J Inf Process Syst, Vol.14, No.1, pp.150~161, February 2018
https://doi.org/10.3745/JIPS.04.0059

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

DeepAct: A Deep Neural Network Model for Activity Detection in Untrimmed Videos

Yeongtaek Song* and Incheol Kim*

Abstract

We propose a novel deep neural network model for detecting human activities in untrimmed videos. The process of human activity detection in a video involves two steps: a step to extract features that are effective in recognizing human activities in a long untrimmed video, followed by a step to detect human activities from those extracted features. To extract the rich features from video segments that could express unique patterns for each activity, we employ two different convolutional neural network models, C3D and I-ResNet. For detecting human activities from the sequence of extracted feature vectors, we use BLSTM, a bi-directional recurrent neural network model. By conducting experiments with ActivityNet 200, a large-scale benchmark dataset, we show the high performance of the proposed DeepAct model.

Keywords

Activity Detection, Bi-directional LSTM, Deep Neural Networks, Untrimmed Video

1. Introduction

In recent years, there has been rapid growth in the production and consumption of a wide variety of video data due to the popularization of relatively low priced, high quality video devices such as smartphones, digital cameras, and camcorders. It has been reported that on YouTube about 300 hours of video data updates occur every minute. Along with the growing production of video data, new technologies for convenient consumption of video data are also attracting attention. These include technologies such as video captioning, video question-answering, and video-based activity/event detection.

From input video data, in general, the human activity detection not only guesses what activity is contained in the video (as shown in the top of Fig. 1), but also localizes the regions in the video where the activity actually occurs (as shown at the bottom of Fig. 1). Such human activity detection technologies have many useful applications such as the elderly care service systems, video-based surveillance/security systems, unmanned autonomous vehicles, intelligent home service robots, and others.

Existing state-of-the-art approaches to the video-based activity detection task suffer from one or more of the following major drawbacks: they do not learn deep representations from videos, but rather use hand-crafted features [1,2]. Such features may not be optimal for localizing activities in diverse

Manuscript received May 8, 2017; first revision August 7, 2017; accepted September 1, 2017.
Corresponding Author: Incheol Kim (kic@kyonggi.ac.kr)
* Dept. of Computer Science, Kyonggi University, Suwon, Korea (dudtroc@gmail.com, kic@kyonggi.ac.kr)

video domains, resulting in inferior performance. Furthermore, current methods' dependence on external proposal generation or exhaustive sliding windows leads to poor computational efficiency.

In this paper, we propose a novel deep neural network model for detecting human activities in untrimmed videos. In the proposed DeepAct model, two different convolutional neural network (CNN) models, C3D and I-ResNet, are used to extract rich features that represent unique patterns of each activity from video segments. A new bi-directional recurrent neural network (RNN) model, BLSTM [3], is used to identify human activities from the sequence of extracted feature vectors. We conduct experiments with ActivityNet 200 [4], a large-scale video benchmark dataset, and show the high performance of the proposed DeepAct model. In the subsequent Section 2, the related works are summarized briefly. After the DeepAct model is explained in detail in Section 3, the experimental results for evaluating the performance of the proposed DeepAct model are presented in Section 4. In the final Section 5, the conclusions and some discussions are presented.

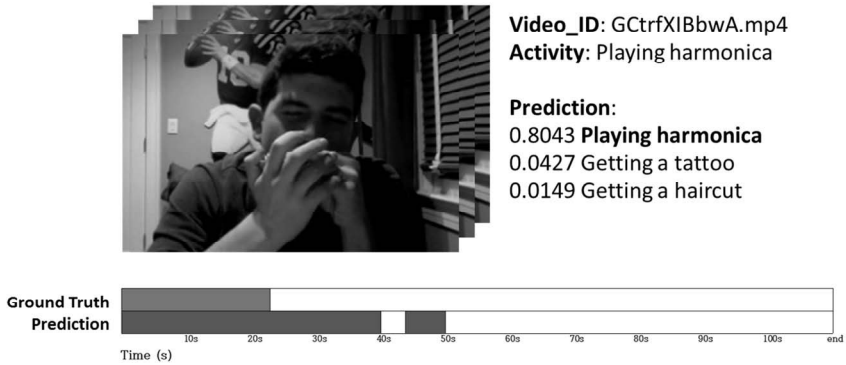

Fig. 1. Example of human activity detection in a video.

