Video based human smoking event detection method

Anna V. Pyataeva[1,2], Maria S. Eliseeva[1]

[1]Siberian Federal University, Krasnoyarsk, Russia
[2]Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

Abstract

The paper proposes a method for recognizing smoking event detection from visual data. The method uses a three-dimensional convolutional neural network ResNet, which provides work with video based spatio-temporal features.

Keywords

Smoking event detection, convolutional neural network, spatio-temporal features.

1. Introduction

According to WHO Framework Convention on Tobacco Control [1] there is no safe level of tobacco smoke exposure. Creating a completely smoke-free environment is the only way to protect people from the harmful effects of breathing even second-hand smoke. Human action analysis based on visual processing is significant for many applications such as intelligent video surveillance, analysis of employee and customer behavior. Recognizing a person's smoking while driving can significantly increase road safety [2]. To recognize smoking activity on the use of smartwatch sensors as a state-transition model that consists of the mini-gestures hand-to-lip, hand-on-lip, and hand-off-lip [3]. Wu et al. [4] proposed the color-based ratio histogram analysis is introduced to extract the visual clues from appearance interactions between lighted cigarette and its human holder. The techniques of color re-projection and Gaussian Mixture Models enable the tasks of cigarette segmentation and tracking over the background pixels. Smoke detection in the area around human faces and hands can be applied to recognition of the smoking action [5, 6, 7]. The reliable smoke detection is a difficult due to great variability of shape, color, transparency, turbulence variance, non-stable motion, boundary roughness, and time-varying flicker effect in the boundaries of smoke as well as artifacts during shooting such as low resolution, blurring, and weather conditions. The key problem of smoking behavior recognition is the irregular shape: different ways to hold a cigarette, types of tobacco products, bad weather and shooting conditions.

2. Smoking event detection method

In this paper spatio-temporal features based smoking activity detection algorithm, which allows recognizing human smoking activity regardless of the person's appearance, the way to hold a cigarette, the type of cigarette, the distance of the object of interest, and movement patterns.

SDM-2021: All-Russian conference, August 24–27, 2021, Novosibirsk, Russia
anna4u@list.ru (A. V. Pyataeva)

2.1. Spatio-temporal features of smoking activity

Smoking activity belongs to a group of atomic actions that can be recognized only if there is a certain set of spatio-temporal features. Four atomic action groups are considered:

- arm position changes. The sequence of actions: the hand rises to the level of the lips, pause, falls down, pause, rises again;
- lip movement on close-up scenes;
- lighting a cigarette:
  - tilt of the head;
  - using a cigarette lighter, using a lighter involves a sequence of actions:
    - bringing the lighter to the face with one hand;
    - the thumb of this hand starts the mechanism (the action can be repeated several times);
    - the other hand can prevent the cigarette from fading and block the view to recognize previous actions (in this case, both hands are at the level of the lips);
    - the hands are lowered;
  - lighting up with matches, the use of matches for lighting cigarettes consists of the following actions:
    - the cigarette is clamped between the teeth;
    - both hands are at chest level or just below the chest;
    - one hand is performed with a small wave (the action can be repeated);
    - one hand remains at chest level, the second changes position, moving higher to the chin or lips,
    - a wave of the hand (to extinguish the match);
    - lowering the hands;
- flicking the ash from the cigarette (the action may not be present in the frame) consists of the following steps: the withdrawal of the hand with the cigarette down and the characteristic movement of the hand or fingers of the hand.

Smoking activity recognition is implemented using a three-dimensional neural network based on the spatio-temporal features in the entire video data.

