Pattern Recognition 85 (2019) 198–206

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Prediction of partially observed human activity based on pre-trained deep representation

Dong-Gyu Lee[a], Seong-Whan Lee[b,*]

[a] Department of Computer and Radio Communications Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea
[b] Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea

ARTICLE INFO

Article history:
Received 26 March 2018
Revised 3 July 2018
Accepted 11 August 2018
Available online 14 August 2018

Keywords:
Pre-trained CNN
Human activity prediction
Human interaction
Sub-volume co-occurrence matrix

ABSTRACT

Prediction of complex human activities from a partially observed video is valuable in many practical applications but is a challenging problem. When a video is partially observed, maximizing the representational power of the given video is more important than modeling the temporal dynamics of the activity. In this paper, we propose a novel human activity descriptor for prediction, which can maximize the discriminative power of a system in a compact and efficient way using pre-trained deep networks. Specifically, the proposed descriptor can capture the potentially important pairwise relationships between objects without prior knowledge or preset attributes. The relationship information is automatically reflected during the descriptor construction procedure based on object's participation ratios, local and global motion activations. Pre-trained Convolutional Neural Networks are utilized without additional model training procedure. From a practical point of view, the proposed method is more cost-effective when implementing a smart surveillance system. In the experiments, we evaluate the proposed methods in two cases: (1) prediction accuracy with different observation ratios, and (2) the effect of pre-trained network and layer selection. Experimental results from five public datasets verified the efficacy of the proposed method by outperforming competing methods with stable high-performance regardless of network selection.

1. Introduction

Automated recognition of human activity is an important part of intelligent computer vision systems. Over the past years, many research studies have been conducted to enable automatic human action/activity recognition in videos. For human activity analysis, an accurate recognition of the atomic activity in a video stream is the primary component of the system, and also the most important, as it affects the performance significantly. Although many recognition studies have been performed in an uncontrolled environment by considering real-world scenarios, recognizing human activity in a surveillance video is still challenging owing to the tremendous intra-class variation in activity caused by different visual appearances, motion variations, temporal variability, etc.

In the past, the bag-of-words (BoW) approach [1] with space-time interest points (STIP) [2] has obtained successful results in human activity recognition [3–6]. This approach extracts the low-level features using local patches by treating sequential images as 3D XYT volumes. However, the BoW approach ignores the important aspects of human activity, namely, the location where the feature was extracted or the relationships between multiple objects. Some research studies have achieved remarkable results using preset motion attributes or key-pose information [7–12]. However, attributes such as "Which actions are important attributes?", "When did the motion occur?" or "How many persons are involved?" are problem-dependent factors, and finding these attributes still remains a problem for video surveillance systems.

Most of the existing activity recognition methods are conducted under the assumption that the videos are in a perfect state, which makes these approaches unsuitable for human activity prediction from a partially observed video stream. However, predicting human activity before it is fully executed has broad applications. For example, an autonomous vehicle would able to prevent an accident occurring by predicting potentially dangerous actions. The system can reduce the damage by escaping from the situation before it occurs. An early detection of criminal activity such as fighting can also prevent situations from becoming more serious. With these kinds of applications, prediction on early stage can have an advantage over recognition in the entire video. This is also important because, in real world applications, it is not guaranteed that the system can observe the entire video: The human or interacting object can be occluded, video signals can drop off, and the objects can move out of a field of view.

* Corresponding author.
E-mail addresses: dg_lee@korea.ac.kr (D.-G. Lee), sw.lee@korea.ac.kr (S.-W. Lee).

https://doi.org/10.1016/j.patcog.2018.08.006
0031-3203/

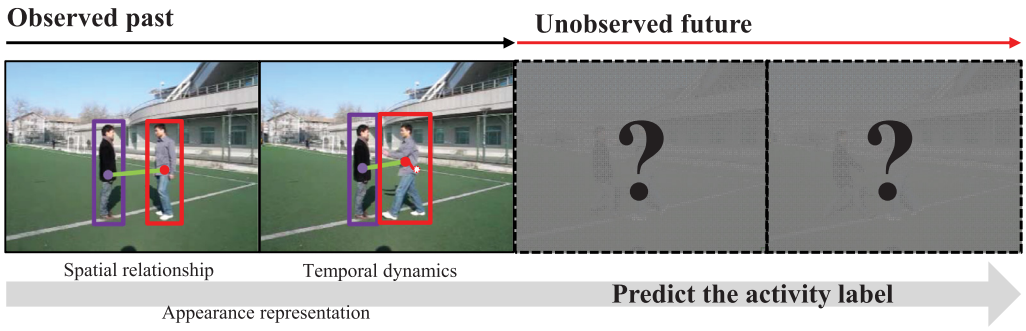

Fig. 1. Illustration of the human activity prediction problem from a partially observed video. Both the appearance of individual objects and the spatio-temporal relationship between objects are important factors.

Since Ryoo [4] tackled the problem of inference of ongoing activity given temporally incomplete observation, the prediction of human activity before it is fully executed has become an interesting field [5,13–17]. Fig. 1 depicts an overview of the human activity prediction problem from a partially observed video. The spatial and temporal relationships between objects are as important as their appearance and motion. The major difference between human activity recognition and prediction is that the temporal location of end point is not given in the prediction task, while the video data is entirely given in the recognition task. Therefore, for an accurate prediction, the discriminative power of video descriptor from small amount of initial frame has to be maximized. We have solved the prediction problem by expressing the partially observed behavior robustly at each time step. Spatio-temporal segment of each object, called sub-volume, that play more important role are automatically given more weights based on its activation ratio.

In this paper, we propose the sub-volume co-occurrence matrix (SCM) method to predict the partially observed human activity label. The proposed method constructs a mid-level descriptor by considering the relationships between objects and finds important sub-volume by using motion activations. The potentially important relationships automatically reflect their information during the SCM construction procedure.

Recently, the Convolutional Neural Network (CNN) has achieved outstanding results on a variety of computer vision tasks [18–22]. In research on human action video classification, methods such as 3D CNN [23], two-stream CNN [24] and multi-stream CNN [25] have been proposed and have shown good performance. However, despite the excellent achievements of these studies, difficulties remain in the use of a CNN for complex human activity analysis in a surveillance video system.