2. Related Works

There is a long history of activity recognition, or classifying trimmed video clips into fixed set of categories [5–7]. Unlike activity recognition from trimmed videos, activity detection needs to predict the start and end times of the activities within untrimmed and long videos [8–10]. Existing activity detection approaches are dominated by models that use sliding windows to generate segments and subsequently classify them with activity classifiers trained on multiple features [11–15]. Most of these methods have stage-wise pipelines which are not trained end-to-end. Moreover, the use of exhaustive sliding windows is computationally inefficient and constrains the boundary of the detected activities to some extent. In [11], a new CNN-based activity detection method has been proposed. With temporal segment network (TSN), this method attempts to overcome the long-range temporal structure modeling problem when a CNN model is applied to video-based activity detection. After an input video is divided into K segments, a short snippet is randomly selected from each segment. Instead of working on single frames, TSN, which is a pair of the spatial CNN and the temporal CNN, operate on a sequence of short snippets sparsely sampled from the entire video. Each snippet in this sequence produces its own prediction of the activity classes. The class scores of different snippets are fused by the consensus function to yield segmental consensus, which is a video-level activity prediction. Predictions from all

modalities, such as temporal and spatial consensus, are then fused to produce the final activity classification. This method has the advantage of complementary employment of features, such as optical flow features and RGB features, for activity classification. However, due to the nature of the CNN-based classification model, this method cannot make use of rich contextual information over adjacent segments to identify each video segment's activity. In [12], global video-level features, such as ImageNet Shuffle and motion boundary histogram (MBH), are used for untrimmed video classification task. On the other hand, frame-level features, such as C3D [16], are used for activity proposal generation and scoring. On three different types of features, a one-versus-rest linear support vector machine (SVM) is trained for each activity class, and then the resulting SVM scores are used for untrimmed video classification. Untrimmed video classification is achieved by fusing all video level scores using a linear SVM as a meta-classifier. In this work, activity localization is performed by training a binary random forest (RF) classifier for each class on the frame-level C3D features, and casting activity region proposal generation as an optimization problem, which makes use of these binary decisions. The scheme proposed in this work has the advantage that it can share the C3D features both for activity classification and for activity localization. However, it has some computational burden to extract both video-level and frame-level features from a long untrimmed videos.

Recently, some approaches have bypassed the need for exhaustive sliding window search to detect activities with arbitrary lengths. [17,18] achieve this by modeling the temporal evolution of activities using RNNs or Long Short Term Memory (LSTM) networks and predicting an activity label at each time step. The deep action proposal model in [17] uses LSTM to encode C3D features of every 16-frame video chunk, and directly regresses and classifies activity segments without the extra proposal generation stage. In [18], an entire untrimmed video is first divided into multiple segments, and then C3D features are extracted from video segments. The features from a C3D CNN are used as input to train a uni-directional RNN, LSTM, which learns to classify video segments. After segment prediction, the outputs of the LSTM model are post-processed to assign a single activity label to the entire video, and the temporal boundaries of the activity within the video are determined for activity localization. This LSTM-based activity detection method is able to improve the performance to some degree by learning the sequential patterns from successive video segments. However, it has the limitation of not being able to use a greater variety of features, other than the C3D features, for the activity classification.

3. Activity Detection Model

The overall process of activity detection in video with our proposed DeepAct model is illustrated in Fig. 2. This process comprises four stages: preprocessing, feature extraction, activity classification, and activity localization. In the preprocessing stage, the entire video is divided into segments consisting of 16 frames, and then each video segment is resized according to the input format for two different CNN models (C3D, I-ResNet). In the following feature extraction stage, two different CNN models (C3D and I-ResNet) are used to extract mutually complementary features from each video segment, and then these features are merged into one feature vector. In the activity classification stage, the bi-directional RNN model, BLSTM, is used to compute the score of each activity from series of feature vectors, and then identify the most likely activity contained in the video. In the final activity localization stage, the temporal regions where the activity occurs in the video are identified.

