Provisional chapter

Real-Time Action Recognition Using Multi-level Action Descriptor and DNN

Cheng-Bin Jin, Trung Dung Do, Mingjie Liu and
Hakil Kim

Additional information is available at the end of the chapter

Abstract

This work presents a novel approach to the problem of real-time human action recognition in intelligent video surveillance. For more efficient and precise labeling of an action, this work proposes a multilevel action descriptor, which delivers complete information of human actions. The action descriptor consists of three levels: posture, locomotion, and gesture level; each of which corresponds to a different group of subactions describing a single human action, for example, smoking while walking. The proposed action recognition method is able to localize and recognize simultaneously the actions of multiple individuals using appearance-based temporal features with multiple convolutional neural networks (CNN). Although appearance cues have been successfully exploited for visual recognition problems, appearance, motion history, and their combined cues with multi-CNNs have not yet been explored. Additionally, the first systematic estimation of several hyperparameters for shape and motion history cues is investigated. The proposed approach achieves a mean average precision (mAP) of 73.2% in the frame-based evaluation over the newly collected large-scale ICVL video dataset. The action recognition model can run at around 25 frames per second, which is suitable for real-time surveillance applications.

Keywords: multilevel action descriptor, action recognition, video surveillance, deep neural networks

1. Introduction

Visual action recognition—the detection and classification of spatiotemporal patterns of human motion from videos—is a challenging task, which finds applications in a variety of domains including intelligent surveillance system [1], pedestrian intention recognition for advanced driver assistance system (ADAS) [2], and video-guided human behavior research [3]. For delivering complete description about human actions, this work proposes a multilevel action descriptor (**Figure 1**) to solve the existing representation problem of an action. For instance, traditional methods give the action representation of *phoning* for one person who is *phoning while running* and the same action descriptor for another person who is *phoning while sitting*. The action semantics for these two cases should be substantially different. The first difference is posture: one person is *standing*, and the other is *sitting*. The second difference is locomotion: one person is *running*, and the other is *stationary*. The proposed multilevel action descriptor consists of three levels: posture, locomotion, and gesture, which describe different categories of human subactions in a single action to address the above problem. Each level of subaction can be recognized by a corresponding convolutional neural network (CNN)-based classifier, which captures different appearance-based temporal features to represent a human subaction.

Most of the existing works [4, 5] have focused on video-based action recognition ("*Is there a certain action in the video?*") trying to classify the video clip as a whole via globally pooled features. This global feature pooling method works well, however, fails to consider the difference in the actions of multiple individuals that are present at the same time. For instance, one person in the video is *texting* and, besides him, another person is *smoking*. In our work, the problem of action detection in video surveillance is addressed as: "*is there a certain action in the video, and where is it spatially and temporally?*" The rationale behind the action detection strategy is partly inspired by the technique used in a recent paper [6], where the regions of

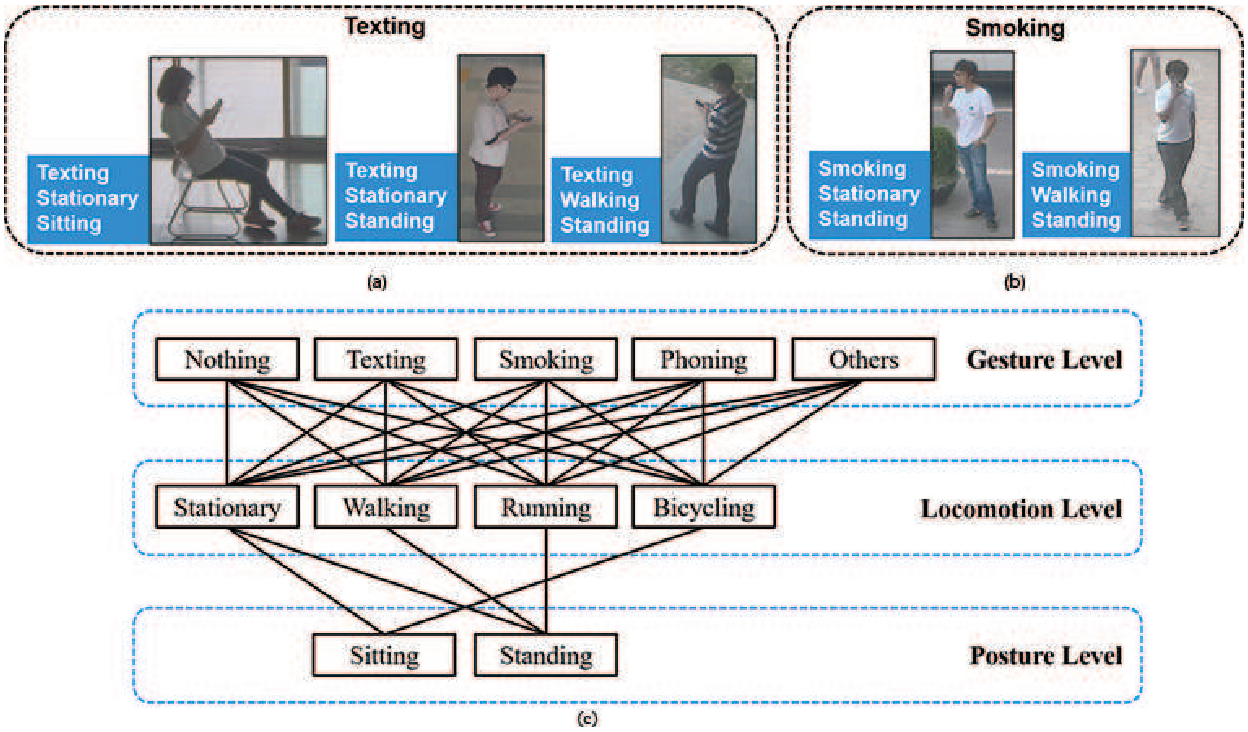

Figure 1. Conventional action representation and multilevel action descriptor: (a) *texting* for three different cases, (b) *smoking* for two different cases, and (c) structure of the multilevel action descriptor.

action are located and then classified to improve the representational power and classification accuracy.

This work aims to develop a real-time action recognition system with localizing and recognizing actions for multiple persons at the same time. Many works have been studied to estimate human pose [7–10] and analyze motion information [11] in real time. However, to the best of our knowledge, the real-time multilevel action descriptor was first introduced by the authors in [12] and this work is the extended version by adding two new actions, *bicycling* and *phoning*, and the evaluation of the processing time.

Figure 2 shows the overall scheme of the proposed real-time action recognition model. Through background modeling, motion-detection, human-detection, and multiple-object tracking, the appearance-based temporal features of the regions of interest (ROIs) are fed into the three CNNs, which make predictions using the shape, the motion history, and their combined cues. In the training phase, the ROIs and the multilevel action annotations are acquired manually in each frame of the training videos, and three appearance-based temporal features, namely—binary difference image (BDI), motion history image (MHI), and weighted average image (WAI)—are computed from the ROIs. Every level of the subaction has its own CNN classifier denoted as PostureNet, LocomotionNet, and GestureNet, respectively.

