Published in final edited form as:

Proc ACM Int Conf Multimed. 2017 October ; 2017: 1059–1067. doi:10.1145/3123266.3123365.

Region-based Activity Recognition Using Conditional GAN

Xinyu Li¹, Yanyi Zhang¹, Jianyu Zhang¹, Yueyang Chen¹, Huangcan Li¹, Ivan Marsic¹, and Randall S. Burd²

¹Rutgers University, Piscataway, NJ, 08854, USA.

²Children's National Medical Center, Washington, D.C., 20010, USA

Abstract

We present a method for activity recognition that first estimates the activity performer's location and uses it with input data for activity recognition. Existing approaches directly take video frames or entire video for feature extraction and recognition, and treat the classifier as a black box. Our method first locates the activities in each input video frame by generating an activity mask using a conditional generative adversarial network (cGAN). The generated mask is appended to color channels of input images and fed into a VGG-LSTM network for activity recognition. To test our system, we produced two datasets with manually created masks, one containing Olympic sports activities and the other containing trauma resuscitation activities. Our system makes activity prediction for each video frame and achieves performance comparable to the state-of-the-art systems while simultaneously outlining the location of the activity. We show how the generated masks facilitate the learning of features that are representative of the activity rather than accidental surrounding information.

Keywords

Activity Recognition; Localization; Generative Adversarial Network; Deep Learning

1 INTRODUCTION

Driven by successes in image and speech recognition, deep learning has also been applied to activity recognition [7]. Most deep learning approaches to activity recognition first extract spatial features with convolutional neural networks (ConvNet), model the temporal associations of spatial features with long-short-term-memory networks (LSTMs), and then treat the activity recognition as a multiclass classification problem. Although deep learning has outperformed representation-based approaches [7], problems still remain. First, *a correct activity prediction does not guarantee that the model learned the representation of the activity* because a deep learning model may extract the features associated with the activity as well as unrelated features. As a result, the system may rely on the background or unrelated features to make the prediction. For example, a system may recognize a gym vault activity by observing a pommel horse instead of the actual activity. This problem is especially evident for small datasets due to the limited variability in the training and testing sets. Second, *making correct activity prediction do not necessary mean that the system is able to locate the activity*. The network may rely on features other than activity performer to make predictions and, the activation map in each convolutional layer may not properly

reveal the location of activity performer. The approach based on region proposal requires a pre-trained network for object recognition and then delimits the regions in input image where the associated activity might occur [4, 18]. Our approach is more general, using activity-associated features to generate the activity location later used for activity recognition.

We introduce an activity recognition system that recognizes activities in two steps (Figure 1). First, it localizes the activity by generating an *activity mask* outlining the location where the activity is expected to occur. For mask generation, we used a conditional generative adversarial network (cGAN) [22]. Given that activities are continuous and usually represented as video clips, we introduced a Conv-LSTM-Deconv structure as a continuous mask generator with cGAN for training. The cGAN-based image generator has shown better performance than a Conv-Deconv-based generator [8]. The ConvNet-LSTM structure has been used to model spatio-temporal associations [9, 24]. In addition to the adversarial loss, we introduced a spatio-temporal loss and implemented perceptual loss [10]. These losses penalize the neural network for pixel-wise errors in the generated mask and discontinuities between masks from consecutive frames. Second, the generated mask is appended to each color channel of its input video frame to delimit the activity region for the activity recognizer. Because the ConvNet-LSTM structure has been used successfully for modeling spatio-temporal activity associations [16, 35], we adopted a VGG-LSTM network for activity recognition.

To train and test our system, we manually created activity masks for two datasets. We selected six activities from a well-known Olympic sports dataset (15 videos per activity) and for 10 frames of each video manually created binary masks outlining the activity performer's location. We also manually created multilevel masks highlighting trauma team members and their roles for depth videos of trauma resuscitations.

Our system achieved state-of-the-art performance compared with previous activity recognition approaches [2, 9, 13, 23]. In addition, it generated activity location masks. Our system works online, as opposed to offline prediction based on features extracted from the entire video [19, 36]. Unlike the approaches that rely on crafted features or purely rely on learned features, our system automatically generates activity masks that promote learning the most representative features. This property makes our system generalizable and scalable. Our visualization of the learned feature maps demonstrates that the generated masks enable the system to learn the features that are associated with activities, unlike a VGG net trained using only video frames. Our contributions are:

- (1) **A novel approach for activity recognition** that first generates activity masks to provide additional information for subsequent activity recognition. The system uses the generated masks as an additional color channel of the original input image to perform per-frame activity recognition and find the activity's bounding box.
- (2) **The Conv-LSTM-Deconv generator for cGAN** used to generate an output sequence of video masks based on the spatio-temporal associations in a sequence of input video frames. Our **spatio-temporal loss function** can tune the

mask generator based on pixel-wise errors of the generated masks and discontinuities in the mask sequence.

- (3) **A cascade-structure mask generator** that generates the activity mask before the activity is recognized using a VGG-LSTM network. We introduced **activity recognition loss** to tune the weights of the conditional GAN generator based on the loss from activity recognition.
- (4) **Two labeled datasets** with manually generated masks for an Olympic sports dataset [23] and a trauma resuscitation dataset. These datasets could be used in future research on segmentation, classification, or activity recognition.

In the rest of the paper, Section 2 reviews the related work on activity recognition and image-to-image mapping. Section 3 describes our approach and its implementation. Section 4 presents the experimental results on different datasets. Section 5 shows our model's ability to learn activity-associated features, and discusses limitations and possible extensions. Section 6 concludes the paper.

2 RELATED WORK

Activity recognition approaches can be roughly divided into two classes: representation based and deep learning based [7]. Representation-based approaches rely on crafted features and descriptors that face generalization issues. The body skeleton and joints provided by the Kinect sensor [31] are often used for activity recognition as a representation of body posture [24]. Due to the variety of application environments, these features and skeletons may fail to generalize and lead to recognition errors.

Deep learning has been recently applied to activity recognition, initially using single images independently [38], that ignores the temporal associations of activities. Subsequent ConvNet research used approaches that learned the temporal associations by feeding in video frames stacked over a time window [12]. This approach could only model short-range temporal associations of activities in short video clips. Fusion-over-time strategies partially addressed this issue by stacking the features extracted from video frames and using slow fusion [12]. Recurrent neural networks and long-short-term memory networks (LSTMs) also were used for modeling long-range temporal associations. The ConvNet-LSTM structure was used for activity recognition with different types of input (RGB video, mobile sensor data) [16, 30, 35].

Most existing research treated the activity recognition as a classification problem performed within a black-box. It is difficult to tell whether this model learned the actual activity related features or over-fitted to the background features. Previous research has focused on visualizing and understanding the trained ConvNet [37] and LSTM [11], and the learned feature maps demonstrated that the neural network learned the features of both activity and its surroundings. Generating the correct label on the testing set with limited testing data does not prove the model learned the activity; therefore, this system may not be generalizable. This problem is particularly noticeable for small-size datasets.

Our approach to activity recognition draws from several existing ideas. The region proposal method [20, 38] requires a pre-trained object recognition model for region generating (usually trained on ImageNet [28]). The regions containing the target (and possibly unrelated people or objects) are generated, and a secondary network takes the generated regions and their relative location for textual image description or activity recognition. Unlike this approach, we only generated the activity region for training the activity recognition model. Other research has relied on the use of trajectory information to distinguish people from background, and multiple descriptors for action detection [19, 36], or has used an LSTM autoencoder to help the supervised learning [34]. Because not all the people in the scene may be involved in the activity, we first generate masks that only outline the activity location or the performer's role in a team instead of simply segmenting the people from the background.

Our approach first mines additional information (activity location or team members' roles) and then does activity recognition based on both inputs and this additional information. We used a per-pixel mapping strategy for activity localization or team role labeling, similar to semantic segmentation that distinguishes foreground and background, or image-to-image mapping [8]. Image segmentation previously used fully-convolutional networks with convolution and deconvolution [20]. More recent state-of-the-art research used GAN-based generative structures [5, 22] for the image-to-image mapping [8] and segmentation [21]. Our work is novel in that we trained the system to find the activity location information from input data for activity recognition.

3 METHODOLOGY

Our system has two network structures (Figure 2), one for activity-mask generation that includes a generator and discriminator and one for activity recognition. For selected video frames in our datasets, we manually created a binary activity mask outlining the activity performer, or a multilevel mask for team members with different roles. We used these masks as the ground truth for system training.

3.1 Model Structure

We used a cGAN to generate the activity masks and a VGG-LSTM structure for activity recognition (Figure 2). Generating the mask is comparable to other space-transformation or mapping tasks, such as text-to-image mapping [26], image-to-image mapping [8], or image super resolution (low to high-resolution mapping) [14]. The current state-of-the-art approaches to these problems are GAN-based. We developed our model based on cGAN to generate the continuous activity masks.

Unlike single image segmentation which relies only on the spatial associations between inputs and outputs, generating activity masks for continuous video frames relies on both spatial and temporal associations. The generated mask should spatially match the ground truth mask. The generated mask should temporally be continuous over time (i.e., smoothly transitioning). We introduced a Conv-LSTM-Deconv structure as the generator of the conditional GAN model (Figure 2, generator). The Conv-Deconv network with fully convolutional structure was used [20] to avoid the use of fully-connected layers, making the

model more memory-efficient and better at preserving spatial features. The LSTM [3] has been widely used to model temporal associations and recently for acoustic modeling [29]. We inserted the LSTM layers between the convolutional and deconvolutional networks to model the temporal associations of spatial features learned by the ConvNet. This approach increased the average precision of activity recognition by 6% compared with the Conv-Deconv generator alone (Figure 2).

The adversarial loss from the discriminator (L_A in Figure 2) can only tune the weights in the mask generator based on the feature-wise difference between the generated and ground-truth masks. To better tune the generator, we built a cascade of the mask generator followed by activity recognizer (Figure 2) which produces activity recognition error to penalize the generator (L_C in Figure 2). This structure takes the generated mask and original image, appends the mask after each color channel, and uses it for activity recognition. Because recognizing some activities requires background features, we decided to *append* the generated mask as an additional “color channels” of the input image after each regular RGB color channel, instead of directly applying the mask on the input image. This approach allows the system to learn some features not directly associated with the activity performer. Our activity recognition module favors the area covered by the activity mask but also allows inclusion of the whole background. The activity recognition loss is backpropagated for tuning the generator (L_C in Figure 2).

As previously shown [30], deep Conv-Deconv structures are hard to converge. To improve the convergence, we adopted the “U-Net” [27] to add extra connections between the associated convolutional and de-convolutional layers.

3.2 Loss Functions

The generative adversarial network (GAN) [5] learns the mapping of input data z to output mask y , denoted as $G: z \rightarrow y$ [5]. The regular GAN model can be extended to a conditional model that uses additional information x . The conditional GAN [22] establishes the mapping function based on both the input data z and conditional information x , denoted as $G: \{x, z\} \rightarrow y$. The additional information x can be any auxiliary information. We used the manually generated masks which contained ones or categorical labels at pixels related to the activity, and zeros for the unrelated pixels as the condition x .

Our system was designed for activity recognition, so that loss could be measured spatially and temporally. In addition to the regular adversarial loss L_A , we used three other losses:

$$L^* = L_A(G, D) + \alpha L_{sr}(G) + L_p(G) + \beta L_C(G, C) \quad (1)$$

where α, β are the manually defined weights. We empirically determined that $\alpha = 0.8$ and $\beta = 0.2$ showed the best performance.

The first term $L_A(G, D)$ is the regular adversarial loss from the mask discriminator (Figure 2). The discriminator uses the ConvNet structure to measure the feature-wise similarity

between the generated and activity masks, and normalizes the similarity to the range of zero to one. The adversarial loss of cGAN model is:

$$L_A(G, D) = E_{x, y \sim p_{data}(x, y)} [\log D(x, y)] + E_{x \sim p_{data}(x), z \sim p_z(z)} (1 - \log(1 - D(x, G(x, z)))) \quad (2)$$

where G denotes the mask generator and D denotes the discriminator. The training adjusts the parameters of G to minimize the objective while the discriminator D maximizes the objective.

The second term $L_{st} = L_d + L_t$ is the proposed spatio-temporal loss calculated by the generator itself. L_d denotes the pixel-wise loss between the generated activity mask and the ground truth, which is commonly used in GAN image generation tasks as:

$$L_{\ell_1} = \frac{1}{CWH} \sum_{c=1}^C \sum_{w=1}^W \sum_{h=1}^H \left\| [\phi_G(x)]_{c, w, h} - y_{c, w, h} \right\|_1 \quad (3)$$

where image dimensions are the number of color channels C , width W and height H . In our case of activity masks, $C = 1$ because of single mask per frame, whether the mask is binary or multilevel. The $\phi_G(\cdot)$ denotes the generated mask, and y denotes the ground-truth mask. $\|\cdot\|_1$ denotes the pixel-wise L1 norm which favors similarity between the generated activity mask and ground truth for every frame. The L_t is the temporal loss. To make the system generate masks for continuous activities, the L1 distance of the consecutive masks should have minimal variation. We thus introduced L_t to model the temporal loss between frames as:

$$L_t = std([L_{\ell_1}]) \quad (4)$$

where the $std(\cdot)$ denotes the standard deviation of the list $[L_d]$ of the pixel-wise L1 distances across frames.

The third term, $L_p(G)$, denotes the perceptual loss for the generator [10]. We define $\hat{y} = \phi_G(x)$ as the output of the generator. We followed the definition:

$$L_{feat}^i(\hat{y}, y) = \frac{1}{C_i W_i H_i} \left\| \phi_{vgg}^i(\hat{y}) - \phi_{vgg}^i(y) \right\|_2^2 \quad (5)$$

$$L_{style}^i(\hat{y}, y) = \left\| G_{vgg}^i(\hat{y}) - G_{vgg}^i(y) \right\|_F^2 \quad (6)$$

where $C_i \times W_i \times H_i$ defines the shape of the feature map, L_{feat}^i denotes the feature loss and L_{style}^i denotes the style loss at corresponding i^{th} convolution layer of the VGG net [32]. The $\|\cdot\|_2^2$ denotes the squared L2 norm, and $\|\cdot\|_F^2$ denotes the squared Frobenius norm. The $\phi_{vgg}^i(\cdot)$ term denotes the channel-wise Gram matrix at i^{th} convolution layer of VGG net. The element in position $1 \leq m, n \leq C_i$ of Gram matrix for any feature map f is defined as:

$$[G_{vgg}^i(f)]_{m,n} = \frac{1}{C_i W_i H_i} \sum_{w=1}^{W_i} \sum_{h=1}^{H_i} (f_{m,w,h} \times f_{n,w,h}) \quad (7)$$

In addition, we applied a total variant regularization $L_{TL}(\hat{y})$ to generator output $\hat{y}$ for smoothing [10]. The perceptual loss consists of feature-reconstruction loss, style loss and total variation regularization, which have been successfully used for GAN-based image generation [10, 14]. We used this loss for generator tuning.

The last term $L_C(G, C)$ is our proposed classification loss. We designed the generator to take the additional loss from the error of the activity recognition network C . The activity recognition loss is defined as the categorical cross-entropy from the activity recognizer. The cross-entropy is calculated using the stacked generated masks and input images as input and activity ground truth as output. This classification loss can be expressed as:

$$L_C = - \sum_N y'_n \log y_n \quad (8)$$

where N is the total number of classes, y'_n is the predicted probability distribution of class n , and y_n is the ground truth for class n . For each iteration, we first trained the activity recognizer using the stacked ground truth mask and the input image as input, and used the activity label to tune the weights of the recognizer. To generate the activity recognition loss, we used the generated mask instead of the ground truth for activity recognition and calculated the categorical cross-entropy of activities.

To study the impact of different losses, we ran the comparison experiments with versions of our system trained on different losses. The results show that the proposed spatio-temporal loss has a positive contribution to activity recognition and segmentation (Table 1). We used the per-pixel accuracy [1], as well as per-frame activity recognition accuracy for this comparison.

3.3 Activity Recognition and Localization

Once the system is trained, we append the activity recognizer after the generator and run feed-forward for activity recognition. It is possible that a dataset contains a large number of video clips, but only a few of them have manually-generated masks for training. A system trained with a limited amount of generator-training data is more likely to make mistakes when generating masks. To avoid the mask generation errors propagating to the activity

recognition results and help the system tolerate these errors, we chose to append the generated mask to the original image after each channel instead of directly applying the mask to the original image by performing their multiplication. In this way, information from the original input image is preserved even when the generated mask contains errors.

The activity recognition network was not well trained during mask generator training due to the limited number of frames with ground truth masks. We then used the data with only activity class labels and no activity masks to further fine-tune the activity recognition network (VGG). The system first generated the mask for each input frame and then appends the generated mask to this input frame. This stacked image is then fed into the activity recognizer, which is trained using ground truth activity labels. We set the weights of the mask generator as untrainable during this fine tuning so that we can only tune the weights of the activity recognizer.

Activity localization is an end product of our method that can be simply accomplished by calculating the bounding box of the fully-connected regions in the generated masks and applying the bounding box to the input image. During activity prediction, we performed an image morphology close operation to the generated masks to eliminate the noise before calculating the bounding box.

3.4 Model Implementation

We implemented our system with Keras using the Theano backend. We used the fully convolutional structure [20] with He initialization [6]. *Leaky ReLU* and *tanh* were used as activation functions. The batch normalization and dropout strategy [33] was used to avoid overfitting. The generator was tuned three times in each training iteration: first using the adversarial loss from the discriminator, then using the spatio-temporal and perceptual losses [10], and finally using the classification loss from the activity recognizer.

We trained the network on the single GTX1080 GPU with 8GB memory. The loss functions were manually defined using Keras backend programming. Training with 128×128 frames took two days to converge. Because we used *Leaky ReLU* and *tanh* as activation functions, we first normalized the input data in $[-1, 1]$.

4 EXPERIMENTAL RESULTS

4.1 Dataset

Olympic Sports Dataset with Masks: We selected six activities from a Olympic sports dataset for labeling (Table 2). The dataset contains videos recorded with moving backgrounds and moving athletes. Because the original videos had different frame rates, and most of the frames within each second were visually similar before the activity started, we down sampled the original color video to three frames per second and selected only the most representative frames (covering different stages of activity performance) for manual labeling. The video frame selection can be automatically accomplished based on the pixel difference between adjacent frames by selecting the frames most different than previous frames. The frame difference can be estimated by calculating the optical flow. We manually

selected masks for 15 videos for each of the six activities, resulting in a total of 900 frames. We kept generating more masks for all 16 activities in the Olympic sports dataset.

Trauma Resuscitation Dataset with Masks: The original trauma resuscitation dataset contained depth videos captured during actual events. We sampled each original depth video to one frame per second. Because we used a fixed camera, much less background movement was observed than in videos of sports activities. Instead of using the difference of adjacent frames as an indicator, we selected the frames based on the resuscitation process phase. The resuscitation phases are the major steps of the process, with each trauma resuscitation contains six phases: pre-arrival, patient-arrival, primary, secondary, post-secondary and patient-leave [17]. We randomly selected five consecutive frames within each phase in 20 videos and labeled four roles of the trauma team (Table 2) and the patient bed (where the resuscitation activities occur) with categorical labels (Table 2). Six hundred images were manually labeled for the trauma resuscitation dataset.

4.2 Activity Recognition in Sports Videos

We evaluated three aspects: the overall performance, the performance breakdown for each activity and the system generalizability.

We first evaluated the overall activity recognition performance using accuracy and the commonly used F-measure with 30% videos of each activity (Table 3). We compared our system with baseline activity recognition models (ConvNet, CNN-LSTM) using the same dataset (Table 3). We also compared the system performance with a different configuration for activity recognition (Table 3) to analyze the contribution of different losses to our proposed system. The results shown that: ① Our baseline VGG net and VGG-LSTM structure achieved performance similar to representation-based approaches [13, 23] (Table 3, left side). While the regular deep-learning models have the ability to learn representative features, this finding showed that, they are also limited by their simple structure when generalizing to complex tasks such as recognizing activities where unrelated people are present. The VGG-LSTM achieved an accuracy increase of about 5% compared with ConvNet only, due to its ability to model the temporal associations of extracted features. ② We also observed that different network models generated masks of different quality (Table 1). The cGAN with our proposed losses generated the most accurate masks for most video frames (Figure 3). We also quantitatively evaluated the generated masks. Because we do not have the ground truth masks for all videos, we adopted the 80% - 20% ratio to split the training and testing sets for the frames that had ground truth masks (Table 1). ③ The quality of generated masks affected the activity recognition results (Table 1, 3). Our proposed spatio-temporal loss increased the activity recognition accuracy by $\approx 3\%$ compared to a regular cGAN trained only using the adversarial loss (Table 3). Including the classification loss into our structure increased the activity recognition accuracy by $\approx 2\%$ (Table 3). ④ The state-of-the-art activity recognition strategies [19, 36] used representation based on multiple features and moving trajectories to describe the entire video. These methods require computationally expensive calculations for feature representation from the entire video. Our proposed approach can make per-frame activity predictions at 20fps by running feedforward networks, and simultaneously provide the

activity location information. We believe runtime operation offers significant advantage though with slightly lower performance.

Second, we evaluated the recognition performance for each of six sports activities and compared that with previous research on the same dataset (Table 4). Our structure outperformed existing methods [12, 13, 20, 23, 29], particularly for the activities during which unrelated people were present (shaded in Table 4), such as judges and observers. For these activities, activity location was critical for recognition. The gym vault shares a similar starting posture (running) with other activities, and the running step lasts longer than other steps. This fact makes the gym vault appear similar to some other activities, causing low predictive performance.

We finally tested generalization in our system. We used the generator trained for the six sports activities to generate masks for 16 activities that included the original six. We then tuned the activity recognizer to recognize all 16 activities. For all 16 activities, our system achieved 7% lower performance than for the six it trained on (Table 3), which was caused by errors in the generated masks. Our system was still able to achieve performance comparable to existing systems [2, 9, 13, 23], showing that it generalizes and to some extent tolerates errors in the generated masks.

4.3 Activity Recognition in Trauma Resuscitation Depth Videos

The trauma resuscitation setting poses a greater activity recognition challenge compared to sports videos. In sports, videos are recorded by RGB cameras and usually contain only a single performer (i.e., no cooperation is required). In contrast, the trauma resuscitation data, includes the teamwork captured by a depth camera. Medical settings are highly privacy-sensitive, precluding data collection by wearable sensors or RGB cameras. The resuscitation process phase consists of multiple tasks, which require teamwork by several roles in the team. We introduced different labels for different members of the trauma team based on their role instead of a binary labeling for all performers (Figure 4).

We compared the process phase detection accuracy and F-score with the previous research using the same dataset and other baseline models (models using VGG-only and VGG-LSTM for activity recognition). Our system achieved an average of 80% accuracy for team role labeling. Its mean average precision of 81% for phase detection outperformed the system without masks (Table 5). We also found that the system performed better with team-role multilevel masks compared with binary masks (i.e., providers and patient bed are labeled with ones and everything else with zeros). This finding likely occurred because the rough location of the team roles varies during resuscitation phases, and role mask provides the location information for different roles. The improvement by role mask was not significant because the cGAN will also likely create more role-mask errors (incorrect values) than in a binary mask.

The comparison (Table 5) shows that the system using only ConvNet and Conv-LSTM achieved the lowest resuscitation phase detection performance. This finding is likely because of the lack of detailed texture information in our low-resolution depth images, as well as its difficulties with handling multiple people performing complex concurrent activities.

Previous approaches have introduced additional sensors to enrich the phase information. For example, the sound captured by the microphone array proved useful for resuscitation phase detection [15, 17]. Our research directly generated the team role masks and not only outperformed previous work on the same dataset but also provided team-role masks that can be used in other applications.

5 DISCUSSION

5.1 Efficiency of Feature Learning

Our system uses masks to facilitate learning features associated with the activity, instead of relying on the background. To further illustrate this, we extracted the feature maps learned in the seventh convolutional layer ($8 \times 8 \times 256$) of our activity recognizer (Figure 2, the layer pointed to by the arrow). We summed up the activation maps along the channel dimension, resized them, and overlaid the generated activation maps on the input image.

The activation map visualization (Figure 5) shows that both networks (the network with or without using the mask as additional input channel) learn features useful for activity recognition. The overlaid heat maps, however, show that most of the features learned by our system using inputs with masks are directly associated with the activity performer (Figure 5, top row), while the mask-less system learned features are associated with background objects (e.g. pommel (led) horse for gym vault, or score board for layup) (Figure 5, bottom row). Although such features often coexist with certain activities and may indicate their occurrences, the system may suffer from generalization issues because these features may not exist in other video clips of the same activity. Given videos recorded in different environments, the system may fail to recognize the activity because of its inability to extract features directly associated with the activity performer. The experimental results confirmed that our system using the generated mask can generate more representative features that are associated with the activity itself, as opposed to background objects or unrelated people.

5.2 Error Analysis

The quality of generated masks affects the activity recognition performance. We analyzed the errors in generated masks and found that they could be roughly categorized into four types (Figure 6, top): ① The generated mask covered the activity and unrelated background targets, because the activity performer was visually or feature-wise similar to background objects. ② The generated mask partly covered the activity performer, usually caused by insufficient training data, or by a blurred or too-small activity performer in the input frame. ③ The mask covered unrelated people instead of the activity performer, because the performer appeared too small or occluded by unrelated people. ④ The system failed to generate the mask and did not capture any target, because the lack of training data caused the system to fail to identify any activity in the frame. The mask generator could be enhanced using more training data, a deeper network, or other types of GAN structure [5].

In addition to the generated mask errors, the activity recognition errors were usually caused by visual similarity between different activities (Figure 6, bottom). For example, the posture

of the starting stage of basketball layup is similar to the starting stage of gym vault, causing the system to misclassify the two activities. The misclassification also happened between a layup and bowling, long jump, high jump and gym vault. Even human observers may find it difficult to distinguish activities initially when the performer's posture may be similar. As the activity progresses and its appearance becomes more distinct, the system can make corrections to the previous prediction. Using the bidirectional-LSTM instead of a regular LSTM may help our system to predict the activity at a certain time using the input frame before and after this time. The tradeoff is that the system would not be able to make activity predictions in real-time because future frames are required for bidirectional-LSTM.

5.3 Limitations and Extensions

Although our system achieved the performance comparable to the state-of-the-art systems, it has several limitations. First, training the activity mask generator requires labor-intensive manual labeling of ground-truth masks. We created two labeled activity datasets and will continue adding more, but constructing a large-scale dataset is tedious and time-consuming. Our next step is to train a generative model able to generate masks for foreground and background using a modified published image-segmentation dataset. This type of generator could be used to generate masks for the activity performer or other targets in video frames. Second, our system relies on the generated activity masks for team-role information or location of the activities, making the activity recognition results partly dependent on the quality of the generated activity masks. The mask generation could be considered as a pixel-to-pixel mapping problem, for which we used the cGAN-based approach commonly applied in similar domains. Generating more accurate masks would improve system performance, which is why our future work includes exploring different approaches for pixel-to-pixel mapping.

We presented two applications of our proposed system: activity recognition and process-phase detection. Our framework, however, is generalizable to a wider range of activity recognition problems. The idea of generating a mask can be compared to region proposal [18, 25], which was used for target recognition and tracking. Instead of generating the activity-location or team-role masks, we could generate masks for any people or objects. One could also generate the masks to distinguish different activity performers and achieve concurrent activity recognition together with their locations. Aside from generalizability, running time is another important factor for real-world implementations. Based on our experiments on $128 \times 128px$ input, our system achieved mask generation and activity recognition speed of 120fps (without considering I/O time) using a simple Conv-Deconv structure, and 20fps using our proposed cGAN structure. The mask generation and activity recognition speeds using simple conv-deconv are comparable to the state-of-the-art region proposal strategy YOLO with 145fps on $224 \times 224px$ input images [25]. Even using the cGAN mask generator, our system achieved twice the speed of R-FCN, which achieved 6fps on $448 \times 448px$ images [18].

6 CONCLUSION

We presented a new approach for activity recognition that first performs pixel-to-pixel mapping from the original image to a generated activity mask, then uses the mask to predict either an activity or a process phase. This work contributes to the community:

1. The Conv-LSTM-Deconv generator for generating continuous masks that delimit the activity location using a conditional GAN and the spatio-temporal loss function.
2. A novel activity recognition and localization strategy, which achieved the state-of-the-art recognition performance while additionally outlining the activity location.
3. The cascade structure and the activity recognition loss used for mask generation and activity recognition.
4. Two manually-labeled activity recognition datasets with evaluations that can be used in future research.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health under Award Number R01LM011834. The authors would like to thank Dr. Vishal M. Patel and He Zhang for suggestions, and the anonymous reviewers for their review and comments.

REFERENCES

- [1]. Csurka Gabriela, Larlus Diane, Perronnin Florent, and Meylan France. 2013 What is a good evaluation measure for semantic segmentation. In *BMVC*, Vol. 27. Citeseer, 2013.
- [2]. Gaidon Adrien, Harchaoui Zaid, and Schmid Cordelia. 2014 Activity representation with motion hierarchies. *International journal of computer vision* 107, 3 (2014), 219–238.
- [3]. Gers Felix A, Schmidhuber Jürgen, and Cummins Fred. 1999. Learning to forget: Continual prediction with LSTM. (1999).
- [4]. Girshick Ross. 2015 Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.
- [5]. Goodfellow Ian, Pouget-Abadie Jean Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron, and Bengio Yoshua. 2014 Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
- [6]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. 2016 Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [7]. Herath Samitha Harandi Mehrtash, and Porikli Fatih. 2017 Going deeper into action recognition: A survey. *Image and Vision Computing* 60 (2017), 4–21.
- [8]. Isola Phillip, Zhu Jun-Yan, Zhou Tinghui, and Efros Alexei A. 2016 Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004* (2016).
- [9]. Jiang Yu-Gang, Dai Qi, Liu Wei, Xue Xiangyang, and Ngo Chong-Wah. 2015 Human action recognition in unconstrained videos by explicit motion modeling. *IEEE Transactions on Image Processing* 24, 11 (2015), 3781–3795.
- [10]. Johnson Justin, Alahi Alexandre, and Fei-Fei Li. 2016 Perceptual losses for realtime style transfer and super-resolution In *European Conference on Computer Vision*. Springer, 694–711.
- [11]. Karpathy Andrej, Johnson Justin, and Fei-Fei Li. 2015 Visualizing and understanding recurrent networks. *arXiv preprint arXiv: 1506.02078* (2015).

- [12]. Karpathy Andrej, Toderici George, Shetty Sanketh, Leung Thomas, Sukthankar Rahul, and Fei-Fei Li. 2014 Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 1725–1732.
- [13]. Laptev Ivan, Marszalek Marcin, Schmid Cordelia, and Rozenfeld Benjamin. 2008 Learning realistic human actions from movies. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 1–8.
- [14]. Ledig Christian, Theis Lucas, Huszár Ferenc, Caballero Jose, Cunningham Andrew, Acosta Alejandro, Aitken Andrew, Tejani Alykhan, Totz Johannes, Wang Zehan, et al. 2016 Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802 (2016).
- [15]. Li Xinyu, Zhang Yanyi, Li Mengzhu, Chen Shuhong, Austin Farneth R, Marsic Ivan, and Burd Randall S. 2016 Online process phase detection using multimodal deep learning. In Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE Annual. IEEE, 1–7.
- [16]. Li Xinyu, Zhang Yanyi, Zhang Jianyu, Chen Shuhong, Marsic Ivan, Farneth Richard A, and Burd Randall S. 2017 Concurrent Activity Recognition with Multimodal CNN-LSTM Structure. arXiv preprint arXiv:1702.01638 (2017).
- [17]. Li Xinyu, Zhang Yanyi, Zhang Jianyu, Chen Yueyang, Chen Shuhong, Gu Yue, Zhou Moliang, Farneth Richard A, Marsic Ivan, and Burd Randall S. 2017. Progress Estimation and Phase Detection for Sequential Processes. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 3 (2017).
- [18]. Li Yi, He Kaiming, Sun Jian, et al. 2016 R-fcn: Object detection via region-based fully convolutional networks. In Advances in Neural Information Processing Systems. 379–387.
- [19]. Li Yingwei, Li Weixin, Mahadevan Vijay, and Vasconcelos Nuno. 2016 Vlad3: Encoding dynamics of deep features for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1951–1960.
- [20]. Long Jonathan, Shelhamer Evan, and Darrell Trevor. 2015 Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3431–3440.
- [21]. Luc Pauline, Couprie Camille, Chintala Soumith, and Verbeek Jakob. 2016 Semantic Segmentation using Adversarial Networks. arXiv preprint arXiv:1611.08408 (2016).
- [22]. Mirza Mehdi and Osindero Simon. 2014. Conditional generative adversarial nets. arXiv preprint arXiv: 1411.1784 (2014).
- [23]. Nieves Juan Carlos Chen Chih-Wei, and Fei-Fei Li. 2010 Modeling temporal structure of decomposable motion segments for activity classification In European conference on computer vision. Springer, 392–405.
- [24]. Piyathilaka Lasitha and Kodagoda Sarath. 2013 Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on. IEEE, 567–572.
- [25]. Redmon Joseph, Divvala Santosh, Girshick Ross, and Farhadi Ali. 2016 You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 779–788.
- [26]. Reed Scott, Akata Zeynep, Yan Xincheng, Logeswaran Lajanugen, Schiele Bernt, and Lee Honglak. 2016 Generative adversarial text to image synthesis. In Proceedings of The 33rd International Conference on Machine Learning, Vol. 3.
- [27]. Ronneberger Olaf, Fischer Philipp, and Brox Thomas. 2015 U-net: Convolutional networks for biomedical image segmentation In International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 234–241.
- [28]. Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, et al. 2015 Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 3 (2015), 211–252.
- [29]. Sak Hasim, Senior Andrew W, and Beaufays Françoise. 2014 Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Interspeech. 338–342.

- [30]. Shahroudy Amir, Ng Tian-Tsong, Gong Yihong, and Wang Gang. 2016 Deep multimodal feature analysis for action recognition in RGB+ D videos. arXiv preprint arXiv:1603.07120 (2016).
- [31]. Shotton Jamie, Sharp Toby, Kipman Alex, Fitzgibbon Andrew, Finocchio Mark, Blake Andrew, Cook Mat, and Moore Richard. 2013 Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013), 116–124.
- [32]. Simonyan Karen and Zisserman Andrew. 2014 Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556 (2014).
- [33]. Srivastava Nitish, Hinton Geoffrey E Krizhevsky Alex, Sutskever Ilya, and Salakhutdinov Ruslan. 2014 Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
- [34]. Srivastava Nitish, Mansimov Elman, and Salakhutdinov Ruslan. 2015 Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*. 843–852.
- [35]. Tian Yi, Ruan Qiuqi, An Gaoyun, and Fu Yun. 2016 Action Recognition Using Local Consistent Group Sparse Coding with Spatio-Temporal Structure. In *Proceedings of the 2016 ACM on Multimedia Conference ACM*, 317–321.
- [36]. Wang Heng and Schmid Cordelia. 2013 Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*. 3551–3558.
- [37]. Yosinski Jason, Clune Jeff, Nguyen Anh, Fuchs Thomas, and Lipson Hod. 2015 Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015).
- [38]. Zhu Yuke, Groth Oliver, Bernstein Michael, and Fei-Fei Li. 2016 Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4995–5004.

CCS CONCEPTS

- **Computing methodologies** → **Activity recognition and understanding;** *Image representations*; Classification and regression trees; Neural networks;

Figure 1:

Two examples of our two-step approach for activity recognition. The system first generates activity mask (middle column) and uses the mask as an additional image channel for activity recognition (right column).

Figure 2:

Our Conv-LSTM-Deconv cGAN structure with U-Net connection (dark shaded boxes on the left [27]). The numbers above convolution layers indicate the number of filters in each convolution kernel. (See digital version for color codes.) The arrow in the activity recognizer box points to the 7th layer neurons for which the activation maps are shown in Figure 5.

Figure 3:

Left: The generated masks and ground truth for a sequence of input frames. Right: The input video frames with activity bounding box generated using different generator structures. (See digital version for colors.)

Figure 4:
Generated role masks for trauma resuscitation dataset. (See digital version for colors.)

Figure 5:
The activation maps of neurons in the 7th convolutional layer of the activity recognizer (Figure 2) overlaid on input frames. Top row used masks and input video frames, while bottom row used video only. (See digital version for colors.)

Figure 6:

Top: the four types of mask-generating errors. Bottom: activity recognition error caused by visually similar frames. (See digital version for colors.)

Table 1:

Activity recognition accuracy and generated mask per-pixel accuracy for different loss combinations.

Loss functions	Activity rec acc'y	Per-pixel acc'y
L_A	62.21%	69.29%
$L_A + L_C$	64.68%	71.03%
$L_A + L_{st} + L_C$	78.16%	81.37%
$L_A + L_p + L_C$	70.92%	83.89%
$L_A + L_{st} + L_p + L_C$	81.75%	86.71%

Table 2:

Activities and background labels in our datasets.

Olympic Sports	Label #		Trauma team roles	Label #
Background	1		Background	1
Lay Up	2		Patient Bed	2
Bowling	3		Nurse Left	3
Weight Lifting	4		Respiratory Therapist	4
High Jump	5		Examine Provider	5
Long Jump	6		Team Leader	6
Lay Up	7			

Table 3:

Averaged accuracy and F-scores of different activity recognition methods for 6 activities in Olympic sports dataset.

Direct activity classification by other systems					Our activity classification approach using activity masks				
System	Acc.	Prec.	Rec.	F1	System	Acc.	Prec.	Rec.	F1
Ref. [13]	na	65.35	na	na	Conv-Deconv	40.10	64.01	40.10	38.77
Ref. [23]	na	74.93	na	na	Conv-LSTM-Deconv	61.92	72.89	61.92	61.45
Ref. [2],[9]	na	74.5/80.0	na	na	Our cGAN ($L_A, L_{st}&L_C$)	78.16	84.88	78.16	79.52
Ref. [19],[36]	na	91.2/91.4	na	na	Our cGAN ($L_A, L_P&L_C$)	70.92	83.63	70.92	73.78
ConvNet	57.38	69.34	57.38	62.80	Our cGAN ($L_A, L_{st}, L_P&L_C$, 6 activities)	81.75	87.79	81.75	84.66
Conv-LSTM	59.15	75.62	59.15	66.38	Our cGAN ($L_A, L_{st}, L_P&L_C$, 16 activities)	73.62	80.01	73.62	76.68

Table 4:

Recognition performance breakdown for individual activities and comparison of mean average precision with existing systems (the best scores are bolded). Shaded are the activities for which videos usually contained unrelated people.

Activity	Ref. [13]	Ref. [23]	ConvNet [12]	Conv-LSTM [29]	Conv-Deconv [36]	Conv-LSTM-Deconv [20]	Our cGAN (L_A & L_{st} & L_C)	Our cGAN (L_A & L_P & L_C)	Our cGAN (L_A, L_{st}, L_P & L_C)
Lay Up	75.8	77.9	85.7	35.48	67.20	64.00	89.92	87.87	96.73
Bowling	66.7	72.7	87.6	70.49	68.60	66.67	92.22	91.42	93.58
Snatch	41.8	69.2	81.3	65.61	76.20	93.83	87.61	80.18	93.33
High Jump	52.4	68.9	40.7	58.48	53.90	69.85	85.76	82.52	96.55
Long Jump	66.8	74.7	35.7	37.61	48.10	48.94	66.70	68.73	74.78
Gym Vault	88.6	86.1	76.5	65.97	42.40	51.91	43.24	45.17	48.83

Table 5:

The evaluation scores for the trauma resuscitation process using different methods.

Model Structure	Acc.	Prec.	Rec	F1 Score
ConvNet	0.51	0.60	0.51	0.55
Conv-LSTM	0.57	0.62	0.57	0.59
Ref. [15]	0.80	0.66	0.64	0.63
Ref. [17]	0.87	0.72	0.69	0.67
Our Model (binary label)	0.83	0.77	0.75	0.76
Our Model (role label)	0.86	0.81	0.78	0.79