2.2. Image pre-processing

Visual information as a result of real-time video shooting may include objects with dynamic behavior, noise of the hardware or transmission lines, as well as artefacts affected by weather conditions (for example, rain or snow, poor luminance in the morning or evening). Because of this, the quality of smoking action recognition significantly degrade. Therefore, scaling and mean subtraction [8] are used to solve this problem. To implement preprocessing algorithms, a computer vision library OpenCV (Open Source Computer Vision Library) was used [9]. Thus, the video sequence preprocessing is performed according to the expressions:

$$R = \frac{R - \mu_R}{\sigma}, \quad G = \frac{G - \mu_G}{\sigma}, \quad B = \frac{B - \mu_B}{\sigma}, \tag{1}$$

where R, G, B are the values of the red, green, blue channels of the image, respectively; $\mu = \{\mu_R, \mu_G, \mu_B\}$ is the average color intensity for each image channel; $\sigma -$ scaling coefficient. The σ value can be the standard deviation over the training set. However, σ can also be manually set to scale the input image space to a specific range

2.3. Neural network architecture

AlexNet [10], VGG [11] and ResNet [9] neural networks are most often used to classify images and video sequences. The ResNet neural network is fully convolutional, so it is used for space-time volume extraction, unlike many architectures with fully connected layers, including AlexNet and VGG-16, which contain several levels of the maximum pool that can damage the actions evaluation. The ResNet network contains only one pool level immediately after the conv1 layer. The reduced number of bonding layers makes ResNet more suitable for visual recognition of smoking, since spatial details must be preserved to recognize this process.

In the work the 34 layers ResNet neural network was used that shows computational efficiency in solving classification problems [12]. In order to use ResNet to estimate multi-frame optical flow, it is necessary to extend this architecture, replacing all $k \times k$ two-dimensional convolutional kernels with an additional time dimension $k \times k \times 3$, as described in article [13]. The pool layers in the decoder are expanded in a similar way. The neural network transformed in this way in the paper is called ResNetM, its composition is presented in Table 1.

In Table 1 the residual blocks are grouped in square brackets. Batch normalization is used after each convolutional layer. The main difference between this architecture and ResNet is the use of 3D kernels and a modified downsampling operation, whereby feature maps in the convolution layer are combined with several adjacent frames in the previous layer, thereby capturing motion information.

The dimensions of the convolutional kernels are $3 \times 3 \times 3$. The network uses 16-frame RGB clips as inputs. The dimensions of the input clips are $3 \times 16 \times 112 \times 112$. Downsampling of inputs is performed periodically in steps of 2.

Table 1
Architecture of the ResNetM neural network.

Layer name	Activation function	Core	Neuron count
Convolutional layer 1	ReLu	$7 \times 7 \times 7$	64
Convolutional layer 2		$\begin{bmatrix} 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix} \times 3$	64
Convolutional layer 3		$\begin{bmatrix} 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix} \times 4$	128
Convolutional layer 4		$\begin{bmatrix} 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix} \times 6$	256
Convolutional layer 5		$\begin{bmatrix} 3 \times 3 \times 3 \\ 3 \times 3 \times 3 \end{bmatrix} \times 3$	512

2.4. Smoking activity detection algorithm

The proposed method uses deep learning network for smoking action detection by recognizing actions that are characteristic of a person who is in the process of smoking. The block diagram of the smoking activity detection algorithm is shown in Figure 1.

Stochastic Gradient Descent (SGD) with momentum is used to train the neural network. Training samples are randomly generated from the videos in the training set. Time positions are selected evenly. Next, 32-frame clips are set around the specified time positions. If the video is shorter than 32 frames, it will loop as many times as necessary to reach the set duration. Then the spatial positions are randomly selected from four corners or one center. In addition to the positions, the spatial scales of each sample are also specified for multi-scale cropping. The frame is cropped at the time-space positions. The size of each sample is 3 channels×32 frames×112 pixels×112 pixels, and each sample is flipped horizontally with a 1/2

