Smoking Action Recognition Based on Spatial-Temporal Convolutional Neural Networks

Chien-Fang Chiu *, Chien-Hao Kuo *, and Pao-Chi Chang *
* National Central University, Jhongli, Taiwan
E-mail: {cfchiu, chkuo, pcchang}.vaplab@gmail.com Tel:+886-3-4226971

Abstract— **In this work, we propose a system that can recognize smoking action. It utilizes data balancing and data augmentation based on GoogLeNet and Temporal segment networks architecture to achieve effective smoking action recognition. The experimental results show that the smoking accuracy rate can reach 100% for Hmdb51 test dataset. For additional irrelevant movie smoking clips, the accuracy can also be as high as 91.67%.**

I. INTRODUCTION

Cigarette smoking increases risk for death from all causes in men and women. The risk of dying from cigarette smoking has increased over the last 50 years worldwide [1]. If one stands next to a smoker, this person still can be infected, called passive smoking. Consequently, smoking is prohibited in many closed public areas such as government buildings, educational facilities, hospitals, enclosed sport facilities, and buses [1]. However, it still often happens that smokers smoke even in highly prohibited places such as hospitals and elementary school campuses.

The objective of this work is to develop a smoking action recognition system based on deep learning, which allows quick discovery of smoking behavior. It is especially useful in a video surveillance intensive used environment.

Deep learning technique has shown outstanding performance in computer vision. Particularly, Convolutional Neural Networks (CNNs) achieve excellent performance from learning useful representations for image classification [1, 2, 3, 4, 5, 6], object detection [7, 8], and video classification. For action recognition [9, 10, 11, 12, 13], CNNs obtain better results compared with traditional methods with hand crafted features [14].

There exist three types of architectures for video representations of action recognition: (1) two-stream CNNs [9], (2) 3D CNNs [11], and (3) 2D CNNs with temporal structure [10]. Two-stream CNNs combine appearance and motion information from RGB color images and optical flow images. They train separate network for spatial network and temporal network. Finally, fusing prediction scores from two-stream network. However, 2-stream CNNs need much time to train and crop optical flow. 3D CNNs directly use 3D convolution and 3D pooling training model from stacked RGB to get spatiotemporal features. However, at the moment, its performance is inferior to two-stream CNNs. 2D CNNs with temporal model can capture long-term temporal information. It satisfies our requirement in getting video temporal action.

II. RELATED WORK ON CNN IN VIDEO CLASSIFICATION

K. Simonyan *et al.* [9] were the first team who proposed two-stream CNNs on action recognition. It calculated the motion information from adjacent video frames to get the optical flow, and utilized multi-frame optical flow and single color image to train temporal CNN and spatial CNN. Imagenet was chosen as the pre-trained model. Both networks used AlexNet [3] as the training model. The results from two trained networks were directly averaged or fused by support vector machine. The fused results were used for classification.

L. Wang *et al.* [10] extended the two-stream CNNs architecture by cutting a video stream into K clips $\{S_1, S_2, \ldots, S_K\}$ (K =3 in this work). A temporal segment network (**TSN**), in which an RGB image $\{T_1, T_2, \ldots, T_K\}$ was chosen randomly from corresponding clip, is described in (1),

$$\begin{aligned}&\mathrm{TSN}(T_1, T_2, \ldots, T_K) = \\ &\mathcal{H}(\mathcal{G}(\mathcal{F}(T_1; \mathbf{W}), \mathcal{F}(T_2; \mathbf{W}), \ldots, \mathcal{F}(T_K; \mathbf{W})))\end{aligned} \quad (1)$$

where $\mathcal{F}(T_K; \mathbf{W})$ is the CNN function with parameters W, which is used to fabricate class scores, $\mathcal{G}$ represents segmental consensus function, which combines the outputs of class hypothesis, $\mathcal{H}$ is the prediction function, which predicts the probability of each class.

In the training stage, a number of RGB image were used equivalently to train K networks simultaneously. Before the final softmax layer, the trained features were processed by average pooling separately and then concatenated.

