Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos

Romero Morais[1*], Vuong Le[1], Truyen Tran[1], Budhaditya Saha[1], Moussa Mansour[2,3], Svetha Venkatesh[1]
[1]Applied Artificial Intelligence Institute, Deakin University, Australia
[2]iCetana, Inc. | [3]University of Western Australia, Australia
[1]{ralmeidabaratad,vuong.le,truyen.tran,budhaditya.saha,svetha.venkatesh}@deakin.edu.au
[2]moussa@icetana.com.au

Abstract

Appearance features have been widely used in video anomaly detection even though they contain complex entangled factors. We propose a new method to model the normal patterns of human movements in surveillance video for anomaly detection using dynamic skeleton features. We decompose the skeletal movements into two sub-components: global body movement and local body posture. We model the dynamics and interaction of the coupled features in our novel Message-Passing Encoder-Decoder Recurrent Network. We observed that the decoupled features collaboratively interact in our spatio-temporal model to accurately identify human-related irregular events from surveillance video sequences. Compared to traditional appearance-based models, our method achieves superior outlier detection performance. Our model also offers "open-box" examination and decision explanation made possible by the semantically understandable features and a network architecture supporting interpretability.

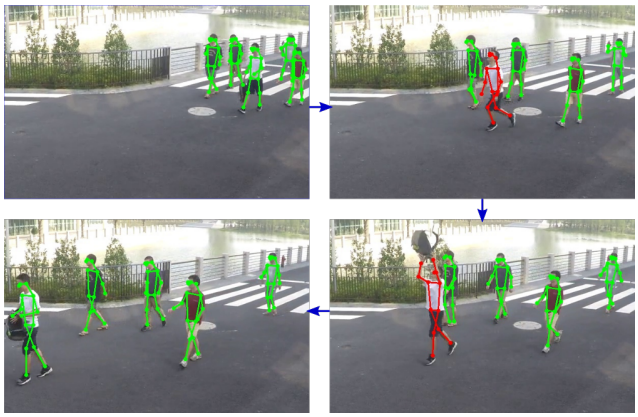

Figure 1. We detect human-related anomalies in video by learning regular spatio-temporal patterns of skeleton features. In this example, we detect the anomalous event of a person catching a backpack. This anomaly is detected by using his unusual skeleton pose and motion compared to those of normal activities. Skeletons in red denote high anomaly scores, while skeletons in green denote low anomaly scores. The order of the frames is specified by the blue arrows.

1. Introduction

Video anomaly detection is a core problem of unsupervised video modeling. An effective solution is learning the regular patterns in normal training video sequences in an unsupervised setting, based on which irregular events in test videos can be detected as outliers. The problem is challenging due to the lack of human supervision and the ambiguous definition of the human-perceivable abnormality in video events. Most current approaches operate on pixel-based appearance and motion features. These features are usually extracted from whole frames [6, 14, 20, 22, 27], localized on a grid of image patches [26], or concentrated on pre-identified regions [7, 15]. Unfortunately, pixel-based features are high-dimensional unstructured signals sensitive to noise, that mask important information about the scene [30]. Furthermore, the redundant information present in these features increases the burden on the models trained on them to discriminate between signal and noise.

Another key limitation of current methods is the lack of interpretability due to the semantic gap between visual features and the real meaning of the events. This limitation can be amplified through processing in deep neural networks [3]. This lack of understanding prevents practitioners from using domain knowledge to customize model architectures and obstructs error analysis.

In this paper, we propose to leverage 2D human skeleton trajectories for detecting abnormal events related to human behavior in surveillance videos. The skeleton trajectories contain the locations of a collection of body joints in the spatio-temporal domain of video sequences, as illustrated in Figure 1. By using skeleton features we explicitly exploit the common structure of surveillance videos, which consists of humans and objects attached to them moving on top of a static background. Compared to appearance-based rep-

resentations, skeleton features are compact, strongly structured, semantically rich, and highly descriptive about human action and movement, which are keys to anomaly detection.

By studying the human skeleton dynamics in a large collection of surveillance videos, we observed that human behavioral irregularity can be factorized into a few factors regarding body motion and posture, such as location, velocity, direction, pose and action. Motivated by this natural factorization, we propose to decompose the dynamic skeleton motions into two sub-processes, one describing global body movement and the other local body posture. The global movement component tracks the dynamics of the whole body in the scene, while the local posture describes the skeleton configuration in the canonical coordinate frame of the body's bounding box, where the global movement has been factored out.

We jointly model the two sub-processes in a novel model called Message-Passing Encoder-Decoder Recurrent Neural Network (MPED-RNN). The network consists of two RNN branches dedicated to the global and local feature components. The branches process their data separately and interact via cross-branch message-passing at each time step. The model is trained end-to-end and regularized so that it distills the most compact profile of the normal patterns of training data and effectively detects abnormal events. In addition to anomaly detection, MPED-RNN supports open-box interpretation of its internal reasoning by providing the weights of the contributing factors to the decision and the visualization of these factors. We trial our method on two of the most challenging video anomaly datasets and compare our results with the state-of-the-art on the field. The results show that our proposed method is competitive in detection performance and easier to analyze the failure modes.