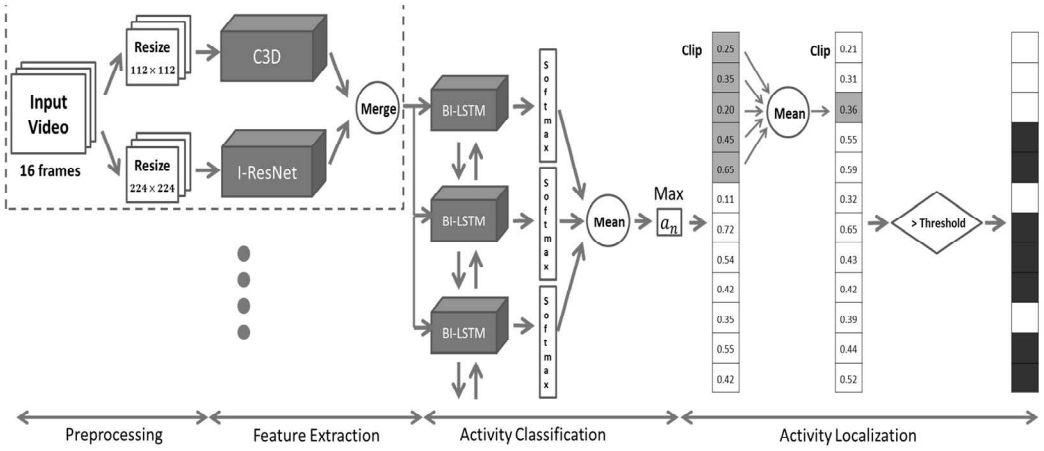

Fig. 2. The process of activity detection in video.

3.1 Preprocessing and Feature Extraction

In the preprocessing stage, the whole video is divided into video segments of a certain size, and the video segments are resized to fit two different convolutional neural networks for feature extraction. First, we divide the whole video into segments of 16 frames. And then each video segment is resized to 112×112 for input to the C3D network, and to 224×224 for input to the I-ResNet network. In the feature extraction stage, C3D and I-ResNet CNN models extract different features from each video segment. Two different features extracted from the same video segment are then merged into one feature vector for activity classification.

While 2D CNN, the two-dimensional convolutional neural network model, is useful for learning features from images, C3D, an extended three-dimensional CNN model, is known to be effective for learning features from videos [3]. The video features extracted by C3D model contain not only the spatial information embedded in each image frame, but also the temporal information across multiple image frames. For this reason, C3D features are very useful in detecting human activities in videos.

We notice that many daily human activities, such as participating in ball games, playing musical instruments, often involve specific objects or tools. Detecting these objects or tools in a video can therefore be of great aid in identifying activities associated with them. In this study, instead of directly finding the objects in the video to detect the associated activity, the proposed model is designed to use additional object features extracted from each video segment. The object features are extracted by using I-ResNet model, which is a convolutional neural network (ResNet) model [19] pre-trained with the large-scale object image dataset called ImageNet. The object features extracted by the I-ResNet model are combined with the activity features extracted by the C3D model to produce one feature vector for identifying the activity contained in each video segment.

Fig. 3 shows the feature extraction process using two different CNN models: C3D and I-ResNet. From a video segment composed of 112×112 sized frames, the C3D model produces the activity features of 4096 dimensions. On the other hand, from a video segment consisting of 224×224 sized frames, the I-ResNet model outputs the object features of 1000 dimensions.