In the testing phase, the prediction of each CNN in the multi-CNN model corresponds to the decision in one subaction level. A motion saliency region is generated using a Gaussian mixture model (GMM) to eliminate regions that are not likely to contain the motion. This

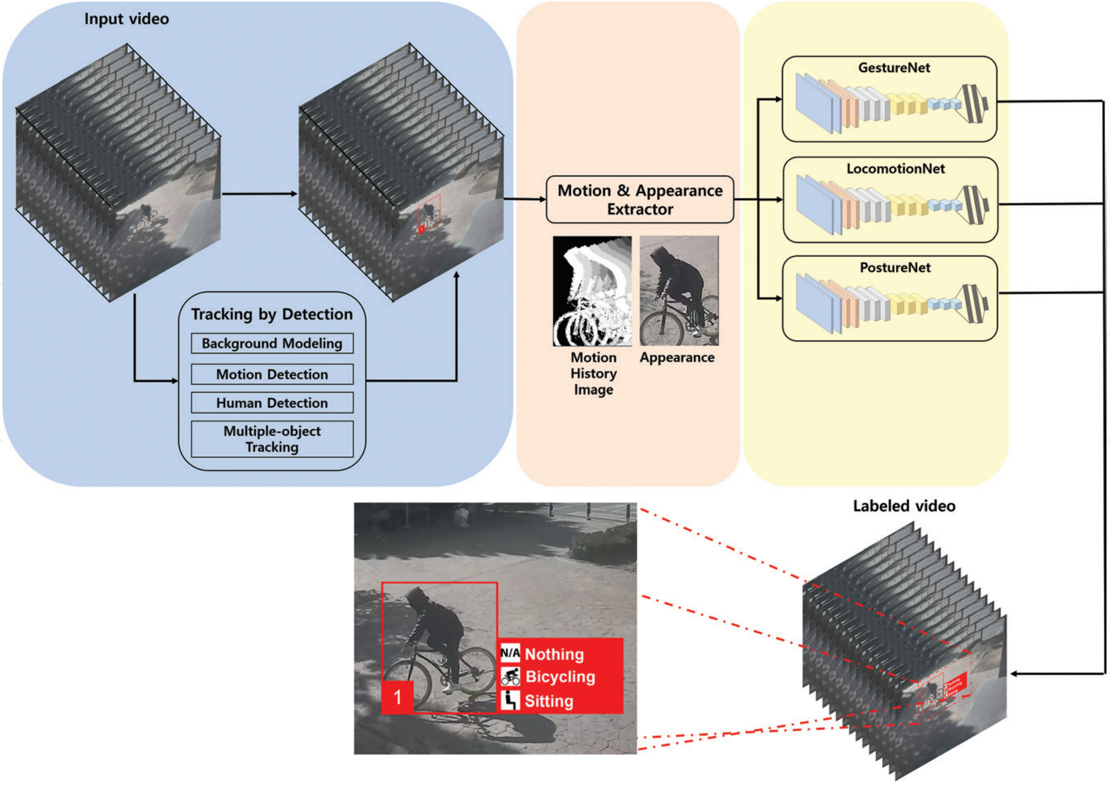

Figure 2. Overall process of the proposed real-time multilevel action recognition model.

leads to a big reduction in the number of regions to be processed. The conventional sliding window-based scheme is used on the motion saliency region as a mask. In the sliding window, a human-detection histogram of oriented gradient (HOG) descriptor [13] with a latent support vector machine (SVM) [14] is used to detect an initial human action in the ROIs. Then, the regions undergo Kalman filtering-based refinement of the locations in the image plane. Given the refined action in the ROI, the shape, the motion history, and their combined cues are used with the aid of the CNNs to predict three subaction categories. Finally, the postprocessing stage checks for any conflicts in the structure of the subaction descriptor and applies temporal smoothing according to the previous action history of each individual for the purpose of noise reduction.

The main contributions of this work can be summarized as follows:

- The multilevel action descriptor is presented for the real-time action recognition. The multilevel action descriptor consists of three levels. The combination of subaction from three levels can describe many different types of actions precisely. Furthermore, new subactions or action-levels can be easily incorporated into the multilevel action descriptor.
- A real-time action recognition model is developed on the basis of appearance-based temporal features with a multi-CNN classifier. Presented in this study is a model for action recognition that simultaneously localizes and recognizes multiple actions of individuals with both low computational cost and high accuracy.

2. Related works

Motion energy image (MEI) and motion history image (MHI) [15, 16] are the most pervasive appearance-based temporal features. The advantage of these methods is that they are simple, fast, and efficient in controlled environments, for instance, when the background of the surveillance video (from a top-view camera) is always static. The fatal flaw in MHI is that it cannot capture interior motions—it can only capture human shapes [12]. In our work, a novel method for encoding these temporal features is proposed, and a study of how many appearance-based temporal features affect performance is provided. Other appearance-based temporal methods are the active shape model, the learned dynamic prior model, and the motion prior model. In addition, the motion is consistent and easily characterized by a definite space-time trajectory in some feature spaces. Based on visual tracking, some approaches use motion trajectories (e.g., generic and parametric optical flow) of predefined human regions or body interest points to recognize actions [17, 18].

Over the past few years, local spatiotemporal feature-based algorithms are the most popular ones for recognizing human actions. Laptev [19] proposed space-time interest point (STIP) by extending the 2D Harris corner to a 3D spatiotemporal domain. Kim et al. [20] introduced a multiway feature pooling approach that uses unsupervised clustering of segment-level HoG3D [21] features. Li et al. [22] extracted spatiotemporal features that are a subset of improved dense trajectory (IDT) features [5, 23], namely, histogram of flow (HoF), motion boundary histogram (MBH), MBH$_x$, and MBH$_y$, by removing camera motion to recognize egocentric actions. However, the disadvantage of the local spatiotemporal algorithms is that it is computationally expensive.

Some alternative methods for action recognition have been proposed. Vahdat et al. [23] developed a temporal model consisting of key poses for recognizing higher level activities. Lan et al. [24] introduced a structure for a latent variable framework that encodes contextual information. Jiang et al. [6] proposed a unified tree-based framework for action localization and recognition based on an HoF descriptor and a defined initial action segmentation mask. Lan et al. [25] introduced a multiskip feature-stacking method for enhancing the learnability of action representations. In addition, hidden Markov models (HMMs), dynamic Bayesian networks (DBNs), and dynamic time warping (DTP) are well-studied methods for speed variation in actions. However, actions cannot be reliably estimated in real-world environments using these methods.