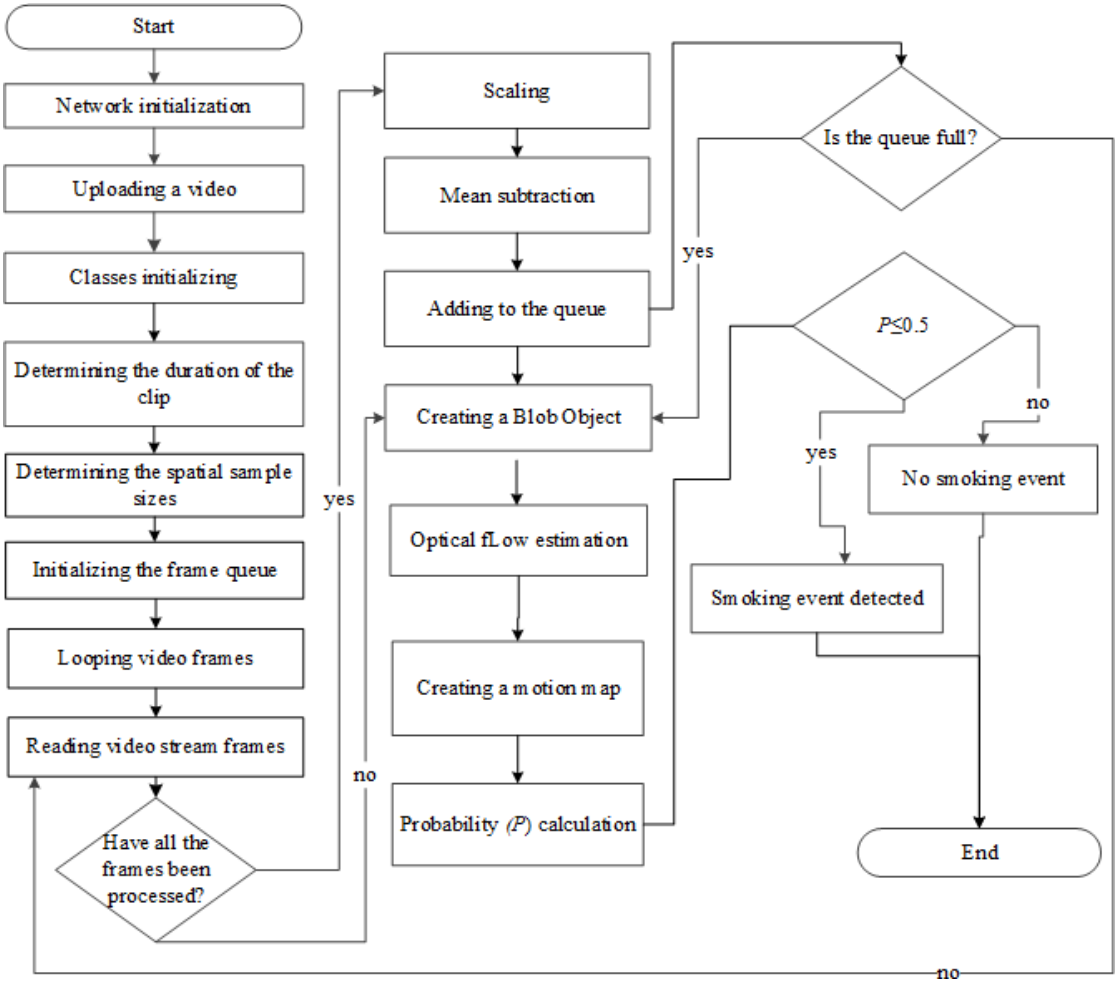


Figure 1: Smoking event detection algorithm.

probability. It also subtracts the average of our dataset from the sample for each color channel. All created samples retain the same class labels as their original videos. Model training uses cross entropy as a function of loss. The training parameters include a damping of 0.001 and 0.9 for the impulse. The learning rate is 0.1 and divided by 10 after saturation of the validation loss. When fine tuning is performed at a learning rate of 0.001, the scale attenuation is 1e−5.

At the first stage, the neural network is initialized, the parameters are set, and after that the video sequence is fed to the input. Initialization of the classes is performed, which allows the classification of the dataset: "smoking", "no smoking". The duration of the sample is determined, that is, the number of frames for classification is 32, and the spatial sizes of the sample are 112×112. To create input clips, the sliding window method is used, in which only the oldest frame in the list is discarded, making room for the newest frame. Each video is then split into non-overlapping 32-frame clips. This operation occurs using a loop that reads frames from the video stream, then checks for frame capture. If a frame is captured, then each clip is cropped around the center position at the maximum scale, an average subtraction is performed and a new frame is added to the queue, otherwise the loop exits. The new cycle allows you to check if the queue is full. At the end of this cycle, a blob object is created. A "blob object" or "blob" is a collection of frames with the same spatial dimensions, expressed in width and height, and the same depth, that is, the number of channels that must be preprocessed in the same way. A blob object has the following dimensions: (3, 32, 112, 112). The number 3 denotes the number of channels in the input frames. 32 — the total number of frames in the "blob". The following numbers represent the height and width respectively.