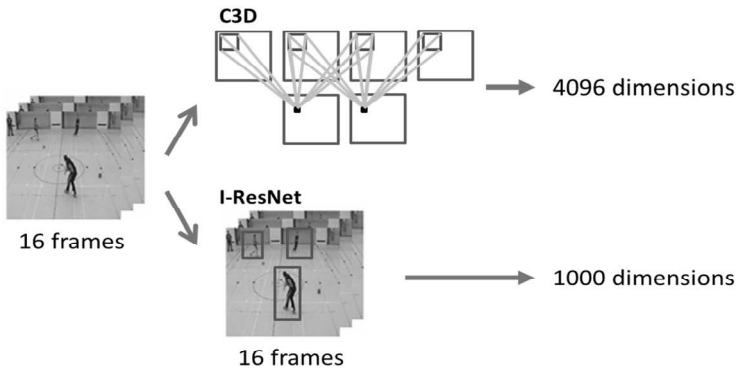

Fig. 3. Feature extraction with two different convolutional neural networks.

3.2 Activity Classification

Through the pre-processing and the feature extraction stages, an untrimmed input video produces a sequence of feature vectors, which individuals are extracted from video segments. In order to identify an activity contained in a video from the sequence of feature vectors, we use a RNN model. Different from CNN models, RNN models like LSTM are known useful for learning the time series patterns. In this study, we adopt a new bi-directional LSTM model, BLSTM, in order to identify an activity from the sequence of video feature vectors.

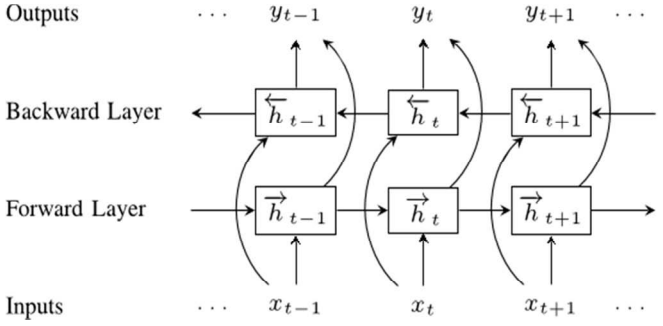

Fig. 4. Bi-directional LSTM (BLSTM) model.

One shortcoming of uni-directional RNN models including LSTM is that they are only able to make use of the previous context. BLSTM models do this by processing the data in both directions with two separate hidden layers, which are then feed forwards to the same output layer. Fig. 4 shows a typical structure of the BLSTM model. By adding the reverse hidden layer to the existing uni-directional LSTMs, the BLSTM model is able to make use of the future context as well as the previous one for deciding the current output. Due to this advantageous characteristic, the BLSTM model adopted in this study can be expected to greatly improve the performance of activity classification. However, because it has more internal parameters to learn than the uni-directional LSTM model, we expect that it will require a little longer training time.

As shown in Fig. 2, the output produced by the BLSTM model for each segment feature vector of a video is converted to the classification score for each activity of the pertaining segment through the fully

connected Softmax layer. And then the final activity contained in the entire video is determined by comparing the averages of the classification scores for each activity obtained from each segment. Fig. 5 shows an example of the classification scores of each activity a_j computed by the BLSTM model and their average calculated by the fully connected Softmax layer for each video segment t_i. In the example in Fig. 5, the activity with the highest average (in this example, 0.6502) a_1 is determined as the main human activity within the entire video.

16 frames

		t1	t2	t3	t4	t5	t6	...	Average
Activity	a_1	0.4561	0.5671	0.6235	0.6333	0.5121	0.2241	...	0.6502
	a_2	0.1212	0.0154	0.1541	0.1211	0.0121	0.0454	...	0.1018
	a_3	0.0111	0.0011	0.0012	0.0041	0.0021	0.0031	...	0.0100
		●							
		●							
		●							
	a_{200}	0.0241	0.0121	0.0052	0.0032	0.0074	0.0049	...	0.0057

Fig. 5. Classification score for each activity (a_j) per video segment (t_i).