Computing handcrafted features from raw video frames and learning classifiers on the basis of the obtained features are a basic two-step approach used in most of the existing methods. In real-world applications, the design of the feature and the choice of the feature are the most difficult and highly problem-dependent issues. Especially for human action recognition, different action categories may look dramatically different according to their appearances and motion patterns. Deep CNNs make some impressive results for the task of action classification [26, 27]. Karpathy et al. [28] trained a deep CNN using 1 million videos for action classification. Gkioxari and Malik [29] built action detection models that select candidate regions using CNNs and then classify them using SVM. Using two-stream deep CNNs with optical flow, Simonyan and Zisserman [30] achieved a result that is comparable to IDT [5]. Ji et al. [31] built a 3D CNN model that extracts appearance and motion features from both spatial and temporal dimensions in multiple adjacent frames.

3. Proposed model for human action recognition

3.1. Multilevel action descriptor

Intraclass variation in the action category is ambiguous, as shown in **Figure 1(a)** and **(b)**. Although the actions of the three persons are *texting* in **Figure 1(a)**, they can be distinguished from a deeper aspect: the first is *texting while sitting*, the second *texting while standing* and is *stationary*, and the third is *texting while walking*. Assigning the same action label (*texting*) is insufficient in video surveillance because they are of different states either in posture or in locomotion for the same action. This is the same problem for the action of *smoking* in **Figure 1(b)**.

The proposed multilevel action descriptor is depicted in **Figure 1(c)**, where the subactions shown in each level are just examples that have been studied in this work and can be easily expanded by adding new subactions. Each of the three action levels, posture, locomotion, and gesture, has a corresponding CNN, and the total three CNNs work simultaneously. The first network, PostureNet, operates on a static cue and captures the shape of the subject of the motion. The second network, LocomotionNet, operates on a motion cue and captures the history of the motion of the subject. And, the third network, GestureNet, operates on a combination of static and motion cues and captures the patterns of a subtle action by the subject. In this descriptor, three levels can be combined to represent many different types of actions with a large degree of freedom.

3.2. Tracking by detection

For real-time applications, a processing time of 20–30 ms for each frame, a stable bounding box for the human action region, and a low false detection rate are the important factors for human detection and tracking. Therefore, we adapt existing methods to provide a stable human action region for subsequent action recognition.

The sliding window is the bottleneck in the processing time of the object detection because many windows, in general, contain no object. To this end, motion detection is performed before object detection to discard regions that are void of motion. The size of the mini motion map is computed with the following equation:

$$\text{size}_{\text{mni-map}} = \frac{\text{size}_{\text{original}} - \text{size}_{\text{detection}}}{\text{stride}}. \tag{1}$$

The default value of $\textbf{size}_{\text{detection}}$ is (64, 128) and that of **stride** is (8, 8) in HOG [12]. **Figure 3** shows the mini motion map. For instance, if the size of the original image is 640 × 360, then the size of the mini motion map is 77 × 34.

In object tracking, three cases exist in the data association problem: (1) adding a new track, (2) updating an existing track, and (3) deleting a track [32]. The procedure for handling multiple detections and tracks is shown in **Figure 4**. When a new track is added, it starts to count the number of frames that the track has updated without detection. If the number is larger than the threshold n_{skip}, the track is considered being disappeared and is therefore deleted.

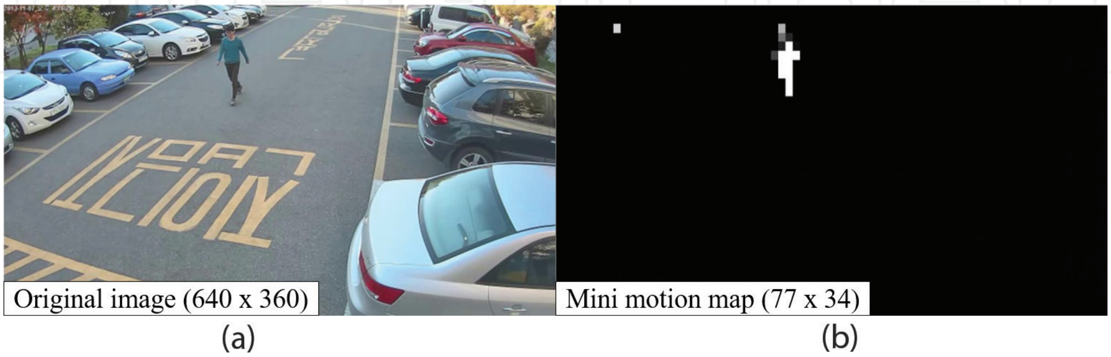

Figure 3. Mini motion map for reducing the unnecessary computation in the HOG-based human detector: (a) original image with a size of 640 × 360 and (b) mini motion map with a size of 77 × 34, which was calculated from the GMM-based motion detection.

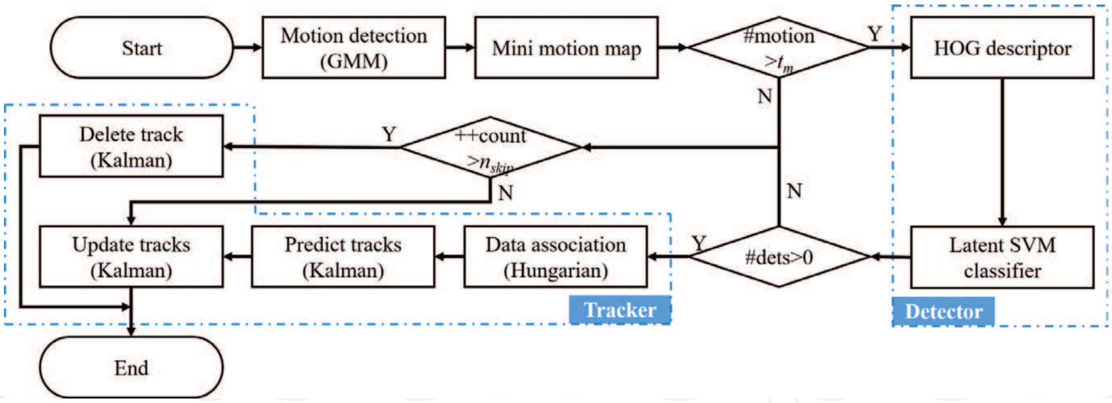

Figure 4. Procedure for multiple detections and tracks.

3.3. Appearance-based temporal features

Appearance-based temporal features are very simple, fast, and work effectively in controlled environments, such as in surveillance systems where the cameras are installed on rooftops or high poles. Therefore, the view angles of the cameras are toward dominant ground planes. A video F is just a real function of three variables:

$$F = f(x, y, t). \tag{2}$$

The frame coordinate (x, y) and t is the index of the video frame. In a multilevel action descriptor, each level has one independent CNN that obtains different appearance-based temporal features. The BDI encodes the static shape information of the subject, denoted as $b(x, y, t)$, and is given by Eq. (3):

$$b(x, y, t) = \begin{cases} 255, & if\, f(x, y, t) - f(x, y, t_0) > \xi_{\text{thr}} \\ 0, & otherwise \end{cases}, \tag{3}$$

It calculates the difference between the current frame $f(x, y, t)$ and the background frame $f(x, y, t_0)$ and compares with a threshold ξ_{thr}. Examples are given in **Figure 5** where BDIs are utilized for the posture level of the subaction descriptor, for example, *sitting* and *standing*.