2. Related Work

2.1. Video anomaly detection

Unsupervised video anomaly detection methods have been an old-timer topic in the video processing and computer vision communities. Traditional approaches consider video frames as separate data samples and model them using one-class classification methods, such as one-class SVM [27] and mixture of probabilistic PCA [16]. These methods usually attain suboptimal performance when processing large scale data with a wide variety of anomaly types.

Recent approaches rejuvenate the field by using convolutional neural networks (CNN) to extract high-level features from video frame intensity and achieve improved results. Some of these methods include Convolutional autoencoder [14], spatio-temporal autoencoder [6], 3D Convnet AE [29], and Temporally-coherent Sparse Coding Stacked-RNN [22]. Acknowledging the limitation of the intensity based features such as sensitivity to appearance noise, Liu *et al*. [20] proposed to use the prediction of optical flow in their temporal coherent loss, effectively filtering out parts of the noise in pixel appearance. However, optical flow is costly to extract and still far from the semantic nature of the events.

Structured representations have recently attracted increased attention for its potential to get closer to the semantic concepts present in the anomalies. In [28], object trajectories were used to guide the pooling of the visual features so that interesting areas are paid more attention to. Towards model interpretability, Hinami *et al*. [15] proposed to use object, attribute, and action detection labels to understand the reason of abnormality scores. Although it works well for a number of events, their method fails in many cases due to the incompleteness of the label sets and the distraction from unrelated information in the labels.

Our method of using skeleton features is another step towards using low-dimensional semantic-rich features for anomaly detection. We also advance the research efforts toward interpretability of the anomaly detection models by providing the ability to explain every abnormal event in our factorized semantic space.

2.2. Human trajectory modeling

Human motion in video scenes is an important factor for studying social behavior. It has been applied in multiple computer vision applications, mostly with supervised learning tasks such as action recognition [8, 19] and person re-identification [9]. Recently, more effort has been invested into unsupervised learning of human motion in social settings [1, 13, 25] and single pose configuration [11]. In this work, we propose to expand the application of skeleton motion features to the task of video anomaly detection. In MPED-RNN, we share the encoder-decoder structure with most unsupervised prediction models. However, instead of perfectly generating the expected features, we aim at distilling only the principal feature patterns so that anomalies are left out. This involves building a highly regulated autoencoder.

Regarding feature representation and modeling, apart from traditional methods that rely on hand-crafted local features and state machines, several recent works proposed to use interacting recurrent networks for motion modeling in social settings [1, 13]. In these approaches, the input data are in the form of whole body xy-location sequences while local postures are ignored. To bridge this gap, in [11, 25] the input of the RNN is extended to the skeleton joint locations. Recently, Du *et al*. [8] proposed to divide the skeleton joints into five parts, which are jointly modeled in a five-branch bidirectional neural network. Different to previous approaches, we factorize skeleton motion based on natural decomposition of human motion into global movement/local deformation and model them jointly in an inter-

active recurrent network.

3. Method

The anomalous human-related events in a surveillance video scene can be identified by the irregular human movement patterns observed in the video. Our method detects those anomalies by learning a regularity model of the dynamic skeleton features found in training videos. We assume the skeleton trajectories have already been extracted from the videos. At each time step t, a skeleton is represented by a set of joint locations in image coordinates $f_t = \left(x_t^i, y_t^i\right)_{i=1..k}$, where k is the number of skeleton joints. This set of temporal sequences is the input to our anomaly detection algorithm.

3.1. Skeleton Motion Decomposition

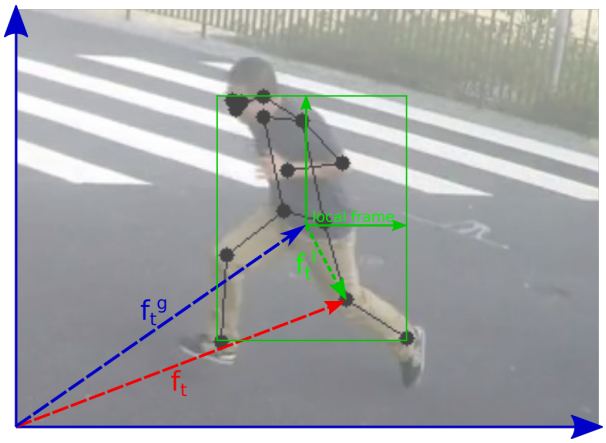

Figure 2. Global and local decomposition of a skeleton in a frame. Based on the canonical local reference frame defined by the green bounding box, the location vector of the left knee joint f_t (dashed red) is decomposed into global f_t^g (dashed blue) and local f_t^l (dashed green) components. The bounding box's width and height are included as global features and used to normalize local features.

Naturally, human motion consists of two factors: rigid movement of the whole body and non-rigid deformation of the skeleton joints. The simplest way to model human motion using a recurrent network is by feeding it the raw sequence of skeleton trajectories in image coordinates $f_t = \left(x_t^i, y_t^i\right)$ [11, 25], implicitly merging the global and local factors together. This solution performs well in videos with uniform skeleton scales and types of activities where the contribution of the two factors is consistent. On realistic surveillance videos, however, the scales of human skeletons vary largely depending on their location and actions. For skeletons in the near field, the observed motion is mainly influenced by the local factor. Meanwhile, for skeletons in the far field, the motion is dominated by the global movement while local deformation is mostly ignored.

Inspired by the natural composition of human skeleton motion and motivated by the factorization models widely used in statistical modeling, we propose to decompose the skeletal motion into "global" and "local" components. The global component carries information about the shape, size and rigid movement of the human bounding box. The local component models the internal deformation of the skeleton and ignores the skeleton's absolute position in relation to the environment.

Geometrically, we set a canonical reference frame attached to the human body (called *local frame*), which is rooted at the center of the skeleton's bounding box. The global component is defined as the absolute location of the local frame center within the original image frame. On the other hand, the local component is defined as the residue after subtracting the global component from the original motion. It represents the relative position of skeleton joints with respect to the bounding box. This decomposition is illustrated in Figure 2 and can be written in 2D vector space as:

$$f_t^i = f_t^g + f_t^{l,i} \tag{1}$$

In 2D image space, xy-coordinates alone poorly represent the real location in the scene because the depth is missing. However, the size of a skeleton's bounding box is correlated with the skeleton's depth in the scene. To bridge this gap, we augment the global component with the width and height of the skeleton's bounding box $f^g = (x^g, y^g, w, h)$ and use them to normalize the local component $f^{l,i} = (x^{l,i}, y^{l,i})$. These features can be calculated from the input features as:

$$\begin{aligned} x^g &= \frac{max(x^i) + min(x^i)}{2}; \quad y^g = \frac{max(y^i) + min(y^i)}{2} \\ w &= max(x^i) - min(x^i); \quad h = max(y^i) - min(y^i) \end{aligned} \tag{2}$$

$$x^{l,i} = \frac{x^i - x^g}{w}; \quad y^{l,i} = \frac{y^i - y^g}{h} \tag{3}$$

The global and local dynamics can be modeled separately as two concurrent sub-processes. In generic videos, these two processes can even manifest independently. For example, a person can move her limbs around while keeping her global location relatively still. Similarly, a person riding a motorbike can move around while having a relatively fixed pose. However, given a specific context, regular human activities contain a strong correlation between these two components. Therefore breaking the cross-component correlation is also a sign of abnormality. In the previous examples, if those actions occurred in the scene where people were normally walking, they would be valid anomaly events. In the next section, we present how both individual dynamic patterns and the relationship between these two components are modeled in our MPED-RNN model.