3.3 Activity Localization

Activity localization is the task of identifying the video segments in which the particular human activity a^* that is found through the activity classification stage actually occurs in the video. In this study, the classification scores for the activity a^* computed in individual video segments during the activity classification stage are first smoothed. After the smoothing process, all segments whose classification score for the activity a^* is higher than the predefined threshold are then identified as temporal segments in which the activity a^* actually occurs.

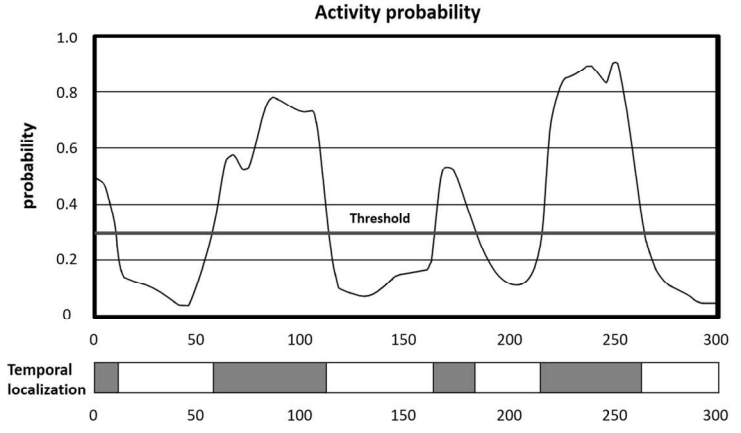

Fig. 6. Threshold-based activity localization.

Fig. 6 shows an example of the threshold-based activity localization. The upper part of Fig. 6 is a probability distribution representing the smoothed classification scores for the activity $a^{\star}$ in a video. The segments with a higher probability than the given threshold, indicated by the horizontal line, are all determined as temporal segments in which the activity $a^{\star}$ actually appears in the video, as shown in the shaded areas at the bottom of Fig. 6.

4. Experiments

The video dataset used for evaluating the performance of our DeepAct model is ActivityNet 200, a large-scale video benchmark, provided by ActivityNet Challenge 2016 [4]. ActivityNet 200 is structured around a semantic ontology which organizes human activities according to social interactions and where they usually take place. It provides a rich activity hierarchy with at least four levels of depth. ActivityNet 200 also provides samples from 200 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours. In our experiment, we use 10,024 video samples for training, 4,926 for validation, and 5,044 for testing, for each activity class.

For the experiments, we implemented the proposed DeepAct model using Keras, which is a deep learning library in Python, in an Ubuntu 14.04 UTS environment. The experiments were performed on a machine equipped with a 4.0-GHz 4 cores, 8 threads CPU and a GeForce GTX TITAN X GPU card. The optimization algorithm for learning the BLSTM model was set to RMSprop. In our experiments, the batch size was set to 256, the number of epochs to 20, and the learning rate was set to 10^{-5}.

In the first experiment, we evaluated the performance of the proposed C3D + I-ResNet feature model for activity classification. For this purpose, we compared the C3D single feature model with the proposed C3D + I-ResNet integrated feature model.

Table 1. Comparison of feature models

	LSTM		BLSTM	
	C3D	**C3D + I-ResNet**	**C3D**	**C3D + I-ResNet**
Time (min)	108	130	38	47
Epoch	100	100	20	20
Time/epoch	1 min 5 sec	1 min 18 sec	1 min 54 sec	2 min 21 sec
Clip accuracy	0.4878	0.5248	0.5140	0.5426
Video accuracy	0.5094	0.5349	0.5118	0.5522

Learning rate=1e-5.

Table 1 shows the results from the first experiment. When two different classification models (LSTM and BLSTM) were used, Table 1 shows the time it took to train the classification model (time), the number of repetitions (epoch), the mean training time per epoch, the classification accuracy for each segment (clip accuracy), and the classification accuracy for the video (video accuracy). The results in Table 1 show that the training time is slightly increased when the integrated C3D + I-ResNet feature model is used, as compared with the single C3D feature model. However, in terms of classification accuracy, the proposed C3D + I-ResNet integrated feature model outperforms greatly the C3D single

model. Based on the experimental results, we are sure that the object features extracted by the I-ResNet model can be supplementary for the spatiotemporal activity features produced by the C3D model for detecting human activities in videos.