In a motion history image, pixel intensity is a function of the temporal history of motion at that point. MHI captures the motion history patterns of the actor, denoted as $h(x, y, t)$, and is defined using a simple replacement and a decay operator in Eqs. (4)–(6) [14].

$$d(x, y, t) = \begin{cases} 255, & if\, f(x, y, t) - f(x, y, t-1) > \xi_{\text{thr}} \\ 0, & otherwise \end{cases} \tag{4}$$

$$h(x, y, t) = \begin{cases} \tau_{\max}, & if\, d(x, y, t) = 255 \\ max(0, h(x, y, t-1) - \Delta\tau) & otherwise \end{cases} \tag{5}$$

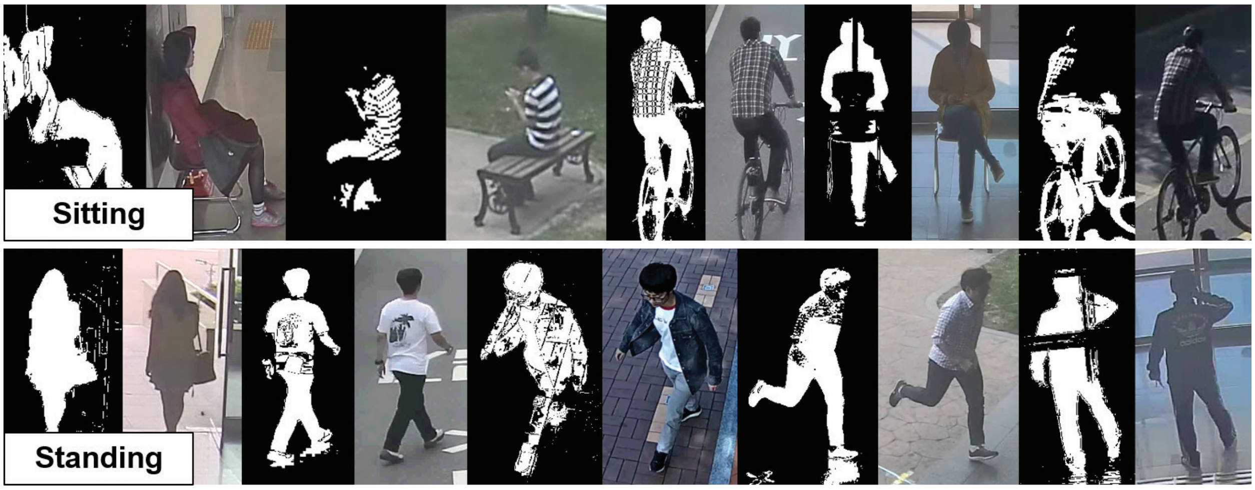

Figure 5. Examples of BDI for different subactions.

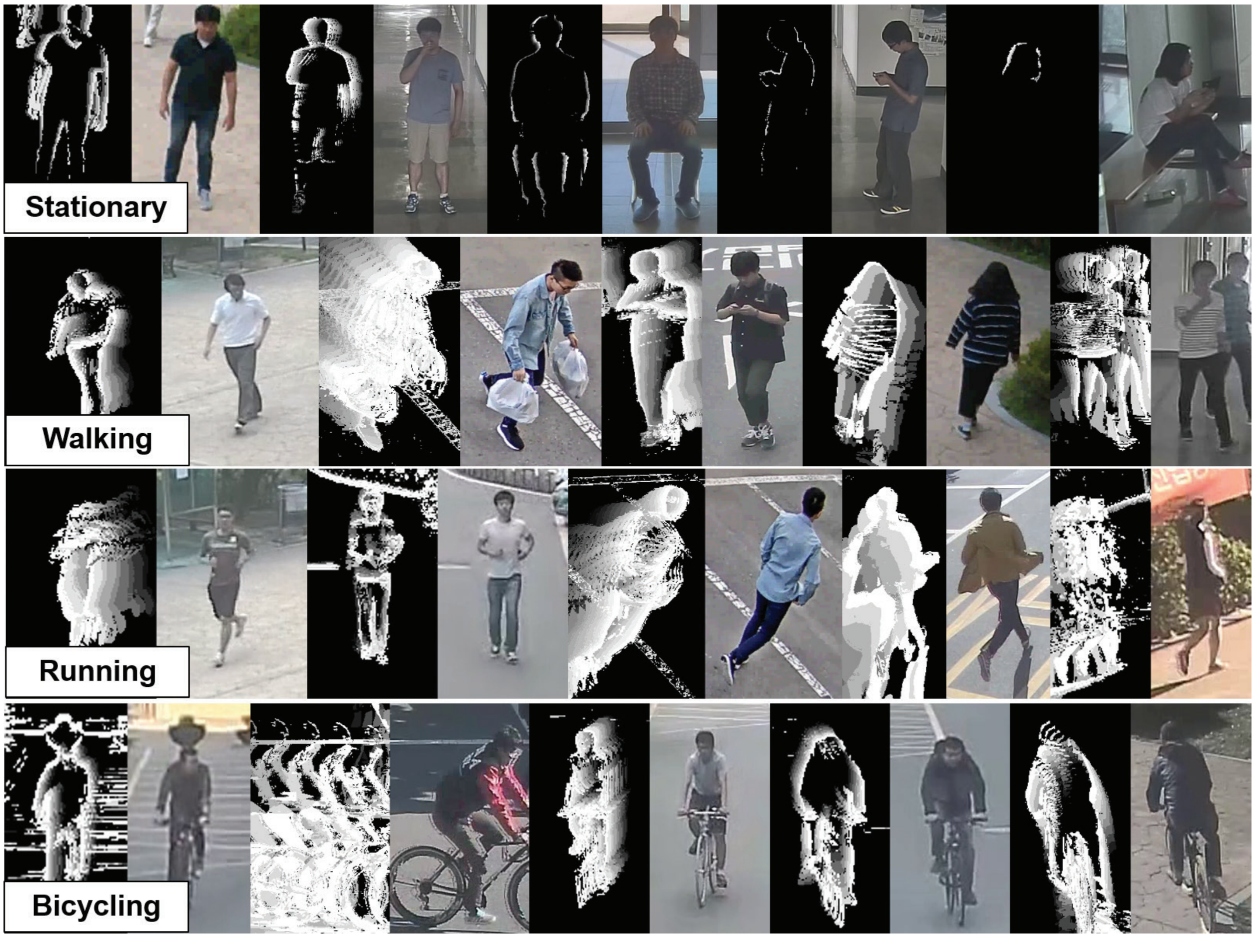

Figure 6. Examples of MHI for different subactions.

$$\Delta\tau = \frac{\tau_{\max} - \tau_{\min}}{n}. \tag{6}$$

MHI is calculated from the difference between the current frame *f*(x, y, t) and the previous frame *f*(x, y, t-1) in Eq. (4). MHI is a vector image of motion, where more recently moving regions are brighter (see **Figure 6**). MHIs are used for the locomotion level of the multilevel action descriptor, which comprises *stationary*, *walking*, *running*, and *bicycling*. MHI captures the motion history cue of the subject, where more recently moving pixel regions are brighter. In Eq. (6), hyperparameter n is critical in defining the temporal range of an action. An MHI with a large n covers a long range of action history; however, it is insensitive to current actions. Similarly, MHI with a small n puts the focus on the recent actions and ignores past actions. Hence, choosing a good n can be fairly difficult.