Firstly, learning a good CNN model requires a very large amount of data. However, human activity datasets have a relatively small number of samples, which is insufficient for the proper training of a deep model. The large scale VIRAT dataset [26], some event only has 20 clips which are not enough to train good network model from scratch. Although it is easy to collect a large number of video clips in the current digital era, it is still difficult to collect and label the proper data. Because, unlike video-level event detection where the entire video clip has a single label, complex human activities includes multiple actions occur continuously and even meaningless frames are included. Considering that the human activity class has a large variation in motion, appearance, and length depending on the environment, this is an important factor that decreases the discriminative power of the CNN. Secondly, a significant amount of hardware resources are also required to train a CNN model. The computation cost for video data is incomparably higher than that for an image dataset. Thus, training individual models for every different scenario is practically unrealistic. To handle this problem, the proposed SCM descriptor can utilize pre-trained CNN models trained using heterogeneous data, with minimum additional computational cost. Lastly, and most importantly, most research studies that use a CNN to perform video classification are limited to frame-level representation using the average fusion technique [27,28]. The average fusion reduces the spatio-temporal characteristics of the individual activities. Moreover, the temporal modeling of the extracted features in a continuous time is effective for the video classification task when the sequences are fully observed, but it is difficult to utilize in a prediction task. Therefore, we will overcome these disadvantages by focusing on the richer relationship information available in the process of constructing the SCM descriptor.

In this paper, our main goal is to address the problem of human activity prediction from a partially observed video by maximizing the discriminative power of a descriptor based on a deep representation. More specifically, we focus on representing the complex ongoing human activities using a pre-trained deep network while considering the pairwise interactions between individuals and their participation ratio in the overall scene. The individual actions were represented by extracting sub-volume feature vectors with a pre-trained CNN model in an efficient way. The co-occurrence of individual actions and their relationships were described by proposed SCM. The potentially important key relationships were automatically reflected in the descriptor configuration process using their participation ratios and activations. The experimental results showed that the SCM outperformed state-of-the-art activity prediction methods.

The main contributions of the paper are two-fold: (1) We propose a novel human activity representation method, called sub-volume co-occurrence matrix, which can automatically reflect important relationships without pre-set attributes. (2) We develop a unified framework in which we can robustly predict partially observed human activity using the pre-trained CNN. The cost of building the video systems can be reduced by utilizing pre-trained model without any additional training procedure. Furthermore, the proposed framework can predict not only human-human interaction, but also human-object interaction owing to representational power of the CNN.

This paper is an extension of our previous work [29]. This extension includes a new formulation of descriptor construction with local and global activation, scoring function, participation ratio measure, sequential data representation, and more experimental results with various pre-trained networks and layer analysis comparisons.

2. Related work

Understanding human activity has gained great interest from researchers. For the past few years, the most popular approach for human activity recognition has used local spatio-temporal features with the BoW paradigm [1–6]. In this approach, a sequence

of 2D images was treated as a 3D XYT volume where the STIPs are located. However, in order to understand complex human activities more precisely, richer information about behavior such as local and global relationships, social context and interactions between people is required [9,30]. In particular, the relationships and interactions between individuals play an important role in complex human activity recognition [10–12,31–33]. Exploiting motion attributes or key-poses in human interaction can be useful for attaining more accurate and robust results. Most promising poses or motions of human activities such as interactive phrases [7], action attributes [8], poselets [9], and discriminative key-components [11] have been detected. However, automatically extracting such rich information is still a challenging task.

Meanwhile, most existing methods focus on after-the-fact activity recognition; Ryoo [4] tackled the human activity prediction problem. Integral BoW and dynamic BoW approaches have been proposed to detect ongoing human activity. The proposed activity descriptors were computed for each progress level by averaging the histogram features in the same category. However, the model was sensitive to outliers, and might not be representative when videos have a large variation of appearance or pose. Kong and Fu [5] predicted action labels by capturing the temporal dynamics of human action considering the history of observed features. Cao et al. [13] proposed an action model by learning feature bases using sparse coding and used the reconstruction error in the likelihood computation. Lan et al. [14] proposed a coarse-to-fine hierarchical representation for action prediction. Xu et al. [15] proposed an activity auto-completion model for human activity prediction from partial videos. A video is represented based on discriminative paths in frames divided into a collection of segments. Yu et al. [16] predicted human activity by proposing Spatial-Temporal Implicit Shape Model (STISM). The early recognition of human activities is accomplished by pattern matching through STISM. Li et al. [34] explored the long-duration action prediction problem with a probabilistic suffix tree. However, their work detected segments by motion velocity peaks, which may not be suitable for complex human activity videos.

Recently, deep learning based representations, especially CNNs, have shown superior results for most computer vision tasks by overcoming the problem-dependent limitation of the hand-crafted features in many tasks [18–22,35,36]. Therefore, there have been several attempts to utilize deep representations for human activity understanding. Ji et al. [23] proposed a 3D CNN model which uses multiple channels of information from an input frame to encode motion information for a single human action recognition. Simonyan et al. [24] proposed a two-stream deep convolutional network, which can incorporate both spatial (single image frame) and temporal (multi-frame optical flow) streams. Deng et al. [32] presented a deep neural network based hierarchical graphical model for individual and group activity recognition in surveillance video. The information about the scene, action and pose are considered by a multi-step message passing structure. Le et al. [35] presented an extension of the Independent Subspace Analysis (ISA) method to learn invariant spatio-temporal features from unlabeled video in an unsupervised manner. However, despite their high accuracy, there are still remaining issues such as requiring a large amount of labeled video data and the computational cost of training an individual model.

Donahue et al. [36] proved that the CNN model pre-trained on the ImageNet dataset can be adapted to different tasks owing to the discriminative power of the CNN. They proved that the activation value of fully connected layers had robust discriminative power for images. Xu et al. [28] proposed a latent concept descriptor using a pre-trained convolutional network. They describe videos using the activation value of pre-trained CNN models. However, such an approach is unsuitable for complex human activity representation owing to the lack of the consideration of relationships among people. Although the previous methods have demonstrated their effectiveness for their given purposes, they mostly focus on frame-level event detection or atomic action recognition.