Next, in order to extract the space-time characteristics, each instance is transmitted through a 3D convolutional neural network. Smoking is recognized by finding multiple optical flow. The optical flow is calculated at each point, then a motion map is formed. Each feature map of a convolutional layer is associated with several consecutive adjacent frames in the upper layer. The next step is to assess the probability of smoking in the clips. The network "scans" the sequence of thirty-two frames, generates motion paths, analyzes the similarity to a known smoking pattern, and finds the probabilities of smoking in each frame, which are then averaged over all clips. The class that has the highest score indicates the action in the given video sequence. If the probability is greater than or equal to 0.5, then smoking in these frames is recognized.

3. Experimental and results

In order to the video-based smoking detection model work, the following specifications are required: a minimum of 2GB NVIDIA graphics card and installed software: CUDA and cuDNN. The model uses Anaconda and Python packages including OpenCV, matplotlib, and Pytorch. Experimental studies were carried out with the characteristics of a laptop Intel (R) Core (TM) i7-6700HQ processor, 2.60 GHz processor clock, 8 GB RAM, Windows 10 operating system, NVIDIA GeForce GTX 960M graphics processor, 2 GB dedicated graphics processor memory. The modified neural network was trained on 6766 videos from the HMDB51 dataset [14]. The video shows actions that can be grouped into five groups:

(1) general face actions: smile, laugh, chew, speak;

(2) actions with object manipulations: smoking, eating, drinking;
(3) general body movements: do a wheel, applaud, climb, climb stairs, dive, fall to the floor, put your hands back, do a handstand, jump, pull up, push up, run, sit down, climb from something, do somersaults , get up, turn around, walk, make a wave;
(4) body movements when interacting with an object: combing hair, catching, drawing a sword, dribbling a ball, playing golf, hitting a ball, picking, pouring, pushing something, riding a bicycle, riding a horse, shooting a ball, shooting bow, shoot a gun, throw a ball;
(5) body movements for human interaction: fencing, hugging, kicking, kissing, punching, shaking hands, sword fighting.

Actions of categories (1)–(5) for experimental research are combined into one class "no smoking". For experimental studies, 70 "smoking" videos were used, in which people of different ages, body types, gender characteristics, different races, differently holding cigarettes, of different shapes and types, were filmed in the process of smoking. and 6766 video with "no smoking" actions. At least two observers to ensure consistency have reviewed each clip. The algorithm results are shown in Table 2.

Tables 3 and 4 shows the frames of some of the video sequences used and the results of smoking recognition. The results of the smoking recognition method are marked with the labels "smoking" — "no smoking".

The test video data is supplemented with videos in which the action is visually similar to smoking, but, thanks to the spatial and temporal features of the neural network and the identified pattern of characteristic smoking movements, it is able to distinguish these actions from smoking action. In video 9 a girl eats a lolipop; video 11 a girl bites a pen; video 15 a man eats ice cream. The training sample was 80%, the test sample was 20% of the total sample. To evaluate the effectiveness of human smoking activity detection and recognition algorithms, the indicators of detection accuracy (TR), false-positive (FAR) and false-negative (FRR) were used. The results of smoking detection for neural network architectures ResNet and modified network ResNetM are shown in Table 5.

Table 2
The algorithm results.

Era	Training loss	Accuracy when training	Test losses	Accuracy when checking
1	1.1552	0.4329	0.7308	0.6699
2	0.9412	0.5801	0.5987	0.7346
3	0.8054	0.6504	0.5181	0.7613
4	0.7215	0.6966	0.4497	0.7984
5	0.6253	0.7572	0.4530	0.7984
		...		
46	0.2325	0.9167	0.2024	0.9198
47	0.2284	0.9212	0.2058	0.9280
48	0.2261	0.9212	0.2448	0.9095
49	0.2170	0.9153	0.2259	0.9280
50	0.2109	0.9118	0.2267	0.9125

Table 3
Description and results of some used videos.