L. Wang *et al.* selected Inception with Batch Normalization (BN-Inception) [15] as the main structure owing to its balance between accuracy and efficiency. Architecture of 5x5 convolutional kernel is replaced by two 3x3 kernels in two layers. They used partial Batch Normalization (regularizing the mean and variance of batch normalization layer except first layer) and dropout after global pooling.

III. SYSTEM FRAMEWORK

The proposed system employs CNN architecture and deep learning method for smoke action recognition. Fig.1 depicts the architecture of the proposed system. The system contains three stages, including pre-processing stage, training stage, and testing stage.

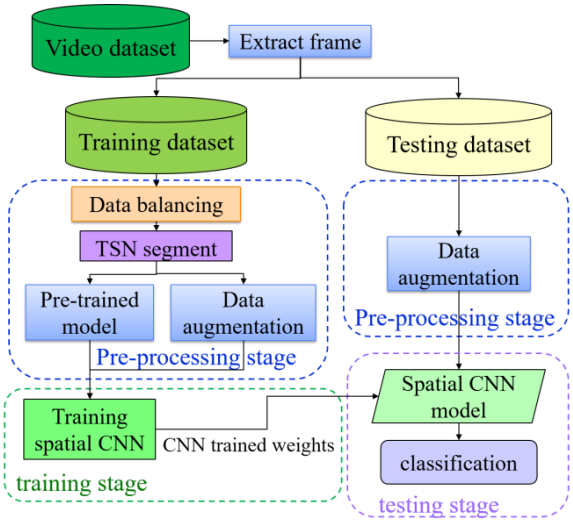

Fig. 1 System architecture of smoking action detection

A. *Pre-processing Stage*

We employ OpenCV [16] to extract frames from all video clips with the frame rate 24 frames/second and set frame size to 340 x 256. **Data balancing** technique is applied to the training data to make each class to have comparable amount of data. Most systems with good performance are analyzed based on randomly and uniformly distributed datasets. The size of the training set for each class is presumed to be equal. Otherwise, the neural network might be trained toward the large date set to minimize the training loss. In our system, HMDB51 dataset has 70 smoking cases vs. 3500 no smoking cases (all other 50 classes are counted as no smoking). The next section shows that the original unbalanced dataset results in low accuracy for smoking. To perform data balancing, only two cases from each non-smoking class are randomly chosen. Overall 100 cases are obtained as no smoking training set. In addition to the original 70 cases for smoking action training, extra 43 cases from daily life smoking action are added. This makes the training set richer in video characteristics and more effective for representing all kinds of smoking actions.

TSN is applied to segment the video clips. Data augmentation [4, 9, 10] is then employed to each frame by cutting four corners and the middle part, as well as horizontal flips, to obtain overall 10 new frames.

B. *Training Stage*

The training process is shown as in Fig. 2. We utilized TSN [10] architecture with the reference setting for smoke detection. We observe that temporal CNN exhibits little help in smoking action recognition, shown as in Table 1. This is probably due to slow or even still motion in smoking action. It leads to smoking optical flow trajectory is difficult to predict. Hence, only spatial CNN is utilized in the proposed system.

We also use the inception model with batch normalization [15] and set segment K =3. The pre-train model follows the TSN HMDB51 model, whose pre-train model is ImageNet. We divide it to 2classes training. Finally, features of three networks are reduced by average pooling and merged by concatenation.

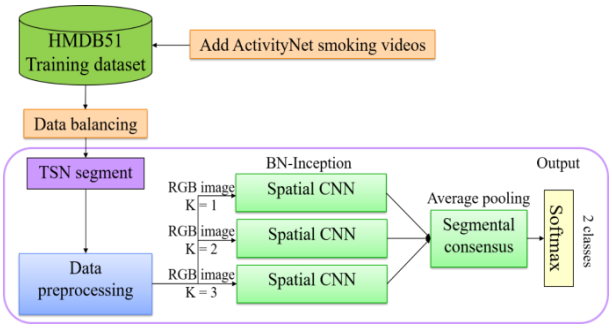

Fig. 2 Training process

C. *Testing Stage*

Following two stream CNNS [9], we sample 25 RGB frames from each video. Through the spatial CNN model, each test data can be recognized as either smoking action or no smoking action.