In the second experiment, we evaluated the performance of the proposed BLSTM model for activity classification. For this purpose, we compared four different classification models: fully connected Softmax (Fc-softmax), simple RNN, LSTM, BLSTM. In this experiment, the C3D+I-ResNet integrated feature model was commonly used for extracting features from video segments.

Table 2. Comparison of classification models

	Fc-Softmax	RNN	LSTM	BLSTM
Time (min)	108	130	130	47
Epoch	100	100	100	20
Time/epoch	1 min 5 sec	1 min 18 sec	1 min 18 sec	2 min 21 sec
Clip accuracy	0.4857	0.4813	0.5248	0.5426
Video accuracy	0.5127	0.5013	0.5349	0.5522

Learning rate=1e-5.

Table 2 shows the results of the second experiment. As shown in Table 2, the classification performance of our BLSTM model (0.5522) was the highest, when compared to those of other classification models (0.5127, 0.5013, and 0.5349). Moreover, both the total time (47 min) and the epochs (20) spent in training our BLSTM model were measured as the shortest. However, we also notice that the training time per epoch of BLSTM model (2 min 21 s) was somewhat longer than those of other models (1 min 5 s, 1 min 18 s, and 1 min 18 s), as it had been expected. Based on these results, we are sure that the high context-dependency of the proposed BLSTM model helps to identify human activities more precisely from the sequence of feature vectors.

In the third experiment, we quantitatively evaluated the activity localization performance of our DeepAct model. For this purpose, we compared the localization performance of our DeepAct model with those of state-of-the-art models. Results are measured by mAP (mean average precision) with different temporal intersection over union (tIoU) thresholds from 0.5 to 0.95. tIoU was calculated as the following way:

$$\text{tIoU} = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{L_g \cap L_p}{(L_g \cup L_p) - L_g \cap L_p} \quad (1)$$

where L_g is the ground truth region and L_p is the predicted region.

Table 3. Comparison with previous models in terms of activity localization

	mAP			
	$\theta > 0.5$	$\theta > 0.75$	$\theta > 0.95$	Average
Wang and Tao [8]	42.28	3.76	0.05	14.85
UPC [18]	34.81	23.08	8.77	22.22
Shou et al. [9]	43.83	25.88	0.21	22.77
Xiong et al. [10]	39.12	23.48	5.49	23.98
DeepAct	37.83	24.82	9.96	24.02

θ = tIoU (temporal intersection over union).

Table 3 shows the results of the activity localization experiment. As shown in Table 3, the average mAP of our DeepAct model (24.02%) was the highest, when compared to those of other state-of-the-art models (14.85%, 22.22%, 22.77%, and 23.98%). We also notice that the localization performance of our DeepAct model was relatively higher than those of other models as the tIoU threshold was increased. Based on these results, we are sure that our DeepAct model has higher localization performance than the state-of-the-art models.