Weighted average images (WAIs) are applied at the gesture level of the multilevel action descriptor, which comprises *nothing*, *texting*, *smoking*, *phoning*, and *others*. For recognizing

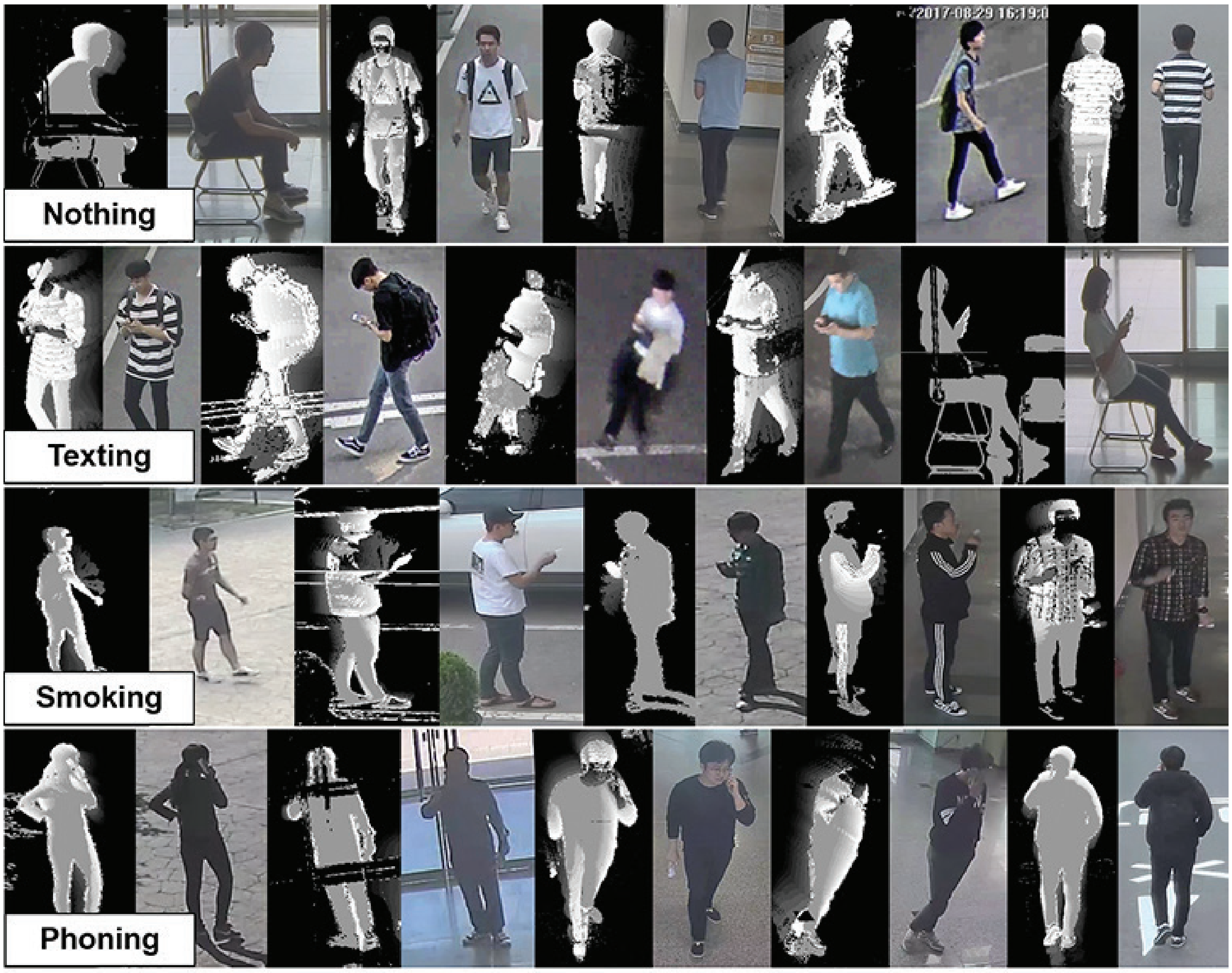

Figure 7. Examples of WAI for different subactions.

subtle actions, the easiest way would be to use the shape or motion history of the actor. It is constructed as a linear combination of BDI and MHI, which is given by Eq. (7):

$$s(x, y, t) = w_1 \cdot b(x, y, t) + w_2 \cdot h(x, y, t), \quad \text{s.t. } w_1 + w_2 = 1. \tag{7}$$

Here, $\mathbf{w} = \{w_1, w_2\}^T$ is another hyperparameter. **Figure 7** shows some examples of WAI for different subactions. WAIs were applied at the gesture level of the subaction descriptor, which comprises *nothing*, *texting*, *smoking*, *phoning*, and *others*. WAI obtained the combined cues of the shape and the motion history. *Texting* (frequently moving fingers) and *smoking* (repeated hand-to-mouth motion) were captured in WAIs.

3.4. Multi-CNN action classifier

In order to reduce the computation time, a lightweight CNN architecture is devised for real-time human action recognition, as shown in **Figure 8**. The architectures of PostureNet, LocomotionNet, and GestureNet are identical with two convolutional layers, two subsampling layers, two fully connected layers, and one softmax regression layer. However, they need to be trained based on the different training data of multilevel action descriptor. The architecture of the network is as follows: Input-Convolution-ReLUs-Max pooling-Convolution-ReLUs-Max pooling-Fully connection-Dropout-Fully connection-Dropout-Fully connection-Softmax regression. The output layer consists of the same number of units as the number of subactions at the corresponding level of the descriptor. If the computational efficiency is not critical, one could use more complicated architectures [33, 34]. In our study, Adam optimizer [35] is used with a learning rate of 1e−3 and $\beta_1 = 0.5$ with a batch size of 256 examples and a weight decay of 5e−4. The networks are trained for 1K iterations [36].