In this work, we represent the ongoing human activity from partially observed video using a pre-trained CNN without additional training. Unlike existing methods, the proposed method automatically captures spatio-temporal relationships between objects without prior knowledge or preset attributes. Specifically, we first extract low-level features using a pre-trained CNN. Using the low-level features, we construct a novel descriptor to represent ongoing complex human activities in a compact and efficient way.

3. Proposed method

In this section, we describe a method for representing a partially observed human activity for early prediction. Given a partially observed human activity video, the proposed method predicts the future activity before the clip ends.

Fig. 2 illustrates the overall flowchart of the proposed method. From a given partially observed video clip, we extract the activation of a fully connected layer weight of each object region for image representation. The coordination of each object is used to measure motion activations. We first generate the sub-volume feature vector for each time segment, then apply a BoW scheme to construct a sub-volume codebook. With the collection of sub-volumes and codebook, an SCM descriptor is constructed for each time step by considering the object appearance, local motion activation, the global motion activation and the relationship between objects. The proposed mid-level SCM descriptor can represent a complex human activity in a single matrix without any constraints. Once the descriptor is constructed, the linear support vector machine (SVM) predicts the class label at the current time step.

3.1. Pre-trained convolutional neural networks

We used publicly available pre-trained CNN models to extract representative image features. Specifically, in this paper, we extracted the low-level features using the MatConvNet toolkit [37] and TensorFlow [38] with the model shared by Simonyan et al. [21]. The network was trained for the ImageNet ILSVRC-2014 classification task [39]. The first 13 layers were convolutional layers, and the last three layers were fully connected layers. For the low-level image representation, unless stated otherwise, we extract the activation value of the first fully connected layer of the network as reported in previous works [20,36,40]. Here, we should note that the activation values after the rectification operation are considered as separate layers. In Section 4.2.2, we compare the classification performance of various pre-trained networks (AlexNet [19], VGG-S [40], VGG-M [40], and VGG-VD16 [21]) or fully connected layers (fc_6, fc_6_relu, fc_7, fc_7_relu).

From a given video clip, we first extract object images for consecutive frames. Here, we should note that it is possible to use any available object detection and tracking algorithms, owing to the fact that main goal of this work was to represent partially observed complex human activities in a single descriptor without any constraint. To utilize the pre-trained CNN models, we resize the images of object regions to 224×224 using bilinear interpolation. The pre-trained CNN models represent each human object image by a 4096-dimensional feature vector, fc_6.

3.2. Sub-volume feature generation

To handle the streaming video in a temporal scale, we divide the video into non-overlapped fixed time duration l, and called sub-volumes. A sub-volume feature vector of object i at the tth

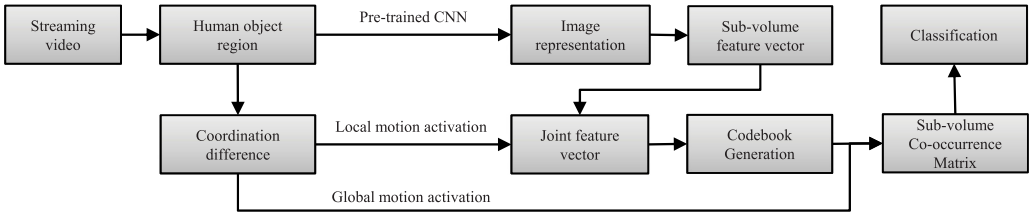

Fig. 2. An overview of the proposed sub-volume co-occurrence matrix construction procedure.

segment is denoted as $\mathbf{f}_i^t = [p, \delta x, \delta y]$, where p denotes the average of the image representation feature vector in a sub-volume. A series of frame-level image feature vectors fc_6 of object i at time t for consecutive l frames are averaged into a 4096-dimensional feature vector. We also use the coordination difference of an object, δx and δy, to denote the movement of the center of the bounding box during the time duration l. The magnitude of coordination difference of each object is calculated as $s_i^t = \sqrt{(\delta x_i^t)^2 + (\delta y_i^t)^2}$, which is called the local motion activation value. The global motion activation is calculated by sum of all local motion activation value in each segment, $\epsilon^t = \sum_{\forall i} s_i^t$. The local and global motion activations denote how actively each object is participating in the activity.

3.3. Bag-of-words assignment

In previous research studies [1,3–5], the BoW paradigm has proved that it can construct a discriminative mid-level descriptor by clustering a set of unknown local patches from spatio-temporal interest points [2]. Our approach also takes advantage of the BoW paradigm to generate the SCM descriptor. Specifically, K-means clustering is performed to generate codewords $\{w_k\}_{k=1}^K$, where k denotes the number of clusters. First, we perform the K-means clustering on sub-volume features of training videos to generate codebook. After the clustering, we assign each sub-volume feature $\mathbf{f}_i^t$ to the corresponding cluster w_k following the BoW paradigm. The index of the corresponding cluster k_i^t is codeword index, which is also the index of the rows and columns of the SCM descriptor during the construction procedure.

3.4. Sub-volume co-occurrence matrix

Assigning sub-volume features using the BoW paradigm for a video representation can lose the temporal dynamics of a video sequence and the spatial relationship between multiple objects. In this section, we describe the process of the proposed SCM construction, which is a mid-level descriptor based on the fully connected layer weights of a CNN model. The SCM can represent the ongoing human interactions using the pairwise relationships of sub-volumes by considering their potential importance.

In each time segment t, we consider the relationships between pairs of all existing sub-volumes while constructing the SCM descriptor. The potentially important pairwise relationships between sub-volumes are automatically considered based on the spatial relationships, global motion activation and the ratio of their participation. From each sub-volume of an object $v_i^t = (\mathbf{f}, x, y, k)$, we first measure the spatial distance between sub-volume i and j as follows:

$$d_{ij}^t = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2}, \tag{1}$$

The overall spatial distance between sub-volume i and the rest j in segment t for $\#pairs$, where $j \neq i$, is aggregated as follows:

$$r^t = \frac{1}{2} \sum_i \sum_{j \neq i} d_{ij}^t \tag{2}$$

Generating a descriptor with a higher weight for a potentially important relationship makes the descriptor more discriminative. Therefore, we assume that if the objects are located close to one another and have a notable activation, which was measured using the ratio of the aggregated distance to the individual distance and the ratio of the local motion activation to global motion activation, it is highly possible for potentially important behavior to occur. The distance between each of sub-volume feature vector $\mathbf{f}$ and codeword w is used to assign the activation value. The difference in distance between sub-volume i and j to the global motion activation represents the participation ratio of the pair in the segment t. The feature scoring function based on sub-volume clustering is calculated as follows:

$$f_p = \log\left(\frac{||w_i^t - \mathbf{f}_i^t|| + ||w_j^t - \mathbf{f}_j^t||}{2} + \psi\right) \tag{3}$$

The SCM descriptor is constructed using the final value considering the ratio of the distance between each sub-volume i, j to the total distance, the ratio of local / global motion activation, and the feature scoring function. After computing all required values between all sub-volumes and relationships, we finally construct the SCM descriptor, as follows:

$$M^t(k_i^t, k_j^t) = \frac{1}{N} \sum_{i, i \neq j} \sum_{1:t} \frac{s_i^t}{\epsilon^t} \frac{r^t}{d_{ij}^t} f_p(\mathbf{f}_i^t, \mathbf{f}_j^t) \tag{4}$$

where N is the normalization term. The value of the sub-volume between i, j is assigned to the SCM using the corresponding cluster index, k_i^t and k_j^t, of each sub-volume. Each of the SCM descriptors is generated for every non-overlapped time step. Thus, from the start, the descriptor is constructed in a cumulative way.

Fig. 3 shows a simple example of the SCM construction procedure for five consecutive time segments when $K = 10$. From the sub-volumes of the first 20% of the entire video, we construct a descriptor that comprises a sub-volume by combining the image feature and the local motion value. Then, we specify the index of each sub-volume using a codebook. Since the sub-volumes of object #1 and object #2 of the first 20% are assigned to the third and sixth clusters in the codebook, respectively, the value of (3,6) in the SCM is increased following Eq. (4). After the 20% of the next 20%, we assign a weight to (9,6) by repeating the same procedure. At this time, the matrix is constructed in a cumulative way. This procedure is repeated for T segments to construct a SCM descriptor for the entire video. When constructing the descriptor, we can see that, in Fig. 3-(2) and -(4), object #1 is allotted to cluster #3, #8 and #9 whereas object #2 is allotted to clusters #6 and #4, respectively. As a result, values are assigned to different indexes and a descriptor becomes distinguishable when there are similar motions and appearances. Furthermore, we can also see that the highest weight is assigned to (8,4) by automatically detecting a potentially important sub-volume, which makes the proposed descriptor more discriminative. Here, we should note that the figure shows a simplified schematic example for the visualization. The matrix M is calculated in a symmetric form. We also should note that (3,6) appears repeatedly; therefore, we can see that relatively high values

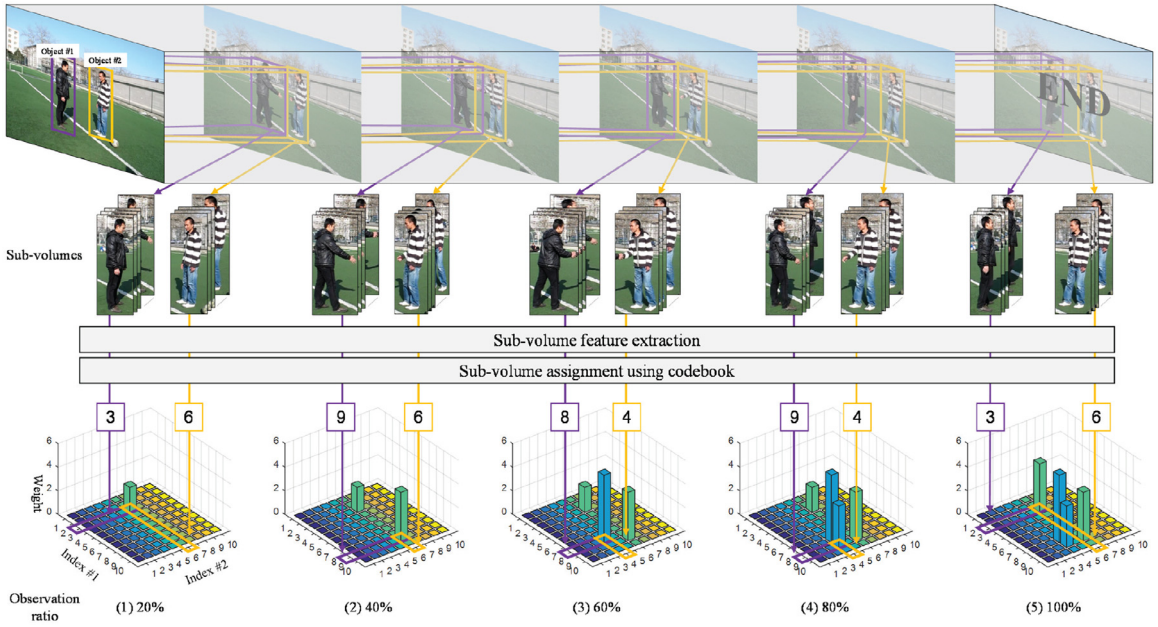

Fig. 3. Example of the proposed SCM construction procedure with five different observation ratios. (1)-(5) denote the constructed SCM descriptor for each of the observation ratios.

are assigned, which is reflected by the standby state before and after taking an action on the dataset. It is not a problem, however, considering that there is no unnatural waiting state in real-world scenarios.

3.5. Classification

In the proposed SCM descriptor, each sub-volume is assigned using their own cluster index of the codebook. Since the human activity is partially observed, in this work, we construct an SCM descriptor at each time segment t. Specifically, in the training stage, we first extract sub-volumes from each object from fully observed videos. We then perform K-means clustering using all sub-volumes from the training data. The codebook w_k is then used to construct the SCM descriptor of the training data. We then train the descriptor using the one-against-all linear SVM found in Eq. (5), where we have the training data $\{\mathbf{x}_n, \mathbf{y}_n\}$.

$$\underset{\mathbf{w}}{\text{minimize}}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_n \max(1 - \mathbf{y}_n \mathbf{w}^T \mathbf{x}_n, 0) \tag{5}$$

In the testing stage, our goal is to predict the class label $\mathbf{y}_n$ of a partially observed video for [1: t]. Given a partially observed video, we extract sub-volume feature vectors at each time segment. We then construct a SCM for use in codebook w_k, which was generated at the training state. The class label $\mathbf{y}_n$ is predicted by a SVM using the descriptor for [1: t] segments.

4. Experiments

4.1. Datasets and experimental setup

To validate the effectiveness of the proposed method, we tested the proposed SCM approach on five datasets: the BIT-Interaction dataset [10], the UT-Interaction dataset (sets #1 and #2) [41], the VIRAT 1.0 Ground dataset and the VIRAT 2.0 Ground dataset [26]. The proposed method was evaluated for classifying videos of incomplete executions using 10 different observation ratios from 0.1 to 1, as done in previous works [5,13,42]. The observation ratio means the proportion of frames used for analysis in each clip. If a full video contains N frames, the accuracy of the proposed SCM descriptor is evaluated by representing it with the first *observation ratio* $\times N$ frames. The half video and full video denote an observation ratio of 0.5 and 1, respectively.

The BIT-Interaction dataset consists of eight classes of human-human interaction: bow, boxing, handshake, high five, hug, kick, pat, and push. In this dataset, each class contains 50 video clips. The videos were captured in realistic scenes with moving objects in a cluttered background, partially observed body parts, illumination conditions, variations in human object scale, unrelated human objects in the scene, etc. For the experiment, 272 videos were selected as the training set (34 clips per class) and the remaining 128 videos were used as testing samples (16 clips per class). For the quantitative comparison, the performance of competing methods and the training/testing clip indexes were provided by the BIT-Interaction dataset provider [5,10].

The UT-Interaction dataset consists of six classes of human-human interaction: push, kick, hug, point, punch, and handshake. The dataset is composed of two sets of video; set #1 and set #2, which were captured at different environments, for example, illumination conditions, human object scale, camera jitters and background movements. Each set contains 10 videos for each class. The UT- Interaction dataset has been tested by several state-of-the-art methods [4,5,7,9,13]. For this experiment, we performed the leave-one-out cross-validation strategy as done in previous works [4,5].

The VIRAT 1.0 Ground dataset and the VIRAT 2.0 Ground dataset include six classes of human-vehicle interaction: person Loading an object to A Vehicle (LAV), person Unloading an object from A car/Vehicle (UAV), person Opening A vehicle/car Trunk (OAT), person Closing A vehicle/car Trunk (CAT), person Getting Into a Vehicle (GIV), person Getting Out of a Vehicle (GOV). The VIRAT 1.0 Ground dataset consists of around 3 h of videos and the VIRAT 2.0 Ground dataset consists of over 8 h of videos. The activities are recorded from different parking lots. For the experiment, half of video clips were used as the training samples and the remaining video clips were used as testing samples following previous works [43–45].

Table 1
Comparison of the prediction results on the BIT-Interaction dataset with 10 different observation ratios.

Method	0.1 (%)	0.2 (%)	0.3 (%)	0.4 (%)	0.5 (%)	0.6 (%)	0.7 (%)	0.8 (%)	0.9 (%)	1.0 (%)
Linear SVM(BoW)	17.19	18.75	37.50	50.00	56.25	57.03	58.59	57.81	60.16	64.06
Dynamic BoW [4]	22.66	25.78	40.63	43.75	46.88	54.69	55.47	54.69	55.47	53.13
MSSC [13]	21.09	25.00	41.41	43.75	48.44	57.03	60.16	62.50	66.40	67.97
MTSSVM [42]	28.13	32.81	45.31	55.47	60.00	61.72	67.19	70.31	71.09	76.56
MMAPM [5]	32.81	36.72	53.90	59.38	67.97	63.28	68.75	75.00	75.78	79.69
CNN_{avg} w. AlexNet	**61.13**	**64.84**	**66.50**	68.16	70.02	72.07	73.54	75.39	75.10	73.83
CNN_{avg} w. VGG-VD16	58.98	61.13	62.30	62.99	63.09	63.28	66.99	72.36	75.59	77.05
Proposed Method w. AlexNet	56.18	61.25	65.87	71.18	78.22	78.93	79.98	83.30	85.16	87.01
Proposed Method w. VGG-VD16	58.37	62.75	65.87	**73.62**	**78.42**	**80.57**	**83.30**	**85.25**	**86.52**	**88.70**

4.2. Performance evaluation

4.2.1. BIT-Interaction dataset

In this experiment, we evaluated the performances of our proposed method on the BIT-Interaction dataset. We compared the performance with state-of-the-art methods, i.e., dynamic BoW [4], mixture of training video segment sparse coding (MSSC) [13], multiple temporal scale support vector machine (MTSSVM) [42], max-margin action prediction machine (MMAPM) [5], and the baseline method linear SVM. We also performed a classification of the average fusion of the activation value feature of sub-volumes with two different CNN models, AlexNet [19] and VGG-VD16 [21]. In this experiment, to extract a frame-level image feature vector, we used the activation values of the first fully connected layer, fc_6. Further information about the CNN models can be found in Section 4.2.2. The K value in the K-means clustering of the sub-volumes was set to $K = 15$. The validation of the effects of the K values is described in Section 4.2.3. The performances of the competing methods were provided by the BIT-Interaction dataset provider [5,10].

The experimental results from the BIT-Interaction dataset are shown in Table 1. The table lists the prediction accuracies measured with observation ratios of 0.1–1. The result indicates that the proposed method achieved the best accuracies in the comparison. The proposed method achieved 78.42% recognition accuracy when only the first 50% of the frames were observed. We also obtained an accuracy of 88.70% from fully observed video sequences. For the first 30% of the observation ratio, the proposed method showed a slightly lower performance than the CNN_{avg}. The performance degradation was because the front part of the interaction clips did not include momentous information. Some clips only showed standing or slightly moving activities at the first 30% of the video. However, in the rest of the experiments, our method outperformed the average fusion of the fc_6 feature because we modeled not only the appearance of the humans but also the spatial relationship between human objects, the local and global motion activation, and the ratio of its participations. This is notable because we can expect better performances in a real-world scenario as there are no waiting motions before taking an action or waiting times to capture a dataset.

4.2.2. Pre-trained CNN models

The proposed method extracts image representation features using pre-trained CNN models learned from heterogeneous data. Thus, in this section, we will show through the experimental results which pre-trained network to use and which layer has the best performance when extracting the activation value. First, we compared the performances between networks using the AlexNet [19], VGG-S [40], VGG-M [40], and VGG-VD16 [21] networks, which were trained on the ImageNet dataset [39]. We also compared the performances between the first fully connected layer fc_6 and the second fully connected layer fc_7. The activation layers after the *ReLU* operation, fc_6_relu and fc_7_relu, were also considered as separate layers in this experiment. To show that the proposed method is better than the average fusion technique, we also compared the performance of the original feature vectors with average fusion.

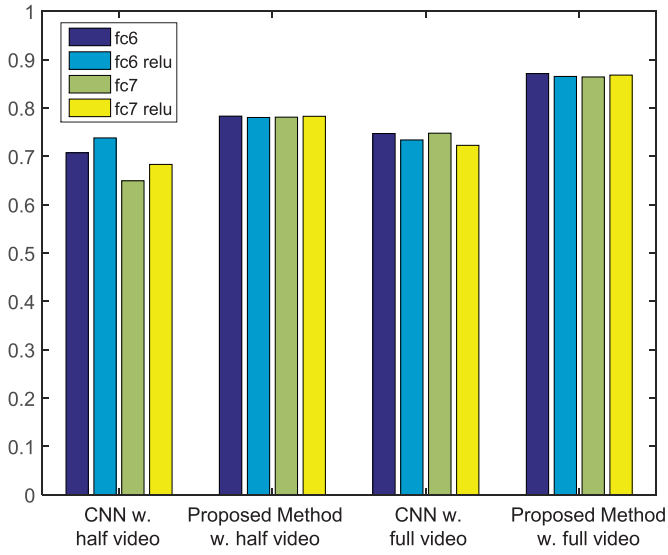

Fig. 4. Average of the classification performance using multiple pre-trained CNN models.

Table 2 shows the prediction performances on both half- and full-video sequences. The average fusion of the activation values showed large performance deviations depending on the selected layer and network. Therefore, we cannot guarantee the prediction performance when a specific network is selected and used. On the other hand, with the proposed method, the system showed stability and high accuracy regardless of which layer or network was used.

Fig. 4 shows the average recognition performance of each layer for the four pre-trained CNN models. In the case of the average fusion of the activation value, the fc_6_relu layer showed the best performance on the half-video sequence, whereas the fc_7 layer showed the best performance on the full-video sequence. This is inconsistent with the fact that fc_6 had the highest performance of 77.66% in the half-video sequence in the VGG-S network and 77.05% in the VGG-VD16 network in the full-video sequence. On the other hand, the proposed method showed the best overall accuracy and was superior not only in performance, but also in terms of reliability. In the rest of the experiments presented in this paper, we used the fc_6 layer of the VGG-VD16 networks, which showed the best performance in the full-video.

4.2.3. Effect of the parameter

In this work, the K-means clustering for generating sub-volume codebook may have an effect on the performance. We conducted an experiment on the BIT-Interaction dataset to validate the im-

Table 2
Comparison of the prediction performances using various networks and layers.

Methods	Half video				Full video			
	fc_6(%)	fc_6_relu(%)	fc_7(%)	fc_7_relu(%)	fc_6(%)	fc_6_relu(%)	fc_7(%)	fc_7_relu(%)
CNN_{avg} w. AlexNet	70.02	72.46	73.24	68.55	73.83	73.34	73.83	71.00
CNN_{avg} w. VGG-S	76.66	74.51	63.09	67.29	73.44	73.05	73.24	70.61
CNN_{avg} w. VGG-M	73.24	73.83	59.77	61.62	74.61	73.44	76.56	73.54
CNN_{avg} w. VGG-VD16	63.09	74.41	63.67	75.88	77.05	73.73	75.49	74.02
Proposed Method w. AlexNet	78.22	78.32	77.93	78.32	87.01	86.62	86.52	86.62
Proposed Method w. VGG-S	78.22	78.03	78.03	78.22	87.11	86.13	86.72	86.62
Proposed Method w. VGG-M	78.42	77.64	**78.52**	78.52	86.62	86.04	86.82	86.82
Proposed Method w. VGG-VD16	78.42	78.22	77.93	78.03	**88.70**	87.30	85.64	87.01

Table 3
Comparison of the classification performances of different K values on the BIT-Interaction dataset.

K	5 (%)	10 (%)	15 (%)	20 (%)	50 (%)	100 (%)
Accuracy	83.4	87.4	**88.7**	86.3	82.6	81.9

Table 4
Comparison of the prediction results on the UT-Interaction dataset (set #1).

Method	Half video (%)	Full video (%)
Bag-of-Words(BoW)	50.00	81.67
Integral BoW [4]	65.00	81.70
Dynamic BoW [4]	70.00	85.00
MSSC [13]	70.00	83.33
SC [13]	70.00	76.67
MMAPM [5]	78.33	**95.00**
CNN_{avg} w. VGG-VD	66.94	77.78
Proposed Method w. VGG-VD	**82.67**	90.22

pact of the K value. The experimental result in Table 3 shows the recognition performance with fully observed videos.

The result shows that the recognition performance was not much affected by K value. The accuracy changed by 6% when the K value was changed from 5 to 100. This lack of effect implies that the key volumes were assigned separately owing to the representational power of the CNN. However, the performance could decrease if the K value is set without consideration of the training data size.

4.2.4. UT-Interaction dataset

We also compared the proposed method on the UT-interaction dataset with state-of-the-art methods, i.e., dynamic BoW [4], Lan et al.'s work [14], sparse coding (SC), MSSC [13], and MMAPM [5]. The experiments were performed on both set #1 and set #2, respectively, as done in previous works [5,7,30]. In this experiment, we used the activation of the fc_6 layer from the VGG-VD16 network.

Table 4 shows the prediction results on set #1 of the UT-Interaction dataset. We can see that the proposed method outperformed all the competing methods by achieving an impressive 82.67% recognition accuracy when only half of the testing videos were observed. However, the proposed method had a slightly lower performance than MMAPM on the fully observed video because not only were some video segments similar in appearance, but also the local motion activation of the other sub-volumes took more weight in the rest of the descriptor construction procedure of the total. Moreover, the environmental complexity of the video from set #1 of the UT-Interaction dataset was clear enough to capture human appearance. The set #1 videos were captured in a gray-colored parking lot background without noise such as jittering, etc. The spatio-temporal local patches with STIP were sufficient to present human activities in these simple environments. However, the backgrounds in set #2 of the UT-Interaction dataset consisted of grass and jittering twigs, which could be noisy local patches. In the following experiment on set #2, we performed prediction task by treating each activity occurring simultaneously as separate clips for evaluation. Table 5 shows the prediction results on set #2 of the UT-Interaction dataset.