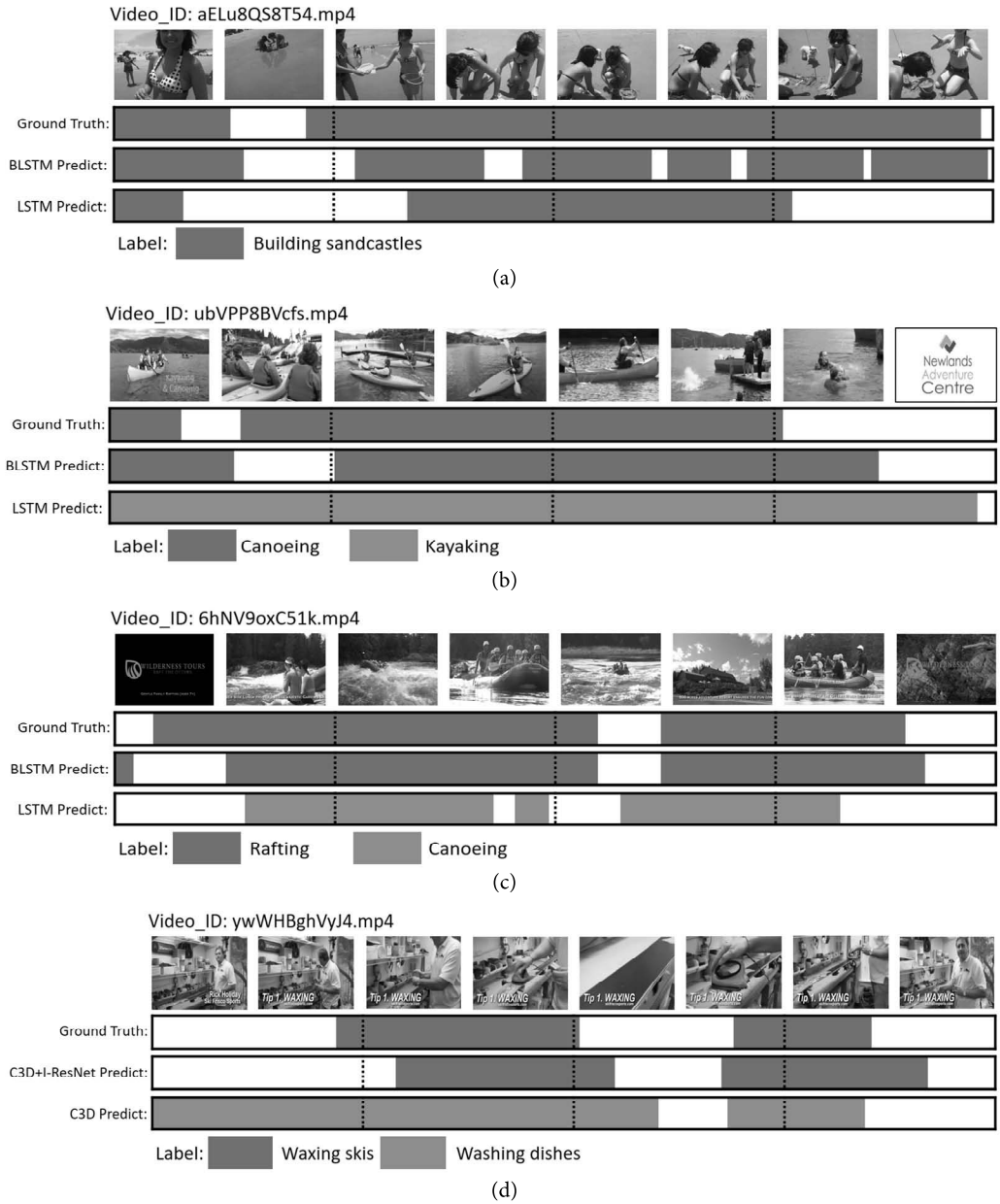

Fig. 7. Evaluation of localization performance. (a)~(c) three examples of localization results with two different classification models, LSTM and BLSTM, and (d) one example of localization results with two different feature models, the C3D only and the C3D+I-ResNet.

In the final experiment, we qualitatively evaluated the localization performance of our DeepAct model, including both the hybrid feature model (C3D + I-ResNet) and the bi-directional LSTM classification model (BLSTM). Fig. 7 shows the results of the final experiment. In particular, Fig. 7(a)–(c) show the experimental results, which are derived from two different classification models, LSTM and BLSTM, but with the same C3D + I-ResNet feature model. Based on the results shown in Fig. 7(a)–(c), it is possible to confirm that the proposed BLSTM classification model is superior to the uni-directional LSTM model in terms of activity localization performance. On the other hand, Fig. 7(d) shows the experimental results, which are conducted from two different feature models, the C3D only and the C3D + I-ResNet, but with the same BLSTM classification model. From the results shown in Fig. 7(d), we are sure that the proposed C3D + I-ResNet feature model is a much better choice in terms of activity localization performance.