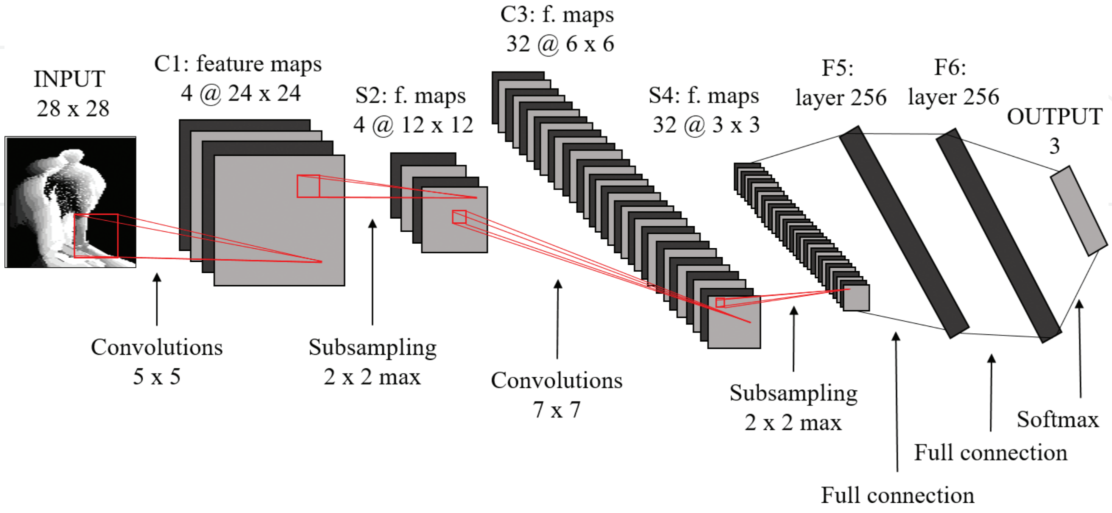

Figure 8. Architecture of a CNN.

4. Experimental results

In this section, an ablation study of the appearance temporal features with the CNN-based approach is presented, and the results of the action recognition are shown with the ICVL dataset. The average processing time was computed based on the ICVL test videos. The experimental results showed that appearance-based temporal features with a multi-CNN classifier effectively recognize actions in surveillance videos.

4.1. Evaluation metrics

To quantify the results, we use the average precision at the frame-based *frame-AP*. The *frame-AP* was used in other approaches at the frame-based evaluation. Frame-AP: recognition is correct if the intersection-over-union (IOU) with the ground truth and detection area at that frame is greater than $\sigma(\sigma = 0.5)$, and the action label is correctly predicted.

4.2. Action recognition on ICVL dataset

LocomotionNet encodes sequential frames as memory capacity to represent actions. However, deciding the number of frames n in Eq. (6) is a highly action-dependent issue. In this work, the number of frames in the MHI was defined by performing a grid search from 5 to 50 frames with an interval of 5. **Figure 9** plots the classification accuracy (mAP) at the frame-based measurement for the subactions at the locomotion level of the multilevel action descriptor. The gray circles are drawn while training LocomotionNet from 100 to 1K iterations with an interval of 100. The circles lie over a 1.96 standard error of the mean and standard deviation in white. The baseline accuracy at $n = 10$ is given by encoding the temporal features. With n equal to 25 frames, LocomotionNet was able to get a performance boost from 1 to 2% of the mAP. This evidence indicates that correctly recognizing one action would need approximately 2 s (15 fps in the ICVL videos).

Table 1 shows the results of each temporal feature with CNN. An ablation study of the proposed approach at the gesture level is presented by evaluating the performance of the two appearance-based temporal features, BDI and MHI, and their combination. Frame-AP is reported for PostureNet, LocomotionNet, and GestureNet. The leading scores of each label are displayed in bold font. As in Eq. (7), WAI is the weighted average of BDI and MHI. GestureNet performed significantly better than PostureNet and LocomotionNet, showing the significance of the combined cues for the task of gesture-level subaction recognition. The GestureNet combines the static and motion history cues to capture specific patterns of the action.

Figure 10 shows the mAP across subactions at the gesture level of the multilevel action descriptor at the frame-based measurement with regard to varying weights on WAI and training iterations of the GestureNet. In the experiment, $w_1 = 0.6$ and $w_2 = 0.4$ show a significant improvement beyond $w_1 = 0.5$ and $w_2 = 0.5$. This implies that the shape cue is more important than the motion history cue in WAI and is quite different from the results in **Table 1**. One possible explanation for this finding is that the motion history cue is more informative than the

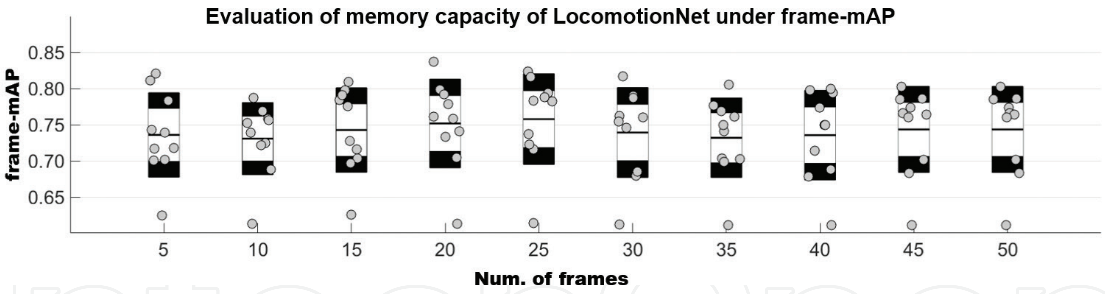

Figure 9. Memory capacity in MHI for the locomotion level of the multilevel action descriptor.

Frame-AP (%)	Nothing	Texting	Smoking	Phoning	mAP
PostureNet	51.3	42.0	11.2	37.9	35.6
LocomotionNet	**62.4**	53.5	14.7	49.2	45.0
GestureNet	53.4	**83.3**	**26.7**	**57.9**	**55.3**

Table 1. Results of the ablation study on the gesture level of ICVL dataset.

shape cue if they are used individually. For the remainder of the experimental results, $w_1 = 0.6$ and $w_2 = 0.4$ were used in WAI.

To evaluate the effectiveness of the action-recognition model, we included the full confusion matrixes as a source of additional insight. **Figure 11** shows that the proposed approach achieved an mAP of 73.2% at the frame-based measurement. The horizontal rows are the ground truth, and the vertical columns are the predictions. Each row was normalized to a sum of 1. The proposed method was able to get most of the subaction categories correct, except for *smoking*. The results of the experiment show that a multilevel action descriptor can eliminate many misclassifications by dividing one action into many subactions that are not at the same levels.