IV. Experimental Result

A. *Dataset and Implementation Details*

Three datasets, HMDB51 [17], smoking a cigarette of ActivityNet [18], and smoke of AVA dataset [19] were used to evaluate the performance. HMDB51 contains 6,766 clips of 51 action classes, each containing at least 101 clips. Following the evaluation scheme from THUMOS13 challenge [20], we took three training/testing splits for evaluation. Smoking a cigarette of ActivityNet contains 53 videos, each video lasts at least 14 seconds. Videos are captured smoking action in real life. Smoke of AVA dataset contains 32 movie clips we selected, each clip lasts 3 seconds long. In addition to smoking action, these clips also include other actions, such as multiple people talking, two people chatting, and one person moving package.

The parameter setting for the training is as follows, mini batch size: 32, learning rate: 0.001, iteration number: 2500. Following TSN, the cross-entropy loss is used for evaluating the training performance.

B. *Results on HMDB51 2classes*

At first, the dataset with 3570 clips is divided into two sets, smoking 70 clips and no smoking 3500 clips, and used for training. The test dataset includes 30 clips smoking and 1500 no smoking. Table 1 shows the smoking detection accuracy with different pre-trained models and different input networks. It shows that the benifit from temporal CNN is very limited. Even with different pre-train model, the accuracy can only be raised to 5.55%. Hence only the spatial CNN is used in this work. It also shows that the change of the pre-trained model from ImageNet to HMDB 51 can achieve better performance, increasing the accuracy from 17.77% to 33.33%. Hence Hmdb51 model is used in the following experiments.

Table 1 The accuracy rate based on different pre-trained model

	Smoking accuracy(%)			
	Temporal CNN		Spatial CNN	
Pre-trained model	**ImageNet**	**HMDB51**	**ImageNet**	**HMDB51**
split1	0	6.66	10.00	30.00
split2	0	0.00	26.66	30.00
split3	0	10.00	16.66	40.00
average	0	**5.55**	17.77	**33.33**

Table 2 shows the results by adding extra smoking datasets in training. It shows that the accuracy merely increases 1.11% by adding short time movie clips, AVA data. On the contrary, the accuracy can increase 18.34% by adding longer time daily life movies ActivityNet smoking. It is also better than the original 51-class training by 0.56%. However, adding both additional datasets for training can achieve 14.45% accuracy increase still lower than adding single data set ActivityNet smoking by 3.89%. It implies that the daily life movies include very important video characteristics that augment the training dataset. Hence, ActivityNet smoking is added to the additional training set.

Table 2 The accuracy rate of additional datasets

Spatial training	**Smoking accuracy (%)**				
	51 classes	**2 classes**			
		HMDB51	**+AVA data (32)**	**+Activity smoking (43)**	**+Activity smoking +AVA data**
split1	50.00	30.00	33.33	52.50	40.00
split2	60.00	30.00	36.67	55.00	53.33
split3	43.33	40.00	33.33	47.50	50.00
average	51.11	33.33	34.44	**51.67**	47.78

All Hmdb51 contents are human action related. Many of these actions are very similar and easy to be confused. One of the problems in the previous experiment is the unbalance of the training data in each class. This will train the system toward the major class, which is no smoking class, and pay less attention to the minor class, which is the smoking class. To solve this problem, the training set of no smoking class is reduced to 100 clips while the smoking class maintains the original 70 clips plus ActivityNet smoking 43 clips. The ratio of the smoking to no smoking is 100: 113, which is more balanced. The test set is the original Hmdb51 test set, which includes 30 smoking and 1500 no smoking clips.