5. Conclusions

In this paper, we presented DeepAct, a novel deep neural network model for detecting human activity in long untrimmed videos. In particular, we proposed the C3D + I-ResNet integrated feature model to extract simultaneously both the spatiotemporal activity features and the associated object features from each video segment. The integration of these supplementing features helps to identify daily human activities more precisely in videos. We also provide the BLSTM model for activity classification. For deciding the current output from the sequence of video feature vectors, the BLSTM model is able to make use of both the previous and the future context. Due to this characteristic, the proposed BLSTM model can improve the performance of activity detection in videos.

Through experiments using ActivityNet 200, a large-scale video benchmark dataset, we showed the high performance of the proposed DeepAct model for detecting human activity in untrimmed videos. In this work, we employed somewhat a simple threshold-based activity localization method. As a future work, we plan to design a more sophisticated method for improving the performance of activity localization in videos.

Acknowledgement

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center support program (No. IITP-2017-0-01642) supervised by the Institute for Information & Communication Technology Promotion (IITP).

References

[1] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in *Proceedings of IEEE International Conference on Computer Vision (ICCV-13),* Sydney, Australia, 2013, pp. 3551–3558.

[2] L. Wang, Y. Qiao, and X. Tang, “Video action detection with relational dynamic-poselets,” in *Proceedings of European Conference on Computer Vision (ECCV-14)*, Zurich, Switzerland, 2014, pp. 565–580.

[3] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[4] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: a large-scale video benchmark for human activity understanding," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR-15)*, Boston, MA, 2015, pp. 961–970.

[5] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS-14)*, Montreal, Canada, 2014, pp. 568–576.

[7] J. Zheng, Z. Jiang, and R. Chellappa, "Cross-view action recognition via transferable dictionary learning," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 2542–2556, 2016.

[8] R. Wang and D. Tao, "UTS at ActivityNet 2016," in *ActivityNet Large Scale Activity Recognition Challenge Workshop*, Las Vegas, NV, 2016, pp. 1–6.

[9] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S. F. Chang, "CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 5734–5743.

[10] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," 2017 [Online]. Available: https://arxiv.org/abs/1703.02716.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: towards good practices for deep action recognition," in *Proceedings of European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 20–36.

[12] G. Singh and F. Cuzzolin, "Untrimmed video classification for activity detection: submission to ActivityNet challenge," 2016 [Online]. Available: https://arxiv.org/abs/1607.01979.

[13] S. Karaman, L. Seidenari, and A. D. Bimbo, "Fast saliency based pooling of fisher encoded dense trajectories," in *Proceedings of European Conference on Computer Vision (ECCV) Workshop*, Zurich, Switzerland, 2014, pp. 1–4.

[14] Z. Shou, D. Wang, and S. F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, 2016, pp. 1049–1058.

[15] L. Wang, Y. Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," in *Proceedings of European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, pp. 1–6.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4489–4497.

[17] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "DAPs: deep action proposals for action understanding," in *Proceedings of European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 768–784.

[18] A. Montes, A. Salvador, S. Pascual, and X. Giro-i-Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," 2017 [Online]. Available: https://arxiv.org/abs/1608.08128.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770–778.

Yeongtaek Song https://orcid.org/0000-0002-1439-574X

He received the B.Sc. degree in Computer Science from the Kyonggi University, Korea, in 2017. He is currently a M.Sc. student of Computer Science Department, Kyonggi University, Korea. His current research interests include machine learning, computer vision, and intelligent robotic systems.

Incheol Kim https://orcid.org/0000-0002-5754-133X

He received the M.Sc. and Ph.D. degrees in Computer Science from the Seoul National University, Korea, in 1987 and 1995, respectively. He is currently a Professor of Computer Science Department, Kyonggi University, Korea. His current research interests include machine learning, knowledge representation and reasoning, task and motion planning, computer vision, and intelligent robotic systems.