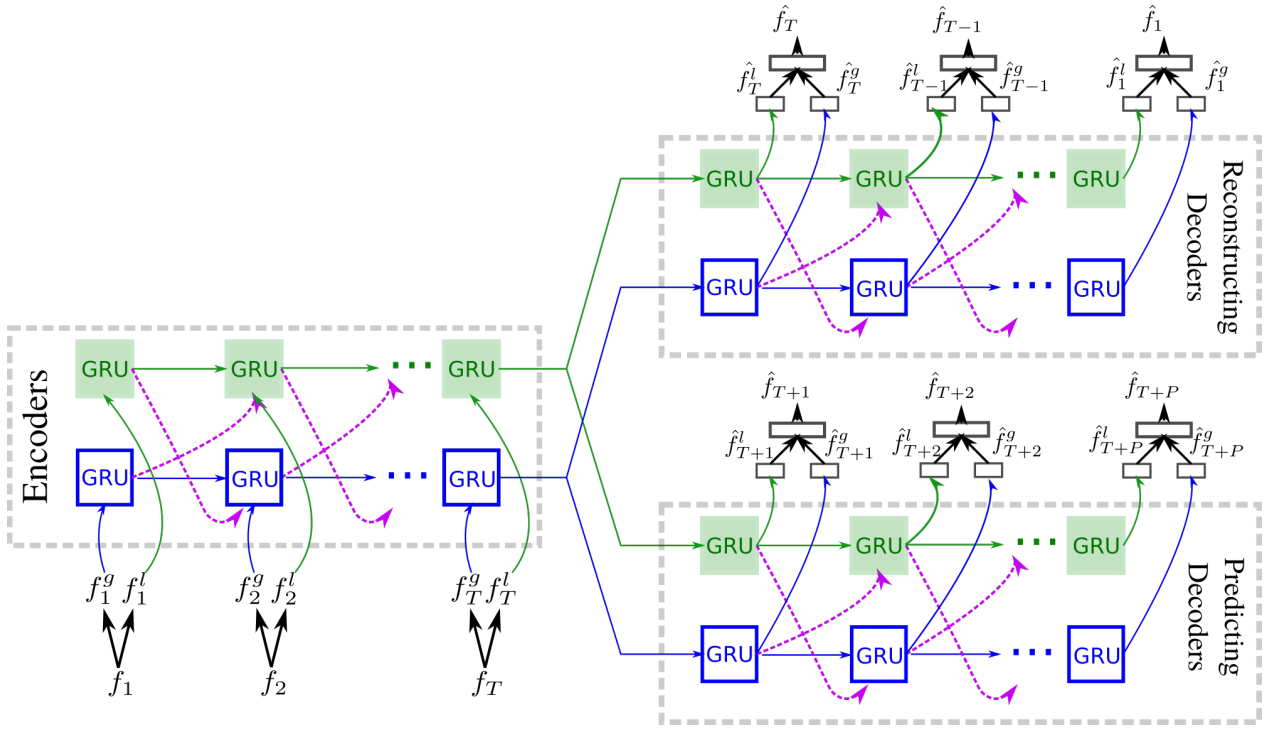

Figure 3. MPED-RNN consists of two interacting branches for two skeleton feature components. The local branch is drawn in green with shaded GRU blocks and the global branch is drawn in blue with transparent GRU blocks. The two components interact through messages (purple dashed) exchanged between the branches. The outputs are generated by a set of MLPs, represented by black rectangles.

3.2. MPED-RNN Architecture

MPED-RNN models the global and local components as two interacting sub-processes where the internal state of one process is used as extra features to the input of the other process. More specifically, the model consists of two recurrent encoder-decoder network branches, each of them dedicated to one of the components. Each branch of the model has the single-encoder-dual-decoder architecture with three RNNs: Encoder, Reconstructing Decoder and Predicting Decoder. This structure is similar to the composite LSTM autoencoder (LSTM AE) of Srivastava *et al*. [24]. However, unlike LSTM AE, MPED-RNN does not only model the dynamics of each individual component, but also the interdependencies between them through a cross-branch message-passing mechanism. We use Gated Recurrent Units (GRU) [5] in every segment of MPED-RNN for its simplicity and similar performance to LSTM [12]. At each time step, the GRU unit of one branch receives a message from the other branch informing its internal state at the previous time step. This message is incorporated into the GRU structure by treating it as an additional input. The same procedure is applied to the other branch. MPED-RNN's architecture is depicted in Figure 3.

Given an input skeleton segment of length T, we first initialize the hidden states of all the GRUs to null. Then, for each time step t, the skeleton f_t is decomposed into f_t^g and f_t^l (using Eqs. (1), (2) and (3)), which are input to the global encoder (E^g) and to the local encoder (E^l), respectively. The messages to be exchanged between the global and local branches are computed as specified by Eqs. 4 and 5 below.

$$m_t^{l\to g} = \sigma\left(W^{l\to g} h_{t-1}^l + b^{l\to g}\right) \tag{4}$$

$$m_t^{g\to l} = \sigma\left(W^{g\to l} h_{t-1}^g + b^{g\to l}\right) \tag{5}$$

For $t = 1, 2, \ldots, T$, the global and local segments are encoded using Eqs. 6 and 7:

$$E^{ge} : h_t^{ge} = \text{GRU}\left(\left[f_t^g, m_t^{le\to ge}\right], h_{t-1}^{ge}\right) \tag{6}$$

$$E^{le} : h_t^{le} = \text{GRU}\left(\left[f_t^l, m_t^{ge\to le}\right], h_{t-1}^{le}\right) \tag{7}$$

After encoding the input segments, the global and local Reconstructing Decoders initialize their hidden states as $h_T^{gr} = h_T^{ge}$ and $h_T^{lr} = h_T^{le}$, respectively, and for $t = T, T-1, ..., 1$, we have:

$$D_r^g : h_{t-1}^{gr} = \text{GRU}(m_t^{lr\to gr}, h_t^{gr}) \tag{8}$$

$$D_r^l : h_{t-1}^{lr} = \text{GRU}(m_t^{gr\to lr}, h_t^{lr}) \tag{9}$$

Similarly, the global and local Predicting Decoders initialize their hidden states as $h_T^{gp} = h_T^{ge}$ and $h_T^{lp} = h_T^{le}$, respectively, and for $t = T+1, T+2, \ldots, T+P$, we have:

$$D_p^g : h_t^{gp} = \text{GRU}(m_t^{lp\to gp}, h_{t-1}^{gp}) \tag{10}$$

$$D_p^l : h_t^{lp} = \text{GRU}(m_t^{gp\to lp}, h_{t-1}^{lp}) \tag{11}$$

In training, the dual decoders in the MPED-RNN's architecture jointly enforce the encoder to learn a compact representation rich enough to reconstruct its own input and predict the unseen future. Meanwhile, in testing, the abnormal patterns cannot be properly predicted because they were neither seen before nor follow the normal dynamics.

In each decoder network, the projected features of the corresponding decoders, $\hat{f}_t^g$ and $\hat{f}_t^l$, are independently generated from the hidden states h_t^g and h_t^l by fully-connected layers. These two projected features are concatenated and input to another fully-connected layer, which generates the projected perceptual feature $\hat{f}_t$ in the original image space. Ideally, $\hat{f}_t$ can be calculated from $\hat{f}_t^g$ and $\hat{f}_t^l$ by inverting Eqs. (2) and (3). However, by being projections into low-dimensional subspaces, a direct computation is unlikely to be optimal. Thus, using a fully-connected layer to learn the inverse mapping allows the computation to be robust to noise. These projected features are used to evaluate the conformity of an input sequence of skeletons to the learned normal behavior and hence are used to build the loss function for training and score function for testing. These procedures are detailed next.