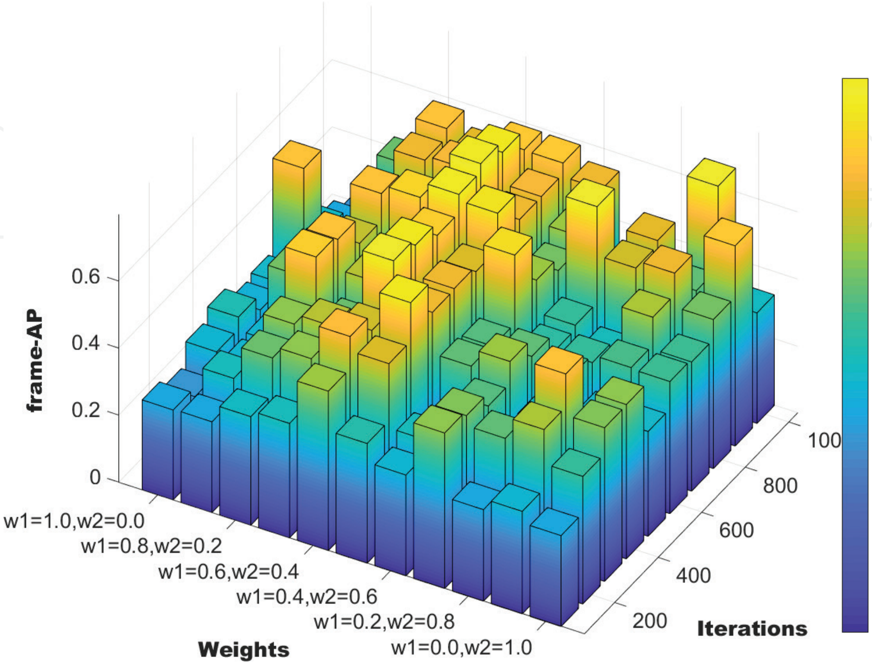

Figure 10. Recognition results with regard to varying weights of WAI and training iterations on GestureNet.

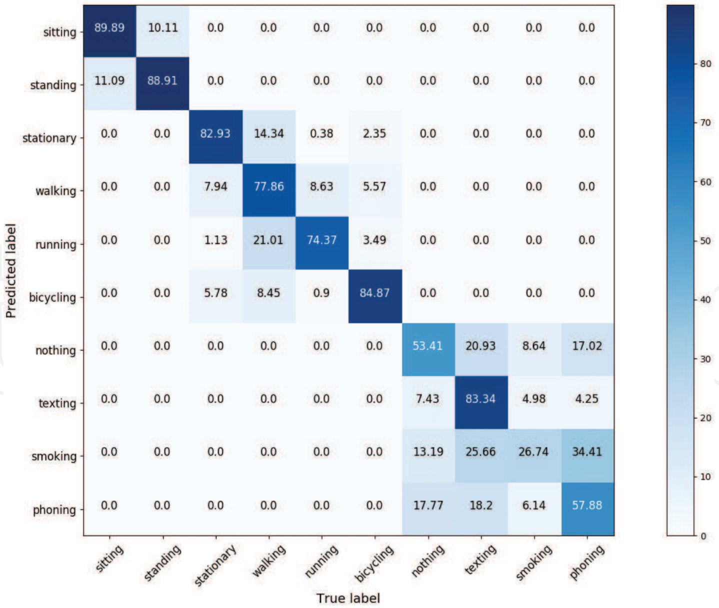

Figure 11. Confusion matrixes of the ICVL dataset at the frame-based measurement for the action-recognition task when using appearance-based temporal features with a multi-CNN classifier.

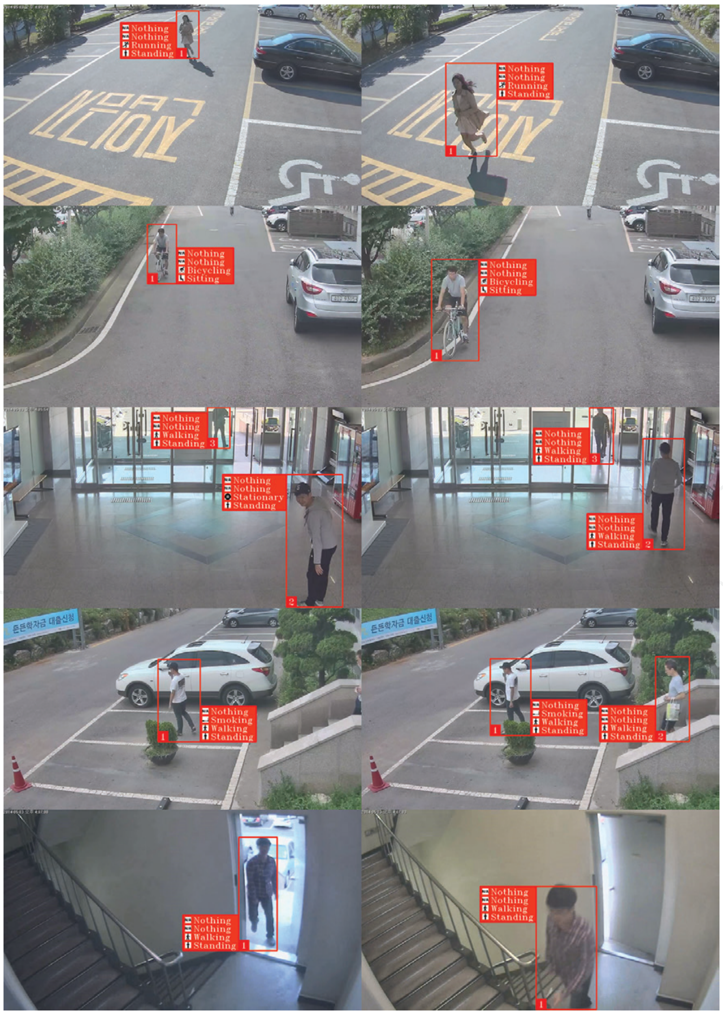
Nothing Nothing Running Standing 1
Nothing Nothing Running Standing 1
Nothing Nothing Bicycling Sitting 1
Nothing Nothing Bicycling Sitting 1
Nothing Nothing Walking Standing 3
Nothing Nothing Stationary Standing 2
Nothing Nothing Walking Standing 3
Nothing Nothing Walking Standing 2
Nothing Smoking Walking Standing 1
Nothing Smoking Walking Standing 1
Nothing Nothing Walking Standing 2
Nothing Nothing Walking Standing 1
Nothing Nothing Walking Standing 1

Figure 12. Examples of action localization and recognition results from the ICVL dataset.

Module	Motion	Detection	Tracking	BDI	MHI	WAI	CNNs	Others	Overall
Processing time (ms)	11.33	11.60	0.28	0.26	0.83	0.12	4.66	12.16	42.93

Table 2. Average processing time of the proposed action detection model.

Figure 12 shows qualitative localization and recognition results using the proposed approach on the test set of the ICVL dataset. Each block corresponds to a video from a different camera. Two frames are shown from each video. The test platform has a PC with an Intel Core i7-4770 CPU at 3.49 GHz with 32 GB memory. The input video was resized to 640 × 480, and the processing time was tested on 72 videos shown in **Table 2**.

5. Conclusions

This work introduced a new approach to real-time action recognition using multilevel action descriptor in video surveillance system. Experimental results demonstrated that a multilevel action descriptor delivers a complete set of information about human actions and significantly eliminates misclassifications by a large number of actions that are built by few independent subactions at different levels. An ablation study showed the effect of each temporal feature when considered separately. Shape and motion history cues are complementary, and the combination of both leads to a significant improvement in action recognition performance. In addition, the proposed action recognition model simultaneously localizes and recognizes the actions of multiple individuals at low computational cost with acceptable accuracy. The model ran at around 25 fps in 640 × 480 frame size, which is suitable for real-time surveillance applications. In future work, we will extend the approach to learn deep motion flow from original frame sequences and combine detecting and recognizing in one network for becoming an end-to-end human action detection framework.