Table 5
Comparison of the prediction results on the UT-Interaction dataset (set #2).

Method	Half video (%)	Full video (%)
Bag-of-Words(BoW)	66.00	80.00
Dynamic BoW [4]	61.00	70.00%
Lan et al.[14]	68.33	83.33
MSSC [13]	71.67	81.67
SC [13]	66.67	80.00
MMAPM [5]	75.00	86.67
CNN_{avg} w. VGG-VD	64.72	81.11
Proposed Method w. VGG-VD	**83.22**	**89.40**

As we expected, the proposed method achieved an accuracy of 83.22% and 89.40% on the half-video and full-video sequences, respectively. It is remarkable that the proposed method outperformed all the competing methods on the both the half- and the full-video sequences with more complex scenarios such as a complex background, partial occlusion, unrelated object movements, etc. Furthermore, the proposed method outperformed all the competing method in half-video observation in the three different datasets. It is noteworthy considering that the temporal location of the atomic human activity is given imperfectly in real-world scenarios.

4.2.5. VIRAT datasets

To demonstrate how the proposed method can be further extended with arbitrary objects, we conducted an experiment for human-object interaction recognition using the proposed SCM descriptor. We compared the performance with state-of-the-art methods, i.e., Zhu et al.'s work [43], Bayesian Network (BN) [44], and Deep Hierarchical Context Model [45]. The experiments were performed on both the VIRAT 1.0 Ground dataset and VIRAT 2.0 Ground dataset, respectively. Each human and vehicle is treated as an object without prior separation. We extracted fc_6 feature using VGG-VD16 [21] network for frame-level image representation. Here, we should note that, in this experiment, there were some frames that human objects are not properly detected occasionally due to the small object size or the occlusion to the vehicle. The image representation feature in the frame is set to be a zero vector at the previously detected position.

The experimental result on VIRAT 1.0 dataset is shown in the Table 6. The table lists the classification accuracies for each of six classes and average. Our proposed method achieved better overall performance over all the other comparison methods with 76.3% of recognition accuracy. Here, we also report the recognition accuracy of CNN_{avg} feature obtained 73.2% accuracy.

Table 6
Comparison of the activity recognition results on the VIRAT 1.0 dataset.

Method	LAV (%)	UAV (%)	OAT (%)	CAT (%)	GIV (%)	GOV (%)	Average (%)
Zhuet al. [43]	52.1	57.5	69.1	72.8	61.3	64.6	62.9
BN [44]	100	71.4	50.0	54.5	45.2	73.5	65.8
DHCM [45]	66.7	85.7	50.0	81.8	64.5	70.6	69.9
CNN_{avg}	84.7	61.0	88.4	93.7	52.6	48.4	73.2
Proposed Method	70.5	80.0	94.7	94.7	53.7	64.2	**76.3**

Table 7
Comparison of the activity recognition results on the VIRAT 2.0 dataset.

Method	LAV (%)	UAV (%)	OAT (%)	CAT (%)	GIV (%)	GOV (%)	Average (%)
SVM-STIP	44.44	51.72	10.00	52.63	58.33	33.33	41.74
BN [44]	77.78	58.62	35.00	63.16	68.75	48.89	58.70
DHCM [45]	66.67	68.97	45.00	89.47	70.83	57.78	66.45
CNN_{avg}	88.76	40.93	53.43	62.67	68.65	59.41	62.31
Proposed Method	90.76	73.91	90.22	88.40	57.61	76.63	**79.53**

We also conducted an activity recognition experiment on the VIRAT 2.0 dataset using same settings as used in the VIRAT 1.0 dataset. The six person-vehicle interaction events were successfully classified as shown in Table 7. The proposed method outperformed other competing state-of-the-art method in four out of six events, and significantly improved the top average recognition accuracy from 66.45% to 79.53%.

In the VIRAT experiments, the sub-volume of vehicle object was assigned to eight of twenty clusters. The rest of the human objects were assigned to fourteen clusters, which is interesting that the vehicle and human objects are sharing two clusters. Nevertheless, our results show that, in the given experiments on VIRAT datasets, the proposed method can work with non-human objects by clustering the sub-volumes. Even when the different objects are assigned to the same cluster, the descriptor can be discriminative by measuring participant, activation ratio, and feature scoring.

5. Conclusion and future work

Predicting the label of a complex human activity from a partially observed video has many practical applications, but is a challenging problem. To address these issues, in this paper we proposed a method for representing the ongoing human activity from a partially observed video in a compact and efficient way. The proposed sub-volume co-occurrence matrix descriptor can maximize the discriminative power of the video using pre-trained CNN models. Utilizing deep representation such as a CNN without additional training procedure can be a cost-effective and practical approach when implementing a smart surveillance system. Owing to the representational power of the proposed sub-volume co-occurrence matrix descriptor for short time segments, we can predict an activity class label even when the video has not ended. For a real application, a smart surveillance system needs to classify the activity label without the exact temporal location of the activity. In our experimental studies on public datasets, i.e., the BIT-Interaction, the UT-Interaction dataset (set #1 and set #2), the VIRAT Ground 1.0 dataset, and the VIRAT Ground 2.0 dataset, we validated the effectiveness of the proposed method, which outperformed the other competing methods from the literature. The superior performance on the half-video classification task and the robustness in a complex environment such as a cluttered background are noteworthy because the proposed method may be more competitive than other methods in a real-world streaming video. We also validated the stability and applicability of the proposed method by showing the pre-trained network and layer comparison. This is a promising advantage for real applications such as a smart surveillance system with limited resources. However, in the experiment on the VIRAT datasets, some activities show lower recognition performance than others. This could be a limitation of the proposed method that can work better on interaction between moving objects than with non-moving objects. Improving expandability of the proposed descriptor will be our forthcoming research.