3.3. Training MPED-RNN

Training setup The trajectory of a person can span many frames in a video. However, recurrent networks are trained on fixed-size sequences. To cope with this issue, we extract fixed-size segments from every skeleton's trajectory using a sliding-window strategy. Therefore, each segment is computed as:

$$\text{seg}_i = \{f_t\}_{t=b_i..e_i} \tag{12}$$

where b_i and e_i are beginning and ending indices of the i-th segment calculated from the chosen sliding stride s and segment length T:

$$b_i = s \times i;\ e_i = s \times i + T \tag{13}$$

During training, batches of training segments are decomposed into global and local features, which are input to MPED-RNN.

Loss functions We consider three loss functions defined in three related coordinate frames. The Perceptual loss L_p constrains MPED-RNN to produce the normal sequences in the image coordinate system. The Global loss L_g and the Local loss L_p act as regularization terms that enforce that each encoder-decoder branch of MPED-RNN work as designed. Each of the losses includes the mean squared error made by the reconstructing and predicting decoders:

$$L_*(\text{seg}_i) = \frac{1}{2}\left(\frac{1}{T}\sum_{t=b_i}^{e_i}\left\|\hat{f}_t^* - f_t^*\right\|_2^2 + \frac{1}{P}\sum_{t=e_i+1}^{e_i+P}\left\|\hat{f}_t^* - f_t^*\right\|_2^2\right) \tag{14}$$

where P denotes the prediction length and $*$ represents one of l, g or p. In case of p notice that it makes f_t^p equal to f_t of Section 3.1. The prediction loss is truncated if the end of trajectory is reached within the prediction length.

The three losses contribute to the combined loss by a weighted sum:

$$L(\text{seg}_i) = \lambda_g L_g(\text{seg}_i) + \lambda_l L_l(\text{seg}_i) + \lambda_p L_p(\text{seg}_i) \tag{15}$$

where $\{\lambda_g, \lambda_l, \lambda_p\} \geq 0$ are corresponding weights to the losses.

In training, we minimize the combined loss in Eq. (15) by optimizing the parameters of GRU cells of the RNN networks, message building transformations in Eqs. 4 and 5, and the output MLPs.

Model regularization When training autoencoder style models for anomaly detection, a major challenge is that even if the model learns to generate normal data perfectly, there is still no guarantee that the model will produce high errors for abnormal sequences [20]. In training MPED-RNN, we address this challenge by empirically searching for the smallest latent space that still adequately covers the normal patterns so that outliers fall outside the manifold represented by this subspace.

We implement this intuition by splitting the normal trajectories into training and validation subsets, and use them to regularize the network's hyperparameters that govern the capacity of the model (*e.g.* number of hidden units). More specifically, we train a high capacity network and record the lowest loss on the validation set. The validation set is also used for early stopping. Then, we train a network with lower capacity and record the lowest loss on the validation set again. We repeat this procedure until we find the network with the smallest capacity that is still within 5% of the initial validation loss attained by the high capacity network.

3.4. Detecting Video Anomalies

To estimate the anomaly score of each frame in a video, we follow a four-step algorithm:

1. *Extract segments*: With each trajectory, we select the overlapping skeleton segments by using a sliding window of fixed size T and stride s on the trajectory, similar to Eqs. (12) and (13).

2. *Estimate segment loss*es: We decompose the segment using Eq. (1) then feed all segment features to the trained MPED-RNN which outputs the normality loss as in Eq. (15).

3. *Gather skeleton anomaly score:* To measure the conformity of a sequence to the model given both the past and future context, we propose a voting scheme to gather the losses of related segments into an anomaly

score for each skeleton instance:

$$\alpha_{f_t} = \frac{\sum_{u \in S_t} L_p(u)}{|S_t|} \tag{16}$$

where S_t denotes the set of decoded segments that contain f_t from both reconstruction and prediction. For each of those segments u, the corresponding perceptual loss, $L_p(u)$, is calculated by Eq. (14).

4. *Calculate frame anomaly score:* The anomaly score of a video frame v_t is calculated from the score of all skeleton instances appearing in that frame by a max pooling operator:

$$\alpha_{v_t} = max\left(\alpha_{f_t}\right)_{f_t \in \text{Skel}(v_t)} \tag{17}$$

where $\text{Skel}(v_t)$ stands for the set of the skeleton instances appearing in the frame. The choice of max pooling over other aggregation functions is to suppress the influence of normal trajectories present in the scene, since the number of normal trajectories can vary largely in real surveillance videos. We then use α_{v_t} as the frame-level anomaly score of v_t and use it to calculate all accuracy measurements.

3.5. Implementation Details

To detect skeletons in the videos, we utilized Alpha Pose [10] to independently detect skeletons in each video frame. To track the skeletons across a video, we combined sparse optical flow with the detected skeletons to assign similarity scores between pairs of skeletons in neighboring frames, and solved the assignment problem using the Hungarian algorithm [17]. The global and local components of the skeleton trajectories are standardized by subtracting the median of each feature, and scaling each feature relative to the 10%-90% quantile range. All recurrent encoder-decoder networks have similar architectures but are trained with independent weights. The regularization of MPED-RNN's hyperparameters is done for each data set, following the method described in Section 3.3.