Acknowledgements

This work was supported by the Industrial Technology Innovation Program, "10052982, Development of multiangle front camera system for intersection AEB," funded by the Ministry of Trade, Industry, & Energy (MOTIE, Korea).

Author details

Cheng-Bin Jin, Trung Dung Do, Mingjie Liu and Hakil Kim*

*Address all correspondence to: hikim@inha.ac.kr

Department of Information and Communication Engineering, Inha University, Incheon,
South Korea

References

[1] Yamin H, Peng Z, Zhuo T, et al. Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recognition Letters. 2017 (available online). DOI: 10.1016/j.patrec.2017.08.015

[2] Andreas S, Rainer S. Pedestrian intention recognition using latent-dynamic conditional random fields. In: Intelligent Vehicles Symposium (IV). Seoul, South Korea: IEEE; June 28–July 01, 2015. pp. 622-627

[3] Michalis V, Christophoros N, Loannis K. A review of human activity recognition methods. Frontiers in Robotics and Artificial Intelligence. 2015;**2**(28):1-28. DOI: 10.3389/frobt.2015.00028

[4] Wang H, Klaser A, Schmid C, et al. Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11). Colorado Springs, USA: IEEE; June 20-25, 2011. pp. 3169-3176

[5] Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV '13). Sydney, Australia: IEEE; December 3-6, 2013. pp. 3551-3558

[6] Jiang Z, Lin Z, Davis L. A unified tree-based framework for joint action localization, recognition and segmentation. Computer Vision and Image Understanding. 2013;**117**(10): 1345-1355. DOI: 1016/j.cviu.2012.09.008

[7] Shotton J, Girshick R, Fitzgibbon A, et al. Efficient human pose estimation from single depth images. IEEE Transitions on Pattern Analysis and Machine Intelligence. 2013;**35**(12):2821-2840. DOI: 10.1109/TPAMI.2012.241

[8] Siddiqui M, Medioni G. Human pose estimation from a single view point, real-time range sensor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11). San Francisco, USA: IEEE; June 13-18, 2011. pp. 1-8

[9] Plagemann C, Ganapathi V, Koller D, et al. Real-time identification and localization of body parts from depth images. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '10). Anchorage, USA: IEEE; May 3-7, 2010. pp. 3108-3113

[10] Cutler R, Davis L. Robust real-time periodic motion detection, analysis, and applications. IEEE Transitions on Pattern Analysis and Machine Intelligence. 2000;**22**(8):781-796. DOI: 10.1109/CVPR.1999.784652

[11] Zhang B, Wang L, Wang Z, et al. Real-time action recognition with enhanced motion vector CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16). Las Vegas, USA: IEEE; June 26–July 1, 2016. pp. 2718-2726

[12] Jin C, Li S, Do T, et al. Real-time human action recognition using CNN over temporal images for static video surveillance cameras. Lecture Notes in Computer Science

(Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015;**9315**:330-339. DOI: 10.1007/978-3-319-24078-7_33

[13] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '05). San Diego, USA: IEEE; June 20-26, 2005. pp. 886-893

[14] Yu C, Joachims T. Learning structural SVMs with latent variables. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV '09). Kyoto, Japan: IEEE; September 29–October 2, 2009. pp. 1169-1176

[15] Bobick AF, Davis JW. The recognition of human movement using temporal templates. IEEE Transitions on Pattern Analysis and Machine Intelligence. 2001;**23**(3):257-267. DOI: 10.1109/34.910878

[16] Davis JW, Bobick AF. The representation and recognition of human movement using temporal templates. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97). San Juan, Puerto Rico: IEEE; June 17-19, 1997. pp. 928-934

[17] Ali S, Basharat A, Shah M. Chaotic invariants for human action recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV '07). Rio de Janeiro, Brazil: IEEE; October 14-20, 2007. pp. 1-8

[18] Fathi A, Mori G. Action recognition by learning mid-level motion feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08). Anchorage, USA: IEEE; June 24-26, 2008. pp. 1-8

[19] Laptev I. On space-time interest points. International Journal of Computer Vision. 2005;**64**:107-123. DOI: 10.1007/s11263-005-1838-7

[20] Kim I, Oh S, Vahdat A, et al. Segmental multi-way local pooling for video recognition. In: Proceeding of the ACM International Conference on Multimedia (ICM '13). New York, USA: ACM; October 21-25, 2013. pp. 37-640

[21] Klaser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the British Machine Conference (BMC '08). Leeds, UK: Inria; September 3, 2008. pp. 275:1-10

[22] Li Y, Ye Z, Rehg JM. Delving into egocentric actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15). Boston, USA: IEEE; June 7-10, 2015. pp. 287-295

[23] Vahdat A, Gao B, Ranjbar M, et al. A discriminative key pose sequence model for recognizing human interactions. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV '11). Barcelona, Spain: IEEE; November 6-13, 2011. pp. 1729-1736

[24] Lan T, Wang Y, Yang W, et al. Discriminative latent models for recognizing contextual group activities. IEEE Transitions on Pattern Analysis and Machine Intelligence. 2012;**34**(8):1549-1562. DOI: 10.1109/TPAMI.2011.228

[25] Lan Z, Ming L, Xuanchong L, et al. Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15). Boston, USA: IEEE; June 7-10, 2015. pp. 204-212

[26] Ni B, Yang X, Gao S. Progressively parsing interactional objects for fine grained action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16). Las Vegas, USA: IEEE; June 26–July 1, 2016. pp. 1020-1028

[27] Yeung S, Russakovsky O, Moi G, et al. End-to-end learning of action detection from frame glimpses in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16). Las Vegas, USA: IEEE; June 26–July 1, 2016. pp. 2678-2687

[28] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14). Columbus, USA: IEEE; June 24-27, 2014. pp. 1725-1732

[29] Gkioxari G, Malik J. Finding action tubes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15). Boston, USA: IEEE; June 7-10, 2015. pp. 759-768

[30] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS '14). Montreal, Canada: NIPS; December 08-13, 2014. pp. 568-576

[31] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Transitions on Pattern Analysis and Machine Intelligence. 2013;**35**(1):221-331. DOI: 10.1109/TPAMI.2012.59

[32] Li S. Human re-identification using soft biometrics in video surveillance [thesis]. Incheon: Inha University; 2015

[33] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS '12). Lake Tahoe, USA: NIPS; December 03-08, 2012. pp. 1-9

[34] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14). Columbus, USA: IEEE; June 24-27, 2014. pp. 580-587

[35] Kingma P, Ba J. Adam: A method for stochastic optimization. 2017. pp. 1-15. arXiv Preprint:1412.6980v9

[36] Jin C-B, Do T, Liu M, et al. Real-time action detection in video surveillance using a sub-action descriptor with multi-convolutional neural networks. Journal of Institute of Control, Robotics and Systems. 2018;**24**(3):298-308. DOI: 10.5302/J.ICROS.2018.17.0243

IntechOpen