same start and end time, see Figure 1.

We make three contributions in this paper. First, we introduce the new task of spatio-temporal interactivity proposal generation in surveillance videos. Second, we introduce an interactivity network. This network estimates the interactivityness between a subject and object using an interactivity module that models the context around subjects and objects, as well as a geometric encoding that models the spatial relations of the pair. Third, we set up an interactivity proposal evaluation, including a dataset distilled from the ActEV surveillance benchmark [2] and interactivity evaluation metrics. Experiments on this evaluation show the effectiveness of our approach, outperforming existing approaches from the temporal and spatio-temporal action proposal literature. We will make the dataset, evaluation protocols, and code publicly available.
2 RELATED WORK

2.1 Action proposals

Temporal action proposals. Proposal methods for temporal action localization form an active research topic [4, 10, 12, 13, 25, 27, 32, 51, 52]. Escorcia et al. [10] utilize LSTMs on extracted CNN features to capture temporal information. Buch et al. [4] adopt the C3D network architecture as a feature extractor with a gated recurrent unit to capture long-term temporal information. Gao et al. [13] collect proposal candidates through a sliding window, which utilizes unit-level information for training. For each proposal, the average unit representation is adopted as proposal representation. Afterwards, temporal regression is performed on the unit-level to refine the start and end times of the proposals. Zhao et al. [51] generate actionness for each frame and group continuous frames with high actionness as proposals. All temporal action proposal methods use whole frames as input. In outdoor surveillance settings, many action and interactions can occur at the same time, hence using whole frames as input is not precise enough. Therefore, we target interactivity proposals in both space and time.

Spatio-temporal action proposals. Spatio-temporal action proposals target the spatio-temporal locations of subjects in videos [15, 17, 21, 30, 41, 49]. One common manner to obtain spatio-temporal action proposals is by clustering local voxels or dense trajectories in a hierarchical manner [21, 30, 41]. Yu et al. [49] generate generic action proposals in unconstrained videos by linking subject detections over time. He et al. [17] propose a tupelet proposal network for action detection, which adopts Faster RCNN [33] to collect boxes with high action score. They link the highest scoring boxes to obtain tubelet proposals. Gleason et al. [15] generate spatio-temporal cuboid proposals by clustering detected boxes in spatio-temporal regions, followed by jittering to collect more proposals for better recall. Where current spatio-temporal proposal methods focus on actions only, we target spatio-temporal proposals of both subjects and objects. More concretely, where a spatio-temporal action proposal is described by a single tube, a spatio-temporal interactivity proposal is described by two tubes with the same start and end time. The tubes represent a subject and an object that should be in interaction.

2.2 Visual human-object interaction

A wide range of works have investigated the relationship between humans (subjects) and objects [5, 11, 14, 18, 50] in images. Gkioxari et al. [14] learn to predict an action-specific density over object locations using detected subjects. Chao et al. [5] capture interaction information in images by measuring relative location information between boxes. Xu et al. [48] utilize semantic regularities for human-object interaction detection in images with knowledge graphs. Gao et al. [31] propose an instance-centric attention module that learns to dynamically highlight regions in an image conditioned on the appearance of each instance. Prest et al. [31] previously studied human-object interaction in actor-centric videos, such as Drinking and Smoking. In this setting, the person boxes generally cover the object boxes. In the surveillance domain, we aim for proposals of interactivities with unique boxes for persons and objects by focusing on the surveillance domain. Wang et al. [43] also investigate interactions in videos, but do so for agent-object animations, while we focus on interactivity detections by proposals.

2.3 Video Surveillance

Recognition in video surveillance is a long-standing challenge [6, 23, 24, 28, 40, 42, 46, 52]. Surveillance settings are often indoor with an explicit focus on subjects, as exemplified by the recent benchmark of Zhao et al. [52]. The works of Maguell et al. [36, 37] relates to our work as they focus on tracking loitering activities across multiple surveillance cameras. Our work focuses on capturing interactivity on single surveillance camera, without considering the explicit interactivity class.

The works of Walker et al. [42] and Misra et al. [28] also relate to our work in that both tackle object localization in space and time. In this work, we focus on outdoor surveillance videos with the ActEv benchmark [2] and we focus on jointly capturing the spatio-temporal localization of subjects and objects in interaction. For spatio-temporal action detection, several datasets have been introduced, such as AAvA [16], UCF-Sports [34], and J-HMDB51 [22].

Current datasets are commonly focused on human-centric actions in non-surveillance domains. Only the annotations of subjects is provided, while the spatio-temporal annotations of objects are absent. Hence, we will not consider these datasets for our experiments. Instead, we will set up an interactivity proposal evaluation, including a dataset distilled from the ActEV surveillance video benchmark [2] and interactivity evaluation metrics.

3 METHOD

In order to obtain interactivity proposals from an input video, our approach consists of three components: 1) obtaining interactivity candidates, 2) computing interactivityness, and 3) generating interactivity proposals. The overview of our method is sketched in Figure 2. We will describe each component in detail next.

3.1 Obtaining interactivity candidates

We first generate an over-complete set of interactivity candidates, where each candidate denotes a pair of subject and object trajectories that potentially interact. Due to the possibly overwhelming number of subjects and objects in a surveillance video, evaluating all possible subject and object pairs is infeasible. Physically, a subject can only interact with an object when they are close enough at some point in time. Hence, in most cases, the interactivity only happens when the subject and the object are in close contact with each other.

Suppose we have obtained \( N \) subject trajectories and \( M \) object trajectories in a video. Each trajectory has consecutive bounding boxes. For example, the subject trajectory \( t_s = \{b^1_s, b^2_s, ..., b^n_s\} \) has \( n \) boxes and the object trajectory \( t_o = \{b^1_o, b^2_o, ..., b^m_o\} \) has \( m \) boxes. A box \( b \in \mathbb{R}^4 \) is denoted by the leftmost, topmost, rightmost, and bottommost coordinates. For each frame \( f \) in the video, we calculate the Intersection over Union (IoU) between subject box \( b^f_s \in t_s \) and object box \( b^f_o \in t_o \). If they overlap with each other, i.e., their IoU score is larger than zero at any point in time, we deem the pair as a potential interactivity. In addition, we compute a union box that
We add the union boxes to the subject-object pairs and obtain spatio-temporal interactivity proposals. The main idea of our method is to capture interaction information to aid recognition. We achieve the goal in two ways: (1) We use two interactivity blocks: one to capture the attention between the subject features and the union features, and one for the attention between the object features and the union features. From the above we know a subject-object pair is composed of continuous triplet boxes \( c = \{ (b^s_{1u}, b^s_{1o}), (b^s_{2u}, b^s_{2o}), ..., (b^s_{ku}, b^s_{ko}) \} \). For each frame, the three boxes are first fed to a backbone convolutional neural network to extract features. For frame \( f \), we obtain three box features: union box features \( F^f_u \), subject box features \( F^f_s \) and object box features \( F^f_o \). The three features then form the input to the interactivity block. Let \( F'_c = (F'_s, F'_o, F'_u) \) denote the combined feature set, then the two individual blocks are given as:

\[
\begin{align*}
IB_s (F'_c) &= c_1 (sm(c_2(F'_s)^T \times c_3(F'_u))) \times c_4(F'_u) + F'_s, \\
IB_o (F'_c) &= c_1 (sm(c_2(F'_o)^T \times c_3(F'_u))) \times c_4(F'_u) + F'_o.
\end{align*}
\]

Here \( c_1, c_2, c_3, c_4 \) are convolutional layers with kernel size 1x1 and \( sm \) denotes the softmax function. The output dimensions of \( c_1, c_2, c_3, c_4 \) are 512. We also incorporate Dropout [39], Rescaling, Layer Normalization [3] and matrix transposition operations. The two interactivity blocks’ convolutional layers share weights during training. The two blocks are combined as follows:

\[
IB(p) = IB_s(p) + IB_o(p).
\]

The details of the interactivity blocks are illustrated in Figure 3. Interactivity block operations do not change the dimensionality of input feature. The dimensionality of input features \( F_s, F_o, F_u \) are all \( \mathbb{R}^{C \times H \times W} \), the output feature \( IB(p) \) remains the same.

With the interactivity block, we force the network to focus on both the subject and the object. At the same time, useful contextual information tightly unifies the subject and object boxes as follows:

\[
b^f_{tu} = (\min(b^f_s[0], b^f_o[0]), \min(b^f_s[1], b^f_o[1]), \max(b^f_s[2], b^f_o[2]), \max(b^f_s[3], b^f_o[3])).
\] (1)

We add the union boxes to the subject-object pairs and obtain \( k \) interactivity candidates, each consisting of a triplet of spatio-temporal trajectories, e.g., for temporal length \( k \) candidate \( c \) is denoted as \( c = \{ (b^s_{1u}, b^s_{1o}), (b^s_{2u}, b^s_{2o}), ..., (b^s_{ku}, b^s_{ko}) \} \).

This procedure is performed for test videos to obtain an initial pool of candidates. During training, we use ground truth trajectories of subjects and objects that are known to interact. The interactivity label itself is ignored, only the trajectories are used.

### 3.2 Interactivity network

Given a subject-object pair from our candidate pool, we need to detect whether this pair has any interactivity. If so, we also want to know when it starts and ends. Here we train a binary classifier to estimate the interactivity likelihoods, called interactivityness. (2) We encode the geometric relation between the object features and the union features. From the above we know a subject-object pair is composed of continuous triplet boxes \( c = \{ (b^s_{1u}, b^s_{1o}), (b^s_{2u}, b^s_{2o}), ..., (b^s_{ku}, b^s_{ko}) \} \). For each frame, the three boxes are first fed to a backbone convolutional neural network to extract features. For frame \( f \), we obtain three box features: union box features \( F^f_u \), subject box features \( F^f_s \) and object box features \( F^f_o \). The three features then form the input to the interactivity block. Let \( F'_c = (F'_s, F'_o, F'_u) \) denote the combined feature set, then the two individual blocks are given as:

\[
\begin{align*}
IB_s (F'_c) &= c_1 (sm(c_2(F'_s)^T \times c_3(F'_u))) \times c_4(F'_u) + F'_s, \\
IB_o (F'_c) &= c_1 (sm(c_2(F'_o)^T \times c_3(F'_u))) \times c_4(F'_u) + F'_o.
\end{align*}
\]

The details of the interactivity blocks are illustrated in Figure 3. Interactivity block operations do not change the dimensionality of input feature. The dimensionality of input features \( F_s, F_o, F_u \) are all \( \mathbb{R}^{C \times H \times W} \), the output feature \( IB(p) \) remains the same.
Figure 3: Interactivity block details. The two interactivity blocks share convolution layer weights with each other. The input are subject box feature \( f_s \), object box feature \( f_o \) and union box feature \( f_u \). Here \( \oplus \) denotes element-wise sum and \( \odot \) denotes matrix product. LN is short for Layer Normalization.

We calculate the temporal Intersection over Union (tIoU) between proposal candidates and ground truths. We collect two types of proposal samples: (1) positive proposals, i.e., those overlap with the closest ground truth with at least 0.5 tIoU; (2) negative proposals, i.e., those that do not overlap with any ground truth. Due to the sparsity of ground truth proposals, the number of negative proposals is much higher than the number of positive proposals. We adopt the weighted cross-entropy loss function to deal with this class imbalance:

\[
\mathcal{L} = -\omega_y(y \log(s) + (1 - y) \log(1 - s)),
\]

where \( s \) denotes the interactivityness output from Eq. 6, \( y \) the ground truth label, and \( \omega_y \) the class-dependent weight used for balancing the positive and negative samples.

### 3.3 Interactivity proposal generation

For a subject-object pair, our network provides an interactivity score per frame. To generate spatio-temporal interactivity proposals, we rely on the 1D-watershed algorithm [35]. The main idea is to find continuous temporal segments with high interactivityness to generate proposals. The watershed algorithm was originally used as a segmentation method and later for temporal action proposal generation [51]. We first feed the boxes from the automatically computed candidate pairs to obtain frame-level interactivityness. Then, we regard the interactivityness score as a 1D terrain with heights and basins. This method floods water on this terrain with different "levels" \( y \), resulting in a series of "basins" filled with water, named by \( G(y) \). Each obtained basin corresponds to a segment with high interactivityness. Starting from the initial basins, we merge consecutive basins until their length is above a temporal threshold \( r \). We uniformly sample \( r \) and \( y \) with step 0.05. By using multiple values for the two thresholds, multiple sets of regions are generated. We average the interactivityness for each region as the proposal score. We repeat this procedure for all selected pairs of subjects and objects. Finally, we apply non-maximum suppression on all generated proposals to remove redundant proposals. The final output is a set of spatio-temporal interactivity proposals for a video.

### 4 EXPERIMENTAL SETUP

#### 4.1 KIEV dataset

To accommodate the new task of spatio-temporal interactivity proposals, we have distilled a subset from the NIST TRECVID ActEV (Activities in Extended Video) dataset, a collection of surveillance videos with spatio-temporal annotations for objects and subject [2]. ActEV is an extension of the VIRAT dataset [29]. Since not all actions in ActEV are interactions, we leverage a subset of ActEV that explicitly focuses on interactivities and call this the KIEV (Key Interactivities in Extended Video) dataset. KIEV includes high-resolution surveillance videos that are 1080p or 720p. In KIEV, the subject is a person and the object could be a person, vehicle or door. We select nine key interactivities from ActEV, namely Closing, Closing Trunk, Entering, Existing, Loading, Opening, Opening Trunk, Unloading and Person Person Interaction. Note that we do not use the interactivity labels in our approach, we are class-agnostic and are merely interested in recognizing their spatio-temporal locations. The training

...
We apply this model on the unseen KIEV validation frames to obtain the geometric features and obtain a 1032-dimensional representation. The temporal context is beneficial for recognizing interactivities.

We also remove pairs whose duration is shorter than one second.

The temporal context is beneficial for recognizing interactivities.

Subject-object pairing. When pairing subjects and objects, we temporally extend each pair with three seconds in both directions. This ensures our choice of generating interactivity candidates based on overlap.

5 RESULTS

We consider three experiments: (i) we ablate the effectiveness of our interactivity networks, (ii) we assess the effect of automatic trackers over ground truth spatial locations, and (iii) we compare to other proposal methods.

5.1 Ablating the interactivity network

In the first experiment, we evaluate the two core components of our interactivity network: the interactivity block and the geometric encoding. The baseline method does not contain these two components. For the baseline we sum the subject feature, object feature and union feature obtained from CNN backbone together. Then we input the summed feature into classifier. We use the Average Temporal Recall (ATR), measures the temporal alignment between proposals and ground truth interactivities. This metric is commonly used for temporal action proposals, e.g. [12, 13, 51]. A proposal is a true positive if its temporal intersection over union (tIoU) with a ground truth is greater than or equal to a given threshold. ATR is the mean of all recall values using tIoU between 0.5 to 0.9 (inclusive) with a step size of 0.05.

For all interactivity instances, the subjects overlap with the object trajectory in interactivity proposal instances of KIEV.

 proposal is not closer to another unmatched ground truth interactivity. The term vIoU refers to the voluminal Intersection over Union and is calculated as vIoU = (tube of overlap) / (tube of union). We report ATR25, ATR50, as well as the AUC (Area Under Curve) to see how well the proposal method works across all thresholds for number of proposals per video.

Average Spatial Recall. The second metric, Average Spatial Recall (ASR), is adapted from the AVA dataset [16]. We compare predicted boxes in each frame with ground truth boxes. If their overlaps are above a threshold of 0.5, we regard the predicted box as a true positive. We evaluate frame by frame to get the final recall.

5.2 Implementation details

Object detection and tracking. We use Faster R-CNN [33] with a ResNet-101 [19] backbone with dilated convolutions and feature pyramids [26] for person and vehicle detection. We use the model provided by [7]. The model is trained on the ActEV training set [2]. We apply this model on the unseen KIEV validation frames to obtain vehicle and person boxes. We rely on the Deep SORT tracking algorithm [47], to generate person and vehicle trajectories. During the tracking procedure, we use the boxes and Region of Interest [18] features from the detection model to link detected subjects and objects into trajectories.

We consider three evaluation metrics, which measure the temporal, spatial, and spatio-temporal proposal quality.

Subject-object pairing. When pairing subjects and objects, we temporally extend each pair with three seconds in both directions. This temporal context is beneficial for recognizing interactivities. We also remove pairs whose duration is shorter than one second.

Subject-object pairing. When pairing subjects and objects, we temporally extend each pair with three seconds in both directions. This temporal context is beneficial for recognizing interactivities. We also remove pairs whose duration is shorter than one second.

Interactivity network. We use the BN-Inception model provided by [51] as the feature extraction backbone. The model is pre-trained on ImageNet [9]. The interactivity network is inserted before the global average pooling layer. We use the features after the global pool layer, whose dimensionality is 1024×7×7. After spatially pooling the feature from the interactivity network, we concatenate them with the geometric features and obtain a 1032-dimensional representation. The backbone, interactivity network, and interactivityness classifier are jointly optimized on the KIEV training set. All boxes are resized to 224 × 224 to meet the input dimension of BN-Inception. We train our model for 100 epochs using Adam with learning rate 1e-5, exponential decay rate 0.9, decay rate 0.999, and weight decay 5e-4. We follow [51] to set other parameters.

Proposal generation. A 1D Gaussian filter with kernel size 3 is applied to smooth the interactivityness sequence. We then apply non-maximum suppression with temporal overlap threshold 0.7 to filter out overlapping proposals.

Table 1: Ablating the interactivity network based on temporal average recall (%). Both the interactivity block and the geometric encoding aid the proposal quality. Their combination works best. The results prove the efficiency of our method.

<table>
<thead>
<tr>
<th>Interactivity Block</th>
<th>Geometric Encoding</th>
<th>Average Temporal Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>ATR25: 6.9 ATR50: 10.9 AUC: 10.1</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓</td>
<td>ATR25: 10.6 ATR50: 15.5 AUC: 9.6</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓</td>
<td>ATR25: 12.4 ATR50: 19.0 AUC: 11.3</td>
</tr>
</tbody>
</table>

Table 2: Effect of automatic tracks on temporal and spatio-temporal proposal quality. For temporal recall, switching from ground truth to automatic trajectories has minimal effect on performance. For spatio-temporal recall, the scores naturally have a larger drop. Automatic tracks are robust enough for temporal proposal quality, but not for spatio-temporal quality.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Temporal ATR25</th>
<th>Temporal ATR50</th>
<th>Temporal AUC</th>
<th>Spatio-Temporal STR25</th>
<th>Spatio-Temporal STR50</th>
</tr>
</thead>
<tbody>
<tr>
<td>ground truth</td>
<td>12.4</td>
<td>19.0</td>
<td>11.3</td>
<td>20.0</td>
<td>23.3</td>
</tr>
<tr>
<td>automatic</td>
<td>11.6</td>
<td>17.6</td>
<td>10.8</td>
<td>6.3</td>
<td>7.8</td>
</tr>
</tbody>
</table>

Figure 5: Ablating the interactivity network by increasing retrieved proposals. When using both the interactivity block and the geometry encoding we obtain best average recall.

Figure 6: Number of example proposals when using automatic trackers for the subject and object trajectories. The qualitative results indicate the difficult nature of the problem of finding spatio-temporal interactivities. Due to occlusions and tiny object sizes, there are some missed detection of interactivity in this dataset, as visualized in Figure 6c. Improved detection will positively affect interactivity proposal generation.

5.2 Effect of automatic tracks

Next, we evaluate the effect of using automatic tracks for subjects and objects on the interactivity proposal quality. We report both the temporal proposal quality (ATR) and spatio-temporal quality (STR) and show results in Table 2.

When evaluating the temporal dimension only, we find that automatic tracks are competitive with ground truth subject and object tubes. Indicating our method is temporally robust to noise in the spatial locations of subjects and objects. Table 2 also shows the spatio-temporal proposal quality is directly impacted by the switch from ground truth to automatic tracks. This is not surprising, since the spatio-temporal evaluation metric is very strict in its spatial evaluation; both the subject and object boxes need sufficient overlap. In Figure 6, we show a number of example proposals when using automatic trackers for the subject and object trajectories. The qualitative results indicate the difficult nature of the problem of finding spatio-temporal interactivities. Due to occlusions and tiny object sizes, there are some missed detection of interactivity in this dataset, as visualized in Figure 6c. Improved detection will positively affect interactivity proposal generation.

5.3 Comparison to prior work

In the third experiment, we compare our approach to several baselines from both the temporal and spatio-temporal action proposal literature, to show that proposing spatio-temporal interactivity locations can not be achieved by existing action proposal methods.

Baseline. We compare to two temporal proposal baselines and one spatio-temporal baseline. The first temporal proposal baseline is TAG from Zhao et al. [51], which proposes temporal regions based on actionness grouping. The second temporal proposal baseline is TURN-TAP from Gao et al. [13], which is based on sliding windows. The spatio-temporal baseline is by Gleason et al. [15], who introduce a spatio-temporal proposal cuboid approach for actions. For a fair comparison, the input object boxes are the same as our approach.
Figure 6: Qualitative results. (a). The top two examples show successful cases, where the proposal highly overlaps in space and time with the ground truth. From top to bottom the interactivities are Entering, Exiting, Closing, Entering and Person Person Interaction. Note that we do not output labels. Here the labels are only for clarifying. The bottom two examples show failure cases, (b). occlusion and (c). small object sizes either result in a low interactivityness or even missed subject and object trajectories. These failure cases highlight the difficult nature of finding interactivities in outdoor settings.
Table 3: Temporal comparison of our interactivity proposals versus regular action proposals. Our method outperforms alternatives.