4. Experiments

We evaluate our method on two datasets for video anomaly detection: ShanghaiTech Campus [22] and CUHK Avenue [21]. Each of these datasets has specific characteristics in terms of data source, video quality and types of anomaly. Therefore, we setup customized experiments for each of them.

4.1. ShanghaiTech Campus Dataset

The ShanghaiTech Campus dataset [22] is considered one of the most comprehensive and realistic datasets for video anomaly detection currently available. It combines footage of 13 different cameras around the ShanghaiTech University campus with a wide spectrum of anomaly types. Because of the sophistication of the anomaly semantics, current methods struggle to get adequate performance on it.

Table 1. Frame-level ROC AUC performance of MPED-RNN and other state-of-the-art methods on the ShanghaiTech dataset and its human-related subset. We use the reported results of the referenced methods on ShanghaiTech and carry out their identical experiments on HR-ShanghaiTech whenever possible.

	HR-ShanghaiTech	ShanghaiTech
Conv-AE [14]	0.698	0.704
TSC sRNN [22]	N/A	0.680
Liu *et al*. [20]	0.727	0.728
MPED-RNN	**0.754**	**0.734**

Most of the anomaly events in the ShanghaiTech dataset are related to humans, which are the target of our method. We left out 6/107 test videos whose abnormal events were not related to humans and kept the other 101 videos as a subset called Human-related (HR) ShanghaiTech. Most of the experiments discussed in this section are conducted on the HR-ShanghaiTech dataset.

4.1.1 Comparison with Appearance-based Methods

We train MPED-RNN on all training videos, which is the practice adopted in previous works. Table 1 compares the frame-level ROC AUC of MPED-RNN against three state-of-the-art methods. We observe that on HR-ShanghaiTech, MPED-RNN outperforms all the compared methods. For completeness, we also evaluate MPED-RNN on the original dataset where non-human related anomalies are present and MPED-RNN still attains the highest frame-level ROC AUC.

To understand how the detection of anomalies is made by all models, we visually compare in Figure 4 the map of anomaly scores produced by MPED-RNN to those produced by Conv-AE [14] and Liu *et al*. [20]. As we can observe, our method avoids many irrelevant aspects of the scene since we focus on skeletons. On the other hand, the other two methods try to predict the whole scene and are more susceptible to noise in it.

4.1.2 Interpreting Open-box MPED-RNN

For a deeper understanding on how MPED-RNN works under the hood, we visualize the features generated from the global and local predicting decoders, and the predicted skeleton in image space. For comparison, we also draw the corresponding features of the input sequence. Figure 5 shows two example sequences from the same scene, a

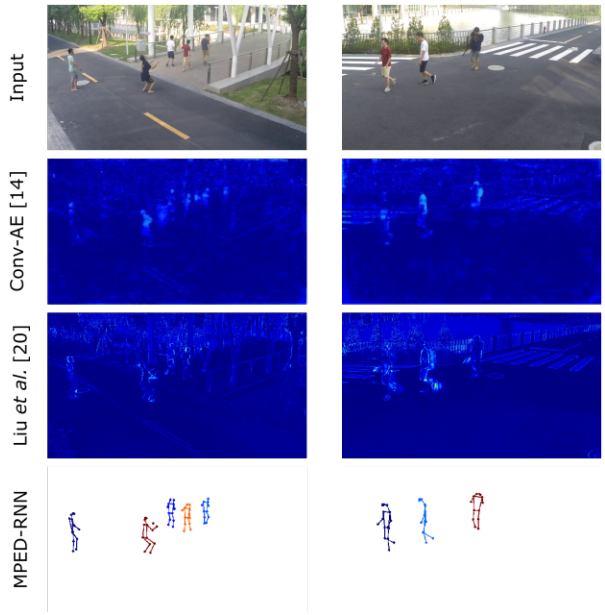

Figure 4. Anomaly score map of Conv-AE [14], Liu *et al*. [20] and MPED-RNN in jet color map. Higher scores are represented closer to red while lower scores are represented closer to blue. The first row shows the original input frames and subsequent rows show the score map of each method. Since MPED-RNN focuses on skeletons it does not produce any score on background pixels.

normal example and an anomalous example. This scene is of the walking area in the campus, where regular activities include people standing and walking casually. In the normal sequence, the predictions follow the input closely in all three domains, which shows that MPED-RNN encodes enough information to predict the normal sequence. On the other hand, the abnormal event contains a person running. Its predicted global bounding box lags behind the input bounding box, indicating that the expected movement is slower than the observed one. The local prediction also struggles to reproduce the running pose and ends up predicting a walking gait remotely mimicking the original poses.

4.1.3 Ablation Study

Table 2 reports the results of simplified variants of MPED-RNN. It confirms that RNN is needed for this problem, and when both global and local sub-processes are modeled, message passing between the sub-processes is necessary. It also shows that the dual decoders are valuable for regularizing the model and detecting anomalies.