Acknowledgment

This work was supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) [No. 2014-0-00059, Development of Predictive Visual Intelligence Technology] and [No. 2016-0-00152 , Development of Smart Car Vision Techniques based on Deep Learning for Pedestrian Safety].

References

[1] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[2] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2–3) (2005) 107–123.
[3] Y. Zhu, N. Nayak, U. Gaur, B. Song, A. Roy-Chowdhury, Modeling multi-object interactions using æstring of feature graphsg, Comput. Vis. Image Underst. 117 (10) (2013) 1313–1328.
[4] M.S. Ryoo, Human activity prediction: Early recognition of ongoing activities from streaming videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1036–1043.
[5] Y. Kong, Y. Fu, Max-margin action prediction machine, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1844–1858.
[6] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: a literature review, Pattern Recognition 48 (8) (2015) 2329–2345.
[7] Y. Kong, Y. Jia, Y. Fu, Interactive phrases: semantic descriptions for human interaction recognition, IEEE Trans Pattern Anal Mach Intell 36 (9) (2014) 1775–1788.
[8] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3337–3344.
[9] M. Raptis, L. Sigal, Poselet key-framing: A model for human activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2650–2657.
[10] Y. Kong, Y. Jia, F. Yunde, Y. Fu, Learning human interaction by interactive phrases, in: European conference on computer vision, Springer, Berlin, Heidelberg, 2012, pp. 300–313.
[11] Y.S. Sefidgar, A. Vahdat, S. Se, G. Mori, Discriminative key-component models for interaction detection and recognition, Comput. Vis. Image Underst. 135 (2015) 16–30.
[12] T. Lan, Y. Wang, W. Yang, G. Mori, Beyond actions: discriminative models for contextual group activities, in: Proceedings of the Advances in Neural Information Processing Systems, 2010, pp. 1216–1224.
[13] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Mark Siskind, S. Wang, Recognize human activities from partially observed videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2658–2665.
[14] T. Lan, T.-C. Chen, S. Savarese, A hierarchical representation for future action prediction, in: Proceeding of European Conference on Computer Vision, Springer, 2014, pp. 689–704.

[15] Z. Xu, L. Qing, J. Miao, Activity auto-completion: predicting human activities from partial videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3191–3199.
[16] G. Yu, J. Yuan, Z. Liu, Predicting human activities using spatio-temporal structure of interest points, in: Proceedings of the ACM International Conference on Multimedia, 2012, pp. 1049–1052.
[17] V. Bloom, V. Argyriou, D. Makris, Linear latent low dimensional space for online early action recognition and prediction, Pattern Recognit. 72 (2017) 532–547.
[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[20] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.
[21] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
[22] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning., in: Proceedings of the AAAI, 4, 2017, p. 12.
[23] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[24] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[25] Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, J. Yuan, Multi-stream CNN: learning representations based on human-related regions for action recognition, Pattern Recognit. 79 (2018) 32–43.
[26] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J.T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al., A large-scale benchmark dataset for event recognition in surveillance video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3153–3160.
[27] Z.-Z. Lan, L. Jiang, S.-I. Yu, S. Rawat, Y. Cai, C. Gao, S. Xu, H. Shen, X. Li, Y. Wang, et al., Cmu-informedia at TRECVID 2013 multimedia event detection, in: Proceedings of the TRECVID 2013 Workshop, 1, 2013, p. 5.
[28] Z. Xu, Y. Yang, A.G. Hauptmann, A discriminative CNN video representation for event detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1798–1807.
[29] D.-G. Lee, S.-W. Lee, Human activity prediction based on sub-volume relationship descriptor, in: Proceedings of the IEEE International Conference on Pattern Recognition, 2016, pp. 2060–2065.
[30] Y. Kong, Y. Fu, Close human interaction recognition using patch-aware models, IEEE Trans. Image Process. 25 (1) (2016) 167–178.
[31] W. Choi, S. Savarese, Understanding collective activities of people from videos, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2014) 1242–1257.
[32] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M.J. Roshtkhari, G. Mori. Deep structured models for group activity recognition. arXiv preprint arXiv:1506.04191 (2015).
[33] P.-S. Kim, D.-G. Lee, S.-W. Lee, Discriminative context learning with gated recurrent unit for group activity recognition, Pattern Recognit. 76 (2018) 149–161.
[34] K. Li, Y. Fu, Prediction of human activity by discovering temporal sequence patterns, IEEE Trans. Pattern. Anal. Mach. Intell. 36 (8) (2014) 1644–1657.
[35] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3361–3368.
[36] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, in: International Conference on Machine Learning, 2014, pp. 647–655.
[37] A. Vedaldi, K. Lenc, Matconvnet: convolutional neural networks for matlab, in: Proceedings of the ACM International Conference on Multimedia, 2015, pp. 689–692.
[38] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, Software available from tensorflow.org.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[40] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014).
[41] M.S. Ryoo, J.K. Aggarwal, UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), 2010, (http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html).
[42] Y. Kong, D. Kit, Y. Fu, A discriminative model with multiple temporal scales for action prediction, in: Proceeding of European Conference on Computer Vision, 2014, pp. 596–611.
[43] Y. Zhu, N.M. Nayak, A.K. Roy-Chowdhury, Context-aware modeling and recognition of activities in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2491–2498.
[44] X. Wang, Q. Ji, A hierarchical context model for event recognition in surveillance video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2561–2568.
[45] X. Wang, Q. Ji, Hierarchical context modeling for video event recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (9) (2017) 1770–1782.

Dong-Gyu Lee received the B.S. degree in Computer Engineering from Kwangwoon University, Seoul, Korea, in 2011. He is currently a Ph.D. student in the Department of Computer Science and Engineering in Korea University. His research interests include computer vision, machine learning, and computational models of vision.

Seong-Whan Lee received his B.S. degree in Computer Science and Statistics from Seoul National University, Seoul, in 1984, and his M.S. and Ph.D. degrees in Computer Science from the Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1986 and 1989, respectively. Currently, he is the Hyundai-Kia Motor Chair Professor and the head of the Department of Brain and Cognitive Engineering at Korea University. He is a fellow of the IEEE, IAPR, and the Korea Academy of Science and Technology. His research interests include pattern recognition, artificial intelligence and brain engineering.