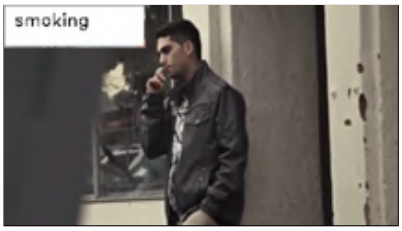
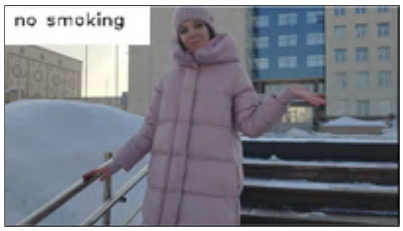
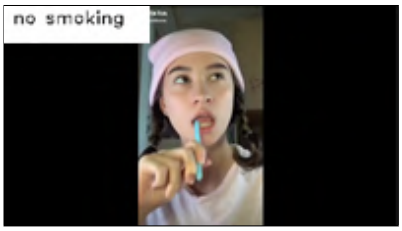
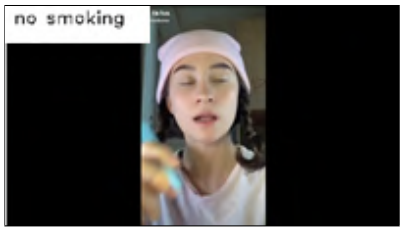
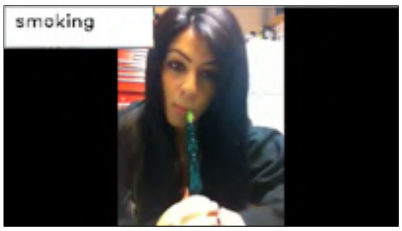
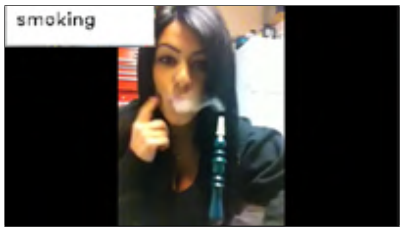
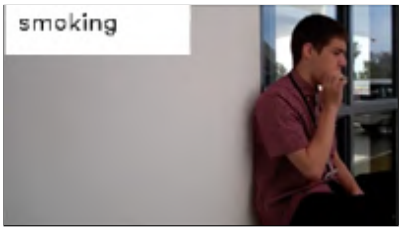
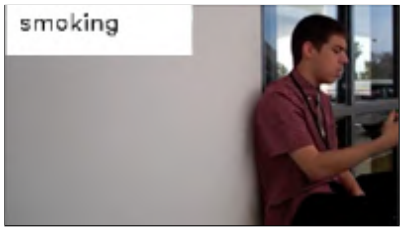
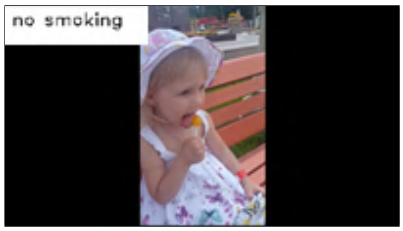
Description of test video	Sample frame 1	Sample frame 2
Alias: Video 1. Number of frames: 107. Resolution: 1280×720. Video duration: 4.47 sec.		
Alias: Video 4. Number of frames: 78. Resolution: 1920×1080. Video duration: 2.63 sec.		
Alias: Video 5. Number of frames: 198. Resolution: 1280×720. Video duration: 7.93 sec.		
Alias: Video 11. Number of frames: 107. Resolution: 1280×720. Video duration: 4.47 sec.		
Alias: Video 12. Number of frames: 117. Resolution: 270×360. Video duration: 3.90 sec.		
Alias: Video 15. Number of frames: 151. Resolution: 1280×720. Video duration: 5.07 sec.		

Table 4
Description and results of some used videos (continued).

