Video-based Concurrent Activity Recognition for Trauma Resuscitation

Yanyi Zhang, Yue Gu, Ivan Marsic
Department of Electrical and Computer Engineering,
Rutgers University,
Piscataway, NJ, USA
{yz593, yg202, marsic}@rutgers.edu

Yinan Zheng, Randall S. Burd
Division of Trauma and Burn Surgery,
Children's National Medical Center,
Washington, DC, USA
{yzheng, RBurd}@childrensnational.org

2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374399

***Abstract*—We introduce a video-based system for concurrent activity recognition during teamwork in a clinical setting. During system development, we preserved patient and provider privacy by pre-computing spatio-temporal features. We extended the inflated 3D ConvNet (i3D) model for concurrent activity recognition. For the model training, we tuned the weights of the final stages of i3D using back-propagated loss from the fully-connected layer. We applied filtering on the model predictions to remove noisy predictions. We evaluated the system on five activities performed during trauma resuscitation, the initial management of injured patients in the emergency department. Our system achieved an average value of 74% average precision (AP) for these five activities and outperformed previous systems designed for the same domain. We visualized feature maps from the model, showing that the system learned to focus on regions relevant to performance of each activity.**

Keywords—concurrent activity recognition, clinical teamwork, video understanding

I. INTRODUCTION

We introduce a video-based activity recognition system for recognizing concurrent activities during clinical teamwork (Fig. 1). Following the success of deep learning in image recognition [1][2][3][4], deep neural networks have been applied to the problem of recognizing activities from videos [5][6][7][8]. Requirements of an activity recognition system in this domain are different from general systems in four aspects. First, the system should give real-time predictions during a relatively long interval that contains many activities, instead of providing a single classification for each video. Second, the system needs to be privacy preserving because RGB videos contain faces of the patients and providers. Third, the model should produce multi-label outputs for the simultaneously performed activities. Finally, the system needs to perform well in a noisy setting in which the participants or objects may be intermittently occluded. The noise in the predictions needs to be removed to merge the correct predictions into activity segments. We solved the first two problems using transfer learning and by segmenting the long video into several clips and feeding these clips into a pre-trained inflated 3D ConvNet (i3D) network [5], a 3D structure for extracting spatio-temporal features. We exported the pre-computed spatio- temporal features from the third stage output of the i3D for further analysis (Fig.1 left side). These features were used for model training. Feature extraction is irreversible, allowing privacy preservation. For concurrent activity recognition, we modified the output activation function to enable the model for the multi-label output. To eliminate noisy predictions (Fig. 1, right side), we applied a filtering algorithm that averages the values in a moving window to smooth out the predictions.

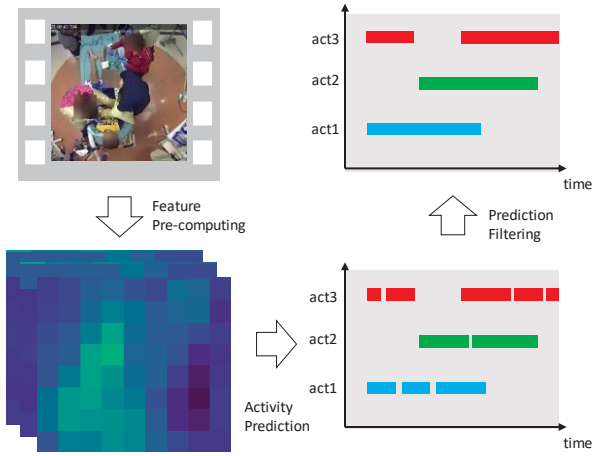

Fig. 1. System overview. The feature pre-computing module extracts spatio-temporal features from the video inputs. The activity prediction module uses the extracted features to make concurrent activity predictions. A prediction filtering module finally smoothens the model predictions. The example frame is from an actual trauma resusciation with faces blured for privacy.

We evaluated our system on five activities during trauma resuscitation. Trauma is the leading cause of mortality in children and young adults. The initial resuscitation of injured patients is critical for identifying and managing life-threatening injuries. To reduce errors in this setting, real time decision support has been evaluated as a method for reducing errors. Automatic activity recognition is needed to align this support with current task performance. The system that we developed for activity recognition achieved an average of 74% mAP on these five trauma resuscitation activities. This result was better than achieved in previous activity recognition systems in this domain. Our contributions are summarized as:

- We extended the an activity recognition algorithm (i3D) [5] to solve the four problems in our domain (real-time prediction, privacy preservation, concurrent activities, and noisy predictions).
- We evaluated our activity recognition system on using actual videos of trauma resuscitation, showing better performance than previous systems.

The rest of the paper is organized as follows. Section II reviews related work. Section III describes our system. Section IV presents the data collection and implementation details. Section V analyzes the experimental results. Section VI concludes the paper and includes proposed future work.

II. RELATED WORK

Medical Activity Recognition using Wearable Sensors: Activity recognition system in healthcare has been

978-1-7281-5382-7/20/$31.00 ©2020 IEEE

studied for several decades. Some systems have used wearable sensors [9][10][11]. For example, information from 3 axial accelerometers has been used as features to recognize activities in the operating room [9][10]. Surgical activities also have been recognized using the locations of objects and medical providers using RFID tags [11]. Sensor-based systems faces challenges for tracking medical workflow because sensors may interfere with work and require maintenance and operation by providers. The additional time required for placing sensors before starting the work may be unacceptable in time-sensitive settings. The position or properties of sensors also may interfere with the performance of clinical activities.

Medical Activity Recognition using Fixed Sensors: To address the issues of wearable sensors, some systems rely on fixed sensors. For example, medical activities have been tracking by using passive RFID tags on the medical tools, using the received signal strength indication (RSSI) collected from the RFID readers as features for activity prediction [13][14]. Speech obtained from fixed microphone also has been used as input for recognizing trauma activities [15]. These systems rely on fixed position detectors in the room and do not require the actions of providers but have limitations on achievable performance. RFID-based systems may require prior tagging of medical tools and cannot be used for activities that do not involve using taggable objects. The use of speech for activity recognition is promising but limited by the cost and time associated with manually generating training transcripts and has challenges for implementation in noisy settings [15]. Extracting representative features from raw data remains an important is challenge for these systems using fixed sensors.

Medical Workflow Analysis using Vision: Several systems have used video for recognizing medical workflow. For example, laparoscopic and ocular surgical videos have been used for recognizing surgical phases [23][25][26]. These previous works used videos focused only on specific regions (e.g., laparoscopic view, microscope view) in which the background was usually stationary. In contrast to activity recognition during trauma resuscitation, these models focus on localized regions. During trauma resuscitation, more than ten providers are moving in the scene and performing different activities. A model for trauma resuscitation needs to extract features correlated to these activities in a noisy environment.

General Activity Recognition: Activity recognition systems have proliferated in recent years because of the availability of deep learning algorithms and a growing experience in their application for image recognition. Many systems have applied deep neural network to the problem of general activity recognition (e.g., Kinetics-400 [16], UCF-101 [17], and Something-Something [18]). For example, the inflated 3D ConvNet (i3D) [5] and the non-local neural network [6] use spatio-temporal structures that have achieved state-of-the-art performance on Kinetics-400, a large scale video sets includes 400 daily life and sports activities. Temporal segment networks (TSN) [7] and temporal-spatial mapping (TSM) [8] have been used on Something-Something, a large collection of videos shows human performing actions using everyday objects. These networks randomly segment the videos and extracting the long-range spatio-temporal features by fusing the branches from different video segments. These systems work on general activity recognition and cannot be directly used for medical activity recognition because of the challenges of real-time prediction, privacy concerns, concurrent activities, and noisy data. Depth videos have been used instead of RGB videos to address privacy concerns but achieved only moderate performance because the depth videos are gray-scale videos and lack sufficient features for recognizing complex activities [27].

III. Methodology

Given a video of a trauma resuscitation case, our system recognizes medical activities in three steps:

- **A feature pre-computing (Fig. 2, left)** module extracts spatio-temporal features by feeding the video clips into a pre-trained i3D network and yielding the output from the third stage of i3D.
- **An activity prediction (Fig. 2 middle)** module uses the last two stages of the i3D network that takes the pre-computed features for extracting high-level features and tunes the weights back-propagated from the fully-connected layer for making activity predictions.
- **A prediction filtering (Fig. 2 right)** module smooths the model outputs to eliminate the noisy predictions.

A. Feature Pre-computing

Videos of trauma resuscitation includes the faces of patients and providers. Feature pre-computing is used to remove identifying facial information before training and testing the model. We feed the videos into the inflated 3D ConvNet (i3D) [5] pre-trained on Kinetics-400 data [16] and obtain the third stage's output of the i3D as the feature representation of the video frames. This method further improves the model by loading the pre-trained weights from a large-scale activity recognition dataset instead of training only on available data.

1) Video Pre-processing: Before feeding the videos into the i3D network for feature pre-computing, we pre-processed the videos to obtain the required input dimensions of the i3D network. The required dimension for the i3D input are "64×224×224×3", which takes 64 continuous frames as an input instance with each frame being "224×224×3." We segmented the video for each case into clips of 64 contiguous frames and resized the frames to the dimensions required by the i3D.

2) Inflated 3D ConvNet (i3D): Inflated 3D ConvNet (i3D) is a structure for extracting spatio-temporal features, which works on video understanding tasks, e.g., activity recognition and video classification [5]. The main idea of the i3D is to extend an existing 2D image recognition model with the time dimension for successive frames using a 3D ConvNet. The i3D model extended the inception v1 [3] 2D recognition model as:

$$f_2(i,j) = \sum_h \sum_w x(j+h, i+w) \cdot W_{c2d}(h,w) \quad (1)$$

$$f_3(k,i,j) = \sum_t \sum_h \sum_w x(k+t, j+h, i+w) \cdot W_{c3d}(t,h,w) \quad (2)$$

where $f_2(i,j)$ denotes each feature element after the 2D convolution filter W_{c2d} is applied on the input frame. After extension to the 3D ConvNet, $f_3(k,i,j)$ denotes each feature

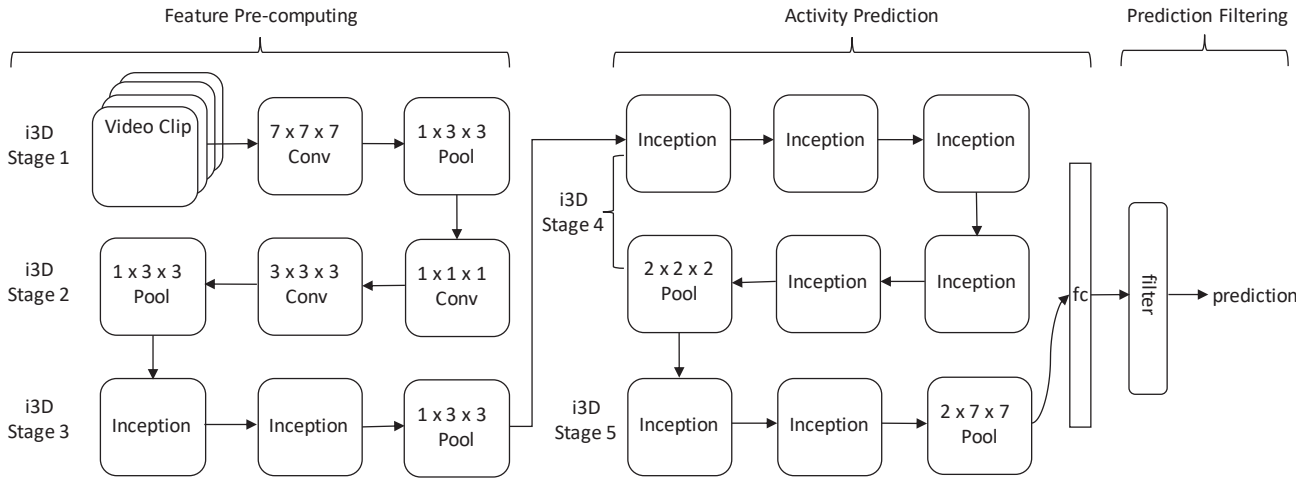

Fig. 2. Detailed composition of each module. The pre-computing module extracts features using the first three stages from the pre-trained i3D network. The ``incept'' element in the diagram is the inception block in Inception v1 [3] after extended into 3D. The pre-computed features are then fed into the last two stages in the i3D and the fully-connected layer for activity prediction. The prediction filtering module applied an average filter to smooth the model predictions.

element produced by applying 3D convolution filter W_{c3d} on the input video clip. The spatio-temporal features extracted by 3D ConvNet represent salient motion in the input clip of the region that is representative of each activity.

The feature pre-computing used this i3D model that was pre-trained on Kinetics-400 [16], a large-scale activity recognition dataset that includes more than 200,000 videos. Each video clip was fed into the pre-trained i3D. The output of the third stage was exported for training and evaluating the model. The features from the third stage contain less information than those from lower stages (first and second stage) because of information loss in the max-pooling layer at the end of each stage. We chose the third stage features because they require less memory storage during the exporting process and do not lead to performance loss. Our system is privacy-preserving by using these pre-computed features instead of the original videos because the convolution operations and nonlinear functions in i3D model are irreversible. The output from the third stage of the i3D has already passed more than 15 convolution and nonlinear functions.

B. Activity Prediction

We next applied the last two stages of the i3D (Fig. 2, Stages 4 and 5) that takes the pre-computed features as the input for extracting high-level features corresponding to each activity. The weights of these i3D stages can be tuned with the loss propagated from the fully-connected layer (Fig. 2, *fc*) that is applied on the output of the pooling layer (Fig. 2, Stage 5). The fully-connected layer is represented as:

$$y_{out}(i,j) = \sigma(f_{3d} \cdot W_{fc} + b) \qquad (3)$$

where y_{out} is the output of our model for activity prediction and σ is the activation function. W_{fc} and b are the weights and bias of the fully-connected layer, respectively. Reviewing the ground truth data (activity labels), we found that the team activities frequently overlapped and that several activities were performed concurrently. To enable the model provide multi-label outputs, we used the "sigmoid" function, which constrains the values of the output neurons to the range [0, 1]. We did not use the "softmax" function because it normalizes the neuron output values to sum to one. With this function, the strongest-predicted activity would exclude any concurrent activities for which the predictions are weaker.

C. Prediction Filtering

After the model made activity predictions for all the video clips, we compared the results to the ground truth and found that the model predictions were noisy and incorrect during some intervals. These errors were mainly caused by visual occlusion, when the individuals or objects related to the activity were blocked from view. When the coders are labeling the ground truth, their knowledge of the process is used to code despite these occlusions. A solution was needed that did not depend on this human interaction. To eliminate these errors, we applied a filtering algorithm that smooths the model results by averaging the point-predictions of each activity over a moving window as:

$$y_{out}{}'(c,a) = \frac{1}{N_a}\sum_{n=-\frac{N_a}{2}}^{\frac{N_a}{2}} y_{out}(c+n,a) \qquad (4)$$

where $y_{out}{}'(c,a)$ is the smoothed prediction for activity a in the c^{th} video clip and N_a is the size of the moving window. Each activity was smoothed using different window sizes. We used the average duration of each activity from the ground truth to set the size for each moving window (Table I). We used the average duration of the activity as its window size to ensure that the activity prediction continues at its specific duration. The evaluation results errors were eliminated achieved around 7% mean average precision (mAP) enhancement (Table II, last row).

IV. Data Collection and Implementation Details

To develop and test our system, we used videos from 230 trauma resuscitation cases: 185 (80%) for training and 45 (20%) cases for evaluation. The videos were recorded using a recording system that includes a camera mounted on the ceiling over the patient bed. The system starts recording when motion is detected in the room. The videos are stored on a secure server in the hospital. The use of videos for research purposes has been approved by the Institutional Review Board at the hospital. The average length of these video was 25 minutes, ranging from 16 to 35 minutes. The recording speed is 30 frames per second (fps) with a

TABLE II. MODEL EVALUATION IN AVERAGE PRECISION

Activity Name	Duration (avg/std)	Freq
c-spine stabilization (CS)	248.6/127.2	0.45
manual blood pressure (MBP)	32.3/12.0	0.02
oxygen administration (OX)	119.4/55.7	0.09
intravenous catheter placement (IV)	131.5/46.1	0.07
back assessment (BK)	43.9/32.4	0.06

resolution of 640×480 pixels. The ground truth data was labeled using manual video review. Reviewers are trained in ground truth coding, only coding videos used for this study after their coding performance was validated. A data dictionary was used to define more than 200 activities relevant to trauma resuscitation. For this study, we focused on five medical activities that are frequently performed and are clinically important during trauma resuscitations: cervical spine (c-spine) stabilization (CS), obtaining a manual blood pressure (MBP), administration of face mask oxygen (OX), placement of an intravenous catheter (IV), and assessment of the back for injuries (BK). The duration of these activities varied from 14 s to 248 s (Table I).

We implemented our model (Fig. 2) using Keras with the TensorFlow backend. The i3D network [5] was implemented based on the published source code [19]. We used batch normalization [20] and ReLU activation in all the convolution layers. We used binary cross-entropy loss and the SGD optimizer with an initial learning rate (LR) 1e-4 and manually decreased the LR to 1e-5 after the training loss was saturated. Dropout was used after the fully-connected layer to avoid overfitting [21]. We used 12 video clips in each batch and trained the model using 3 RTX 2080 Ti GPUs (four for each GPU) for 50k iterations (40k with LR 1e-4 and 10k with LR 1e-5). The model took about 24 hours to converge.

V. EXPERIMENT

We evaluated and analyzed the performance of our system. Medical experts evaluated potential sources of poor performance by reviewing the corresponding source videos. We also compared our system performance with two systems previously developed for activity during recognition trauma resuscitation [14][15].

A. Evaluation Metrics

Because of activity concurrency, we used average precision (AP) as an evaluation metric instead of other metrics, such as accuracy, precision, recall, and F1 score. Accuracy is used for evaluating category classifiers, but our model for concurrent activity recognition produces multi-label outputs. For the binary evaluation metrics (precision, recall, and F1-score), an arbitrary threshold is needed to convert the predicted probabilities into binary predictions. We used average precision (AP) because it has good discrimination and stability when used to evaluate multi-label classification systems without using arbitrary thresholds [22]. The AP is the average of the precision values when thresholding the predictions using the confidence score from the k[th] sample (ranked by confidence score):

$$AP = \frac{1}{|Y^+|}\sum_{k \in Y^+} Precision(R_k) \quad (5)$$

TABLE I. ACTIVITY AVERAGE DURATIONS

Activity Name	Methods & AP	
	i3D	*i3D + filter*
c-spine stabilization (CS)	0.88	0.92
manual blood pressure (MBP)	0.79	0.85
oxygen administration (OX)	0.74	0.82
intravenous catheter placement (IV)	0.26	0.35
back assessment (BK)	0.66	0.75
mean average precision (mAP)	0.67	0.74

where Y^+ is the number of positive ground truth for this activity and R_k is predictions based on a threshold using the confidence score of k^{th} sample.

B. Results Analysis

Based on the AP score for each activity, our system perform well when predicting four activities (CS, MBP, OX and BK) because the model can capture features of these objects related to their corresponding activities from the RGB videos. For example, during manual BP measurement, a large blood pressure measuring instrument was placed near the patient's bed (red rectangle in the top second diagram of Fig. 3). When the activity oxygen was administered, a non-rebreather mask (NRB) is placed on the patient's face (red rectangle in the top third diagram of Fig. 3). These objects are accurately detected by the model. C-spine stabilization (CS) achieved the best AP score among the four activities detected, likely because of a high frequency of performance. In a classification problem, classes having more positive labels will usually achieve higher AP scores. The manual blood pressure (MBP) achieved better performance than the other two activities although having lower frequency. For this activity, the tools used for blood pressure measurement are visually obvious and seldom occluded in the video recordings.

Using mAP (Table II, last row), prediction filtering led to a 7% higher AP than results without using filtering for each activity. We reviewed the model predictions and their corresponding ground truth. Important performance gains resulted from eliminating noisy predictions. The model incorrectly predicted activities when the relevant regions of the video were occluded, a finding discovered by reviewing corresponding RGB videos. The filter that averages the predictions in a moving window eliminated most of these errors unless the occlusion was long (i.e., larger than half of the duration of the activity in TABLE I).

The system achieved the lowest performance for IV placement. We reviewed the model predictions and their corresponding ground truth for this activity and found that the model predictions correctly matched the ground truth in some cases. We asked coders to review the videos in the testing cases in which the model performed well to find possible explanations for these results. IV placement instances were labeled by coders as starting when a tourniquet was placed on the patient's arm and as ending when the tourniquet is removed. The model did not perform well on in most of the cases because the tourniquets were occluded by the providers' arms and the blankets covering the patient after being applied. For other activity types,

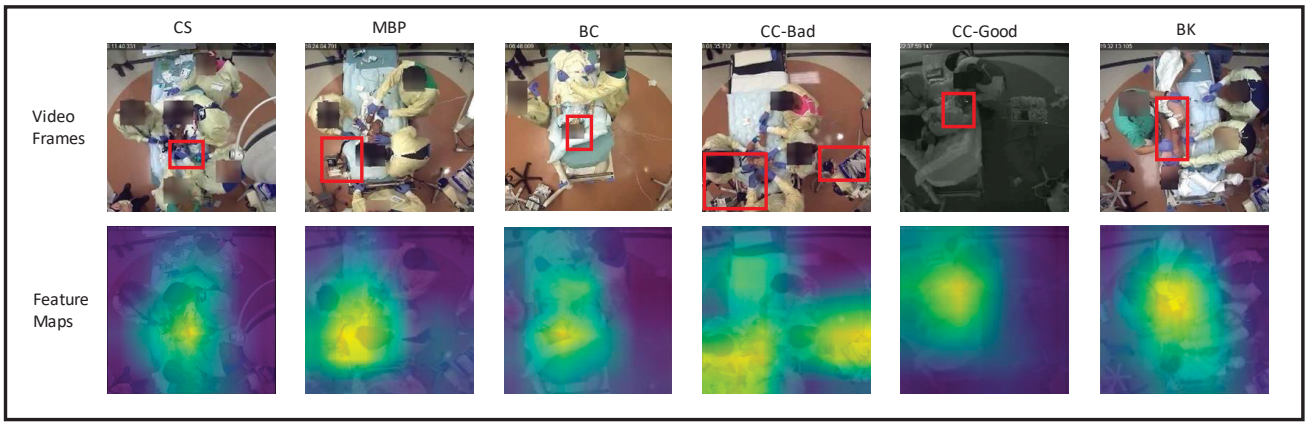

Fig. 3. Model visualization. Examples of feature maps from the output of the last convolution layer after fine-tuningand their corresponding video frames for the five activities (using abbreviations of the activity names). The faces of the patients and medical providers are blured to preserve privacy. The red in the video frames (first row) are the regions that have higher values (>0.5 after normalized) in the corresponding feature maps (second row).

TABLE III. PERFORMANCE COMPARASION WITH OTHER METHODS

Activity Name	Methods & Acc.		
	RFID [14]	*Speech [15]*	*Ours*
c-spine stabilization (CS)	-	0.67	0.76
manual blood pressure (MBP)	0.64	0.77	0.89
oxygen administration (OX)	0.54	0.76	0.80
back assessment (BK)	-	0.68	0.75

relatively large objects were used and were not occluded during activity performance. In some cases of IV placement, the providers turned off the lights and used a flashlight to illuminate the patient's arm to better visualize the vein. In these cases, the model likely performed well because it learned to recognize the activity based on differences in illumination.

To confirm these explanations, we visualized the learned feature maps from the output of the last convolution layer of our model (Fig. 3, second row). We normalized the values of the feature maps and inserted the bounding boxes in the original video frames for the regions where the feature values in the maps were larger than 0.5. The feature maps had high values in the regions around the objects or human gestures relevant to the performed activity. For example, the feature map for the back evaluation (Fig. 3, bottom last map) highlights the patient's back when this activity was performed. The feature map for the IV placement was highlighted in several unrelated regions, suggesting that the model did not always extract features that are representative of this activity (Fig. 3, bottom fourth map). When the model performed well for IV placement, the feature map sometimes highlighted the region that the flashlight illuminated (Fig. 3, bottom fifth map). These visualized feature maps support the explanation of our results.