<table>
<thead>
<tr>
<th>Method</th>
<th>ATR25</th>
<th>ATR50</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhao et al. [51]</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Gleason et al. [15]</td>
<td>1.4</td>
<td>1.6</td>
<td>1.2</td>
</tr>
<tr>
<td>Gao et al. [13]</td>
<td>8.1</td>
<td>12.4</td>
<td>7.4</td>
</tr>
<tr>
<td>This paper</td>
<td>11.6</td>
<td>17.6</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Table 4: Spatio-temporal comparison of our interactivity proposals versus a regular action proposal in terms of Recall (%). Explicitly modeling interactivity results in better spatio-temporal localization.

<table>
<thead>
<tr>
<th>Method</th>
<th>ASR</th>
<th>STR25</th>
<th>STR50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gleason et al. [15]</td>
<td>8.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>This paper</td>
<td>61.5</td>
<td>4.8</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Temporal comparison. Since temporal action proposal methods only provide the start and end times, we first compare our proposals to all baselines using the temporal quality metrics. The results are shown in Table 3 and Figure 7. Our approach performs better than all baselines. In comparison to the best scoring baseline of Gao et al. [13], our method improves the ATR25 by 3.5, the ATR50 by 5.2, and the AUC by 3.4. The approaches of Zhao et al. [51] and Gao et al. [13] fail to generate efficient proposals in this setting because they take the whole frame as input. Since interactivities are only a small part of the video spatially, their representations hardly capture the precise interactions, as expected. These temporal action localization methods fail to solve the interactivity proposal problem. They are capable of localizing temporal boundaries but ignore spatial boundaries. Our approach operates locally in space, which allows for a better estimation of interactivities in time. The approach of Gleason et al. [15] does operate locally in space, but does not explicitly capture contextual and geometric relations between subjects and objects, which results in lower recall scores.

Spatio-temporal comparison. In Table 4, we also compare our approach to Gleason et al. [15] with respect to the spatio-temporal proposal quality. The results show that spatially, the baseline obtains an ASR of 8.4, while we reach a score of 61.5, a considerable gain. Furthermore, the spatio-temporal recall at both 25 and 50 proposals per video is 0 for the baseline, compared to 4.8 and 6.3 for our approach. The reason for this gap in performance is because the baseline generates cuboid-style proposals, leading to coarse spatial localization of subjects and objects. The cuboid-style proposals have low IoUs compared to trajectory-style ground truths. In our evaluation, we care about a precise dynamic alignment in space and time for subjects and objects. Our approach yields more accurate spatio-temporal interactivity proposals, be it the overall spatio-temporal recall is modest. Compared to Gleason et al. [15] we conclude that our approach is better equipped to find interactivities more precisely in space and time.

6 CONCLUSION

This paper introduces interactivity proposals for video surveillance. Rather than focusing on the actions of the subject only, our proposals capture the interplay between subjects and objects in space and time. To that end, we propose a network to compute interactivity between subjects and objects from which we generate class-agnostic proposals. We evaluate the proposals on an interactivity dataset with new overlap metrics, where experiments show the improvement of our approach over traditional temporal and spatio-temporal action proposal methods. Overall, the results are far from perfect, indicating the challenging nature of the problem. To encourage further progress on recognizing interactivity proposals we make the dataset split, evaluation metrics, and code publicly available.

REFERENCES

[17] Jiawei He, Zhiwei Deng, Mostafa SIbrahim, and Greg Mori. 2018. Generic tubelet proposals for action localization. In WACV.
[23] Bingjie Xu, Yongkang Wong, Junnan Li, Qiang Ji. 2016. Actionness estimation using hybrid fully convolutional networks. In CVPR.
[29] Hu Zhang, Yulan Li, and Junsong Yuan. 2015. Fast action proposals for human action detection and search. In CVPR.
[40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In CVPR.
[52] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and real-time tracking with a deep association metric. In ICIP.
[53] Bingjie Xu, Yongkang Wong, Junnan Li, Qiang Ji. 2016. Actionness estimation using hybrid fully convolutional networks. In CVPR.