Description of test video	Sample frame 1	Sample frame 2
Alias: Video 20. Number of frames: 237. Resolution: 640×360. Video duration: 7.93 sec.		
Alias: Video 9. Number of frames: 44. Resolution: 406×720. Video duration: 4.10 sec.		
Alias: Video 10. Number of frames: 128. Resolution: 1920×1088. Video duration: 4.30 sec.		
Alias: Video 19. Number of frames: 87. Resolution: 480×360. Video duration:2.93 sec.		
Alias: Video 18. Number of frames: 81. Resolution: 1280×720. Video duration: 2.73 sec.		

Experimental studies conducted on 20 video sequences obtained in real-world shooting conditions confirm the efficiency of the proposed method for recognizing smoking. The ResNet neural network architecture, modified to a three-dimensional neural network, ensures that the spatial-temporal signs of smoking are taken into account and shows, on average, 15% higher

Table 5
Experimental results.

Video	ResNet			ResNetM		
	TR, %	FAR, %	FRR, %	TR, %	FAR, %	FRR, %
Video 1	80.2	19.8	32.1	87.8	12.0	12.2
Video 3	81.5	18.5	15.6	90.7	9.20	9.32
Video 5	78.0	22.0	31.2	88.8	11.1	11.2
Video 7	86.1	13.9	12.9	97.4	2.41	2.59
Video 8	80.9	19.1	17.9	92.4	7.52	7.57
Video 10	78.7	21.3	20.1	85.9	14.5	14.1
Video 12	90.0	10.0	11.1	98.2	1.74	1.81
Video 14	96.0	4.00	15.0	100.0	0.0	0.0
Video 16	84.5	15.5	16.1	95.4	4.51	4.61
Video 18	81.2	18.8	14.9	92.5	7.42	7.49
Video 20	80.9	19.1	25.7	84.3	1.53	15.7

accuracy in recognizing the smoking actions compared to the basic architecture. The developed software implementation of the smoking recognition method provides real-time operation.

References

[1] WHO framework convention on tobacco control. Available at: https://www.who.int/fctc/text_download/en.

[2] Chien T.C., Lin C.C., Fan C.P. Deep learning based driver smoking behavior detection for driving safety // Journal of Image and Graphics. 2020. Vol. 8. No. 1. P. 15–20.

[3] Odhiambo C., Cole C.A., Torkjazi A., Valafar H. State transition modeling of the smoking behavior using LSTM recurrent neural networks // 2019 International Conference on Computational Science and Computational Intelligence (CSCI). 2019. P. 898–904.

[4] Wu P., Hsieh J., Cheng J., Cheng S., Tseng S. Human smoking event detection using visual interaction clues // 20th International Conference on Pattern Recognition. 2010. P. 4344–4347.

[5] Dunne É. Smoking detection in video footage. A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of MAI (Computer Engineering). Submitted to the University of Dublin, Trinity College, 2018. 43 P.

[6] Iwamoto K., Inoue H., Matsubara T., Tanaka T. Cigarette smoke detection from captured image sequences // Image Processing: Machine Vision Applications III. International Society for Optics and Photonics. 2010. Vol. 7538. P. 82–87.

[7] Iwamoto K., Inoue H., Matsubara T., Tanaka T. Cigarette smoke detection using feature values based on the kernel LMS algorithm // IEICE Technical Report. Circuits and Systems. 2010. Vol. 109(434). P. 237–248.

[8] Varma V.S., Sasidharan K.P., Ramachandran K.I., Nair B. Real time detection of speed hump/bump and distance estimation with deep learning using GPU and ZED stereo camera // Procedia Computer Science. 2018. Vol. 143. P. 988–997.

[9] OpenCV (Open source computer vision library). Available at: https://opencv.org.
[10] Krizhevsky A., Sutskever I., Hinton G.E. ImageNet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems. 2012. P. 1097–1105.
[11] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // CoRR. 2014. abs/1409.1556.
[12] Ji S., Xu W., Yang M., Yu K. 3D convolutional neural networks for human action recognition // IEEE Transactions on Pattern Analysis & Machine Intelligence. 2013. Vol. 35. No. 1. P. 221–231.
[13] Yu G., Li T. Recognition of human continuous action with 3D CNN // International Conference on Computer Vision Systems. Springer, 2017. P. 314–322.
[14] HMDB a large human motion database. Available at: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database.