C. *Comparasion with Existing Systems for Activity Recognition in Trauma Resuscitation*

We compared our system with two previous systems for activity recognition during trauma resuscitation [14][15]. Four of the activities (CS, MBP, OX and BK) detected were also detected by these two systems. Because these systems were not able to recognize concurrent activities, a complete comparison is not possible. To make a fair comparison, we removed all the video clips with concurrent activities in the ground truth and used the "argmax" function to convert the predictions into activity categories. Our system outperformed these systems on the four shared activities (TABLE III), likely because our system used the rich features extracted from RGB videos. In addition, our system did not need human effort beyond ground truth labeling. One system used RFID as an adjunct sensor, requiring placement of RFID tags on objects and maintenance of an ongoing record of tag-to-object correspondence [14]. The second system required manual transcriptions of verbal communication during each resuscitation [15].

VI. CONCLUSION AND FUTURE WORK

We introduce a video-based activity recognition system for recognizing concurrent activities during clinical teamwork. The system gives concurrent activity predictions. Our system is privacy-preserving and outperformed other existing activity recognition systems designed for trauma resuscitation. Our system has two limitations. First, the pre-computed features are ten times larger than the original videos, making additional training reliant on large hardware capacity for storing these features. Second, the system recognizes activities based only on video. Some activities in this domain cannot be recognized using videos only. For example, verbal reports of findings for which observable activities are limited will need to rely on speech recognition. Our next steps will be to develop a runtime implementation of these models to assess performance in a real-world application. We will also develop a more efficient way to store pre-computed features and build a network that fuses visual features and other modalities. Finally, we will develop an efficient approach to integrate new training data that does not rely on complete model retraining.

ACKNOWLEDGMENT

We would like to thank three reviewers' valuable feedback, and the trauma experts from Trauma and Burn Surgery, Children`s National Medical Center for their work on data collection and processing. This research has been funded under NIH/NLM grant 2R01LM011834-05 and NSF grants IIS-1763827, IIS-1763355, and IIS-1763509.

REFERENCES

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[6] Wang, Xiaolong, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[7] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, Cham, 2016: 20-36.

[8] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 7083-7093.

[9] Ahmadi, S. A., et al. "Introducing wearable accelerometers in the surgery room for activity detection." Computer-und Roboter-Assistierte Chirurgie (CURAC) (2008).

[10] Meißner C, Meixensberger J, Pretschner A, et al. Sensor-based surgical activity recognition in unconstrained environments[J]. Minimally Invasive Therapy & Allied Technologies, 2014, 23(4): 198-205.

[11] Bardram J E, Doryab A, Jensen R M, et al. Phase recognition during surgical procedures using embedded and body-worn sensors[C]//2011 IEEE international conference on pervasive computing and communications (PerCom). IEEE, 2011: 45-53.

[12] Chakraborty I, Elgammal A, Burd R S. Video based activity recognition in trauma resuscitation[C]//2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 2013: 1-8.

[13] Li, Xinyu, et al. "Activity recognition for medical teamwork based on passive RFID." 2016 IEEE International Conference on RFID (RFID). IEEE, 2016.

[14] Li X, Zhang Y, Marsic I, et al. Deep learning for rfid-based activity recognition[C]//Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM. 2016: 164-175.

[15] Gu Y, Zhang R, Zhao X, et al. Multimodal Attention Network for Trauma Activity Recognition from Spoken Language and Environmental Sound[C]//2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 2019: 1-6.

[16] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.

[17] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[18] Goyal R, Kahou S E, Michalski V, et al. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense[C]//ICCV. 2017, 1(4): 5.

[19] Carreira, Joao, and Andrew Zisserman. Keras implementation of inflated 3d from Quo Vardis paper + weights (2017), GitHub repository, https://github.com/dlpbc/keras-kinetics-i3d

[20] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

[21] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[22] Wu X Z, Zhou Z H. A unified view of multi-label performance measures[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 3780-3788.

[23] Zia A, Hung A, Essa I, et al. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 273-280.

[24] Twinanda A P, Shehata S, Mutter D, et al. Endonet: A deep architecture for recognition tasks on laparoscopic videos[J]. IEEE transactions on medical imaging, 2016, 36(1): 86-97.

[25] Loukas C. Surgical phase recognition of short video shots based on temporal modeling of deep features[J]. arXiv preprint arXiv:1807.07853, 2018.

[26] Zisimopoulos O, Flouty E, Luengo I, et al. Deepphase: surgical phase recognition in cataracts videos[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 265-272.

[27] Zhang Y, Li X, Zhang J, et al. Car-a deep learning structure for concurrent activity recognition[C]//2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 2017: 299-300.