4.1.4 Error Mode Analysis

Even though MPED-RNN outperforms related methods, it still makes incorrect decisions. To understand the weaknesses of MPED-RNN, we sorted the test sequences by decreasing level of error made by MPED-RNN and looked for the root causes of the major mistakes.

Table 2. Ablation study about the components of MPED-RNN. We show the frame-level ROC AUC of simpler models that compose MPED-RNN on the HR-ShanghaiTech dataset. AE: Frame-level Autoencoder, ED: Encoder-Decoder, G+L: Global and Local features without message passing. The columns stand for different ways the loss is calculated; Rec: reconstruction only, Pred: prediction ony, Rec+Pred: reconstruction and prediction combined.

	HR-ShanghaiTech		
	Rec.	Pred.	Rec. + Pred.
AE/Image	0.674	N/A	N/A
ED-RNN/Global	0.680	0.688	0.689
ED-RNN/Local	0.700	0.714	0.715
ED-RNN/G+L	0.699	0.722	0.713
MPED-RNN	**0.744**	**0.745**	**0.754**

The most prominent source of errors is from the inaccuracy of the skeleton detection and tracking. All of the skeleton detection methods we tried produced inaccurate skeletons in several common difficult cases such as low resolution of human area or unwanted lighting, contrast or shadow. Moreover, when there is occlusion or multiple people crossing each other, the tracking IDs can get lost or swapped and confuse MPED-RNN. Figure 6.a shows an example frame containing a badly detected skeleton.

Apart from the input noise, a small portion of error comes from a more interesting phenomenon when the abnormal action of subjects produce similar skeletons to normal ones. Figure 6.b shows the case of a person slowly riding a bicycle with motion and posture similar to walking, which tricks our model into a false negative. This issue is a predicted downside of geometrical skeleton features, where all appearance features have been filtered out. Augmenting the skeleton structure with visual features is a future work towards solving this issue.

4.2. CUHK Avenue dataset

We also tested MPED-RNN on the CUHK Avenue dataset, which is another representative dataset for video anomaly detection. It contains 16 training videos and 21 testing videos captured from a single camera. Based on earlier error analysis on the ShanghaiTech dataset, we understand that the unstable skeleton inputs are the most important source of inaccuracy. To avoid this issue, we manually leave out a set of video frames where the main anomalous event is non-human related, or the person involved is non-visible (*e.g*. person out of view throwing an object into the scene), or the main subject cannot be detected and tracked. This selection is detailed in Appendix A. We called the re-

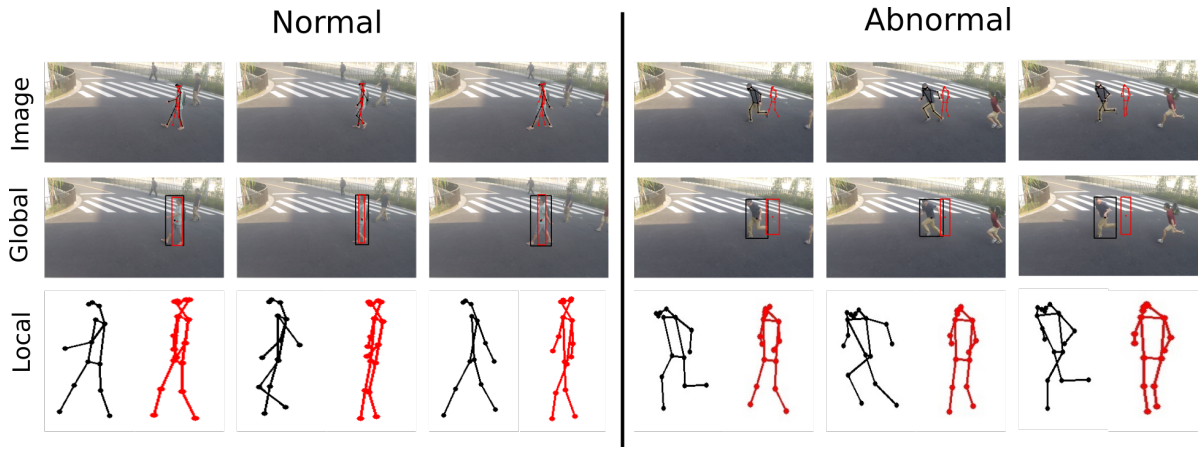

Figure 5. Visualization of the predicted features (red) compared to input features (black) in a sample case for a normal trajectory and an anomalous trajectory.

Figure 6. : Error mode examples. (a) Skeleton detection inaccuracy: bad detection of the person reflected in the glass lead to bad predictions by MPED-RNN. (b) Confusion in feature space: The person riding the bicycle (red) has a moving pattern "similar" to a person walking.

maining dataset HR-Avenue. On HR-Avenue, we achieved a frame-level ROC AUC of 0.863, against 0.862 and 0.848 achieved by Liu *et al*. [20] and Conv-AE [14], respectively.

5. Discussion

With less than a hundred dimensions per frame on average, equal to a small fraction of the popular visual features for anomaly detection (Resnet features of 2048 [14], Alexnet fc7 of 4096 [15]), skeleton features still provide equal or better performance than current state-of-the-art methods. This revives the hope for using semantic guided stage-by-stage approaches for anomaly detection amid the trend of end-to-end image processing deep networks. It also reflects the current evolution of the architectures being modular, with multiple independent modules [2, 18]. Apparently, its performance still depends on the performance of skeleton detection and tracking. This problem becomes more significant in the case of low quality videos. It prevents us from trying our method on UCSD Ped1/Ped2 [23], another popular dataset, whose video quality is too low to detect skeletons. However, the reliability of these skeleton detection modules are constantly increasing [4, 10]. Furthermore, in many cases where skeletons are unavailable, appearance based features can provide complementary information to help. This opens a promising direction of combining these features in a cascaded model, where they can cover weaknesses of each other. Our message-passing scheme can naturally be extended to incorporate sub-processes with non-skeleton features.

Although dynamic movement and posture of single person can reflect the anomalies in most cases, they do not contain information about the interactions between multiple people in the events, and between human and other objects. The global-local decomposition used in our method can be extended to objects by exploring the part-based configuration for each type of them. Toward multi-person/object anomalies, the message passing framework in MPED-RNN is ready to extend support to them, by expanding to inter-entity messages.

6. Conclusions

Through our experiments, we learned that skeleton motion sequences are effective to identify human-related video anomalous events. We observed that the decomposition of the skeleton sequence into global movement and local deformation – combined with our novel message-passing encoder-decoder RNN architecture – appropriately separates anomalous sequences from normal sequences. MPED-RNN is simple, achieves competitive performance and is highly interpretable. Future work includes examining the regularity of inter-human interactions, combining skeleton features with appearance counterparts, and expanding the component based model to non-human objects.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016. 2.2

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *IEEE Conference on Computer Vision*

and Pattern Recognition, pages 39–48, 2016. 5

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3319–3327, 2017. 1

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[5] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN Encoder–Decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014. 3.2

[6] Y. S. Chong and Y. H. Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017. 1, 2.1

[7] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond. Toward abnormal trajectory and event detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):683–695, 2017. 1

[8] Y. Du, Y. Fu, and L. Wang. Representation learning of temporal dynamics for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 25(7):3010–3022, 2016. 2.2

[9] A. Elaoud, W. Barhoumi, H. Drira, and E. Zagrouba. Analysis of skeletal shape trajectories for person re-identification. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 138–149. Springer, 2017. 2.2

[10] H. Fang, S. Xie, Y. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *IEEE International Conference on Computer Vision*, pages 2353–2362, 2017. 3.5, 5

[11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 2.2, 3.1

[12] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017. 3.2

[13] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2.2

[14] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016. 1, 2.1, 1, 4.1.1, 4, 4.2, 5

[15] R. Hinami, T. Mei, and S. Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *IEEE International Conference on Computer Vision*, pages 3639–3647, 2017. 1, 2.1, 5

[16] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, 2009. 2.1

[17] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3.5

[18] B. Liu, S. Yeung, E. Chou, D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *European Conference on Computer Vision*, pages 569–586. Springer, 2018. 5

[19] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2018. 2.2

[20] W. Liu, D. L. W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2.1, 3.3, 1, 4.1.1, 4, 4.2

[21] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 FPS in MATLAB. In *IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 4

[22] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *IEEE International Conference on Computer Vision*, pages 341–349, 2017. 1, 2.1, 4, 4.1, 1

[23] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. 5

[24] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning*, pages 843–852, 2015. 3.2

[25] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *International Conference on Machine Learning*, pages 3560–3569, 2017. 2.2, 3.1

[26] H. Vu, T. D. Nguyen, A. Travers, S. Venkatesh, and D. Phung. Energy-based localized anomaly detection in video surveillance. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 641–653. Springer, 2017. 1

[27] D. Xu, Y. Yan, E. Ricci, and N. Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 1, 2.1

[28] Y. Yuan, J. Fang, and Q. Wang. Online anomaly detection in crowd scenes via structure analysis. *IEEE Transactions on Cybernetics*, 45(3):548–561, 2015. 2.1

[29] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1933–1941. ACM, 2017. 2.1

[30] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012. 1

A. HR-ShanghaiTech and HR-Avenue Datasets

The ShanghaiTech dataset contains anomalies that are not related to humans in 6 out of its 107 test videos. These 6 videos are:

- Camera 01: Videos 0130, 0135 and 0136;
- Camera 06: Videos 0144 and 0145;
- Camera 12: Video 0152.

HR-ShanghaiTech was assembled by removing those videos from the ShanghaiTech dataset.

As for the HR-Avenue dataset, since the original Avenue dataset contains only 21 testing videos, we ignored segments of the videos where the anomalies were not detectable by the pose detector we employed or where the anomaly was not related to a human. The ignored segments were:

- Video 01: Frames 75 to 120, 390 to 436, 864 to 910 and 931 to 1000;
- Video 02: Frames 272 to 319 and 723 to 763;
- Video 03: Frames 293 to 340;
- Video 06: Frames 561 to 624 and 814 to 1006;
- Video 16: Frames 728 to 739.