The smoking detection results include True Positive (TP), smoking and is detected, True Negative (TN), no smoking and is not detected, False Positive (FP), no smoking but is detected as smoking, and False Negative (FN), smoking but detected as no smoking. Table 3 shows that the accuracy of smoking for ActivityNet smoking dataset after data balancing can be as high as 100%. The other class, which has only 1/35 training data, can still reach 71.18% accuracy. This shows how important the data balancing is. The average accuracy is depicted in Fig. 3, the confusion matrix.

Table 3 The detection results with data balancing

Hmdb51 data	FN	TP	TN	FP
	(%)			
split1	0	100	70.13	29.87
split2	0	100	72.80	27.2
split3	0	100	70.60	29.4
average	0	**100**	71.18	28.82

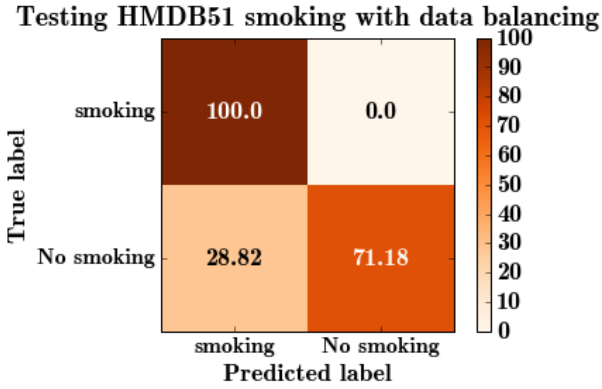

Fig. 3 Confusion matrix of HMDB51 2classes

C. *Results on additional test dataset*

In the above experiments, though the accuracy for smoking detection is extremely high, the training dataset is relatively small. This might raise a concern whether the accuracy can still be kept high for totally different test sets. Consequently, another dataset, AVA dataset including 32 clips, is used to test the model that has been trained. From Table 4, the accuracy for smoking detection can still be as high as 91.67%. This shows the resulting model can effectively detect the smoking action in all sorts of situations, including multi-person talking, other persons in the same frame performing no smoking action.

Table 4 Experimental results on AVA dataset

AVA data	FN	TP
	(%)	
split1	3.125	96.875
split2	9.375	90.625
split3	12.500	87.500
average	8.333	**91.667**

V. CONCLUSIONS

We have proposed a system that is specially designed for cigarette smoking action recognition based on deep learning technique. To compensate for the limited dataset of smoking actions, data balancing and data augmentation (adding ActivityNet smoking data) are shown to be extremely important to achieve high accuracy rate. In our experiment, spatial CNN is more powerful than temporal CNN in smoking action. As results, the proposed system can achieve 100% accuracy for HMDB51 smoking dataset, and 91.67% for irrelevant multi-class video clips.

REFERENCES

[1] U.S. Department of Health and Human Services.The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014 [accessed 2017 Apr 20].

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE 86.11, pp. 2278-2324, 1998.

[3] K. Alex, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, pp. 1097-1105, 2012.

[4] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations (ICLR), 2015.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, 2015.

[6] K. He, Zhang, X., S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[7] R. Girshick, J. Donahue, T Darrell., and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems, pp. 91-99, 2015.

[9] K. Simonyan, and A. Zisserman. "Two-stream convolutional networks for action recognition in videos," Advances in neural information processing systems, pp. 568-576, 2014.

[10] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in European Conference on Computer Vision, pp. 20-36, 2016.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE 86.11, pp. 2278-2324, 1998.networks," Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE, pp. 4489-4497, 2015.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732, 2014.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[14] H. Wang, and C. Schmid, "Action recognition with improved trajectories," In: Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, pp. 3551-3558, 2013.

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

[16] OpenCV: Open Source Computer Vision Library , https://opencv.org/

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," In 2011 International Conference on Computer Vision, pp. 2556–2563, 2011.

[18] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in Computer Vision and Pattern Recognition (CVPR), pp. 961-970, 2015.

[19] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, ... and C. Schmid, "AVA: A video dataset of spatio-temporally localized atomic visual actions," arXiv preprint arXiv: 1705.08421, 2017.

[20] Y. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes."