Activity Prediction Using a Space–Time CNN and Bayesian Framework

Hirokatsu Kataoka[1], Yoshimitsu Aoki[2], Kenji Iwata[1] and Yutaka Satoh[1]
[1]*National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan*
[2]*Keio University, Kanagawa, Japan*
hirokatsu.kataoka@aist.go.jp

Keywords: Activity Prediction, Space-time Convolutional Neural Networks (CNN), Bayesian Classifier, Dense Trajectories

Abstract: We present a technique to address the new challenge of activity prediction in computer vision field. In activity prediction, we infer the next human activity through "classified activities" and "activity data analysis. Moreover, the prediction should be processed in real-time to avoid dangerous or anomalous activities. The combination of space–time convolutional neural networks (ST-CNN) and improved dense trajectories (iDT) are able to effectively understand human activities in image sequences. After categorizing human activities, we insert activity tags into an activity database in order to sample a distribution of human activity. A naive Bayes classifier allows us to achieve real-time activity prediction because only three elements are needed for parameter estimation. The contributions of this paper are: (i) activity prediction within a Bayesian framework and (ii) ST-CNN and iDT features for activity recognition. Moreover, human activity prediction in real-scenes is achieved with 81.0% accuracy.

1 INTRODUCTION

In past years, techniques for human sensing have been studied in the field of computer vision (Moeslund et al., 2011) (Aggarwal and Ryoo, 2011). Human tracking, posture estimation, activity recognition, and face recognition are some examples of these, which have been applied in real-life environments. However, computer vision techniques proposed hitherto have been studied only with respect to "post-event analysis." We can improve computer vision applications if we can predict the next activity, for example, to help avoid abnormal/dangerous behaviors or recommend the next activity. Hence, we need to consider "pre-event analysis.

In this paper, we propose a method for activity prediction within a space–time convolutional neural network (ST-CNN) and Bayesian framework. The approach consists of two steps: activity recognition and data analysis. Human activities are recognized using ST-CNN and improved dense trajectories (iDT), a state-of-the-art motion analysis technique. The method outputs activity tags such as *walking* and *sitting* at each frame. To construct an activity database, the temporal activity tag is accumulated. A naive Bayes classifier analyzes the activity database on a high level and predicts the next activity in a given image sequence. At the same time, our framework combines an activity recognition technique with data mining. The contributions of this paper are: (i) activity prediction on a daily living dataset through high-level recognition and data analysis and (ii) effective human activity recognition with state-of-the-art approaches and improved features. Related work on activity recognition and prediction is discussed below.

Activity recognition. Since Laptev *et al.* proposed space–time interest points (STIP) (Laptev, 2005), we have focused on vision-based classification, especially space–time feature analysis. STIP detects space-time Harris corners in *x*-*y*-*t* image space, then a feature vector is calculated in a bag-of-words (BoW) framework (Csurka et al., 2004). Klaser *et al.* (Klaser et al., 2008) improved the feature descriptor based on space–time histograms of oriented gradients (3D HOG) to achieve a more robust representation for human activity recognition. Moreover, Laptev *et al.* improved the STIP approach by combining histograms of oriented gradients (HOG) and histograms of flows (HOF) (Laptev et al., 2008), which are respectively derived from shape and flow feature space. Niebles *et al.* proposed topic representation in a STIP feature space with statistical modeling (Niebles et al., 2006). In this method, the probabilistic latent semantic analysis model is used to create model topics at each activity for activity classification.

An effective approach for activity recognition is dense trajectories (DT), proposed by Wang *et al.* (Wang et al., 2011) (Wang et al., 2013), which is a feature description using dense sampling feature points in an image sequence. Rohrbach *et al.* experimentally demonstrated that DT is better than other approaches such as the posture-based approach (Zinnen et al., 2009) on the MPII cooking activities dataset (Rohrbach et al., 2012), which consists of fine-grained activities. DT approaches have also been proposed in (Raptis et al., 2013) (Li et al., 2012) (Jain et al., 2013) (Peng et al., 2013) (Kataoka et al., 2014a) (Wang and Schmid, 2013). Raptis *et al.* attempted to generate a middle-level representation, using a simple posture detector with location clustering (Raptis et al., 2013). Li *et al.* translated a feature vector into another feature vector at a different angle using the "hankelet" transfer algorithm (Li et al., 2012). To effectively eliminate extra optical flows, Jain *et al.* applied an affine transformation matrix (Jain et al., 2013) and Peng *et al.* proposed dense optical flows captured in motion boundary space (Peng et al., 2013). Kataoka *et al.* improved the DT feature by adding a co-occurrence feature descriptor and dimensional compression. Wang *et al.* improved DT (Wang and Schmid, 2013) by adding camera motion estimation, detection-based noise cancelling, and Fisher vector (FV) classification (Perronnin et al., 2010). More recently, the combination of CNN features and iDT has achieved state-of-the-art performance in activity recognition (Jain et al., 2014). Jain *et al.* employed per-frame CNN features from layers 6, 7, and 8 using AlexNet (Krizhevsky et al., 2012), which is a well-regarded neural net approach. The combination of iDT and CNN synergistically improve recognition performance.

Activity prediction. Here, we review three types of prediction approaches: trajectory-based prediction, early activity recognition, and activity prediction.

(i) *Trajectory-based prediction:* Pellegrini *et al.* proposed local trajectory avoidance (LTA) (Pellegrini et al., 2009) for prediction systems. LTA estimates a location in the very near future from the positions and velocities of tracked people. The authors in (Kitani et al., 2009) achieved scene analysis and estimation using the state-of-the-art inverse optimal control method. This method dynamically predicts a human's position. The approaches introduced here are mainly used to predict pedestrians trajectories in surveillance situations.

(ii) *Early activity recognition:* Ryoo recognized activities in the early part of the activity (Ryoo, 2011). The framework calculates simple feature descriptions and accumulates histograms for early activity recognition. This method cannot predict activities perfectly because the framework is based on recognition in the early frames of an activity.

(iii)*Activity prediction:* Li *et al.* proposed the latest work in the field of activity prediction (Li et al., 2014). The approach predicts an activity using the causal relationship between activities that occur differently several times. They accomplished several seconds prediction as a "long-duration" activity.

However, the related work that we described here comprise one-by-one activity matching approaches or learning feature and next state correspondence. We propose a completely new framework to understand the context of activity sequences through the activity data analysis of daily living. The data-driven analysis allows us to achieve activity prediction with context in an indoor scene. Therefore, we address and improve activity prediction with a combination of computer vision and data analysis techniques.

The rest of the paper is organized as follows. In Section 2, we present the overall framework for activity recognition and prediction. In Sections 3 and 4, we describe detailed activity recognition and prediction, respectively. In Section 5, we present experimental results on human activity recognition and prediction using a daily living dataset. Finally, Section 6 concludes the paper.

2 PROPOSED FRAMEWORK

Figure 1 presents the workflow of the prediction system. Three attributes are added to the naive Bayes classifier. The classifier calculates a likelihood for each predicted activity. The most likely tag is selected as the predicted activity.

Figure 2 shows the feature descriptor for human activity recognition. The representation consists of ST-CNN and iDT, which is the state-of-the-art approach to activity recognition (Jain et al., 2014). We also employ a concatenated representation of these two feature descriptors.

3 ACTIVITY RECOGNITION

3.1 iDT

We employ Wang's iDT (Wang and Schmid, 2013) to create BoW vectors (Csurka et al., 2004) for activity recognition. The idea of iDT is to densely

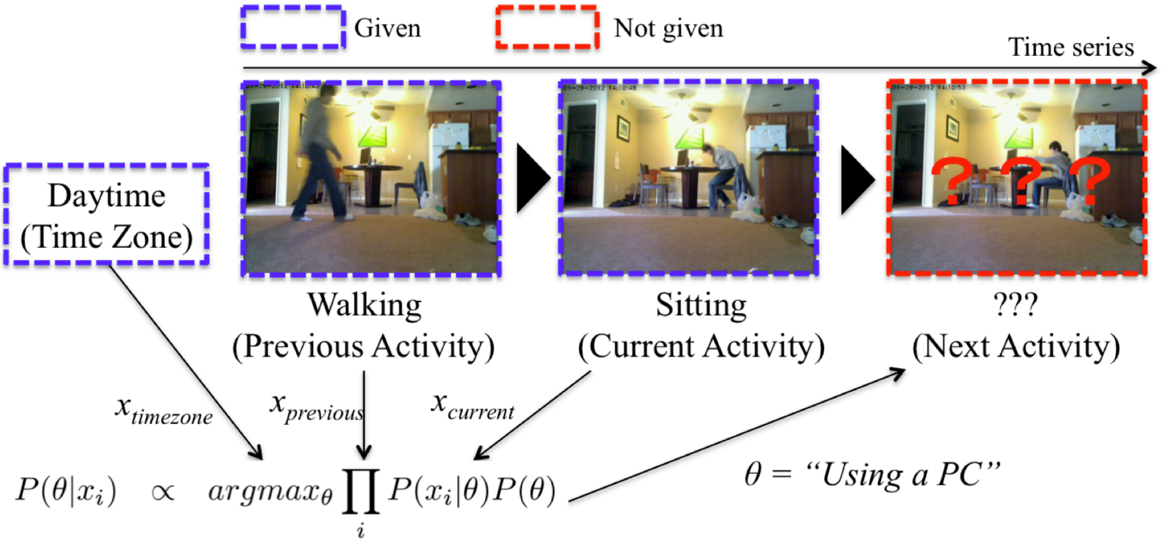


Figure 1: Process flow of proposed activity prediction approach

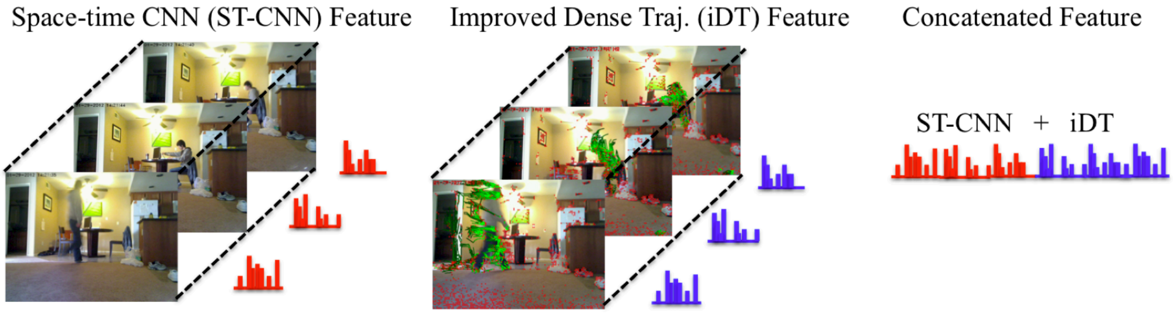

Space-time CNN (ST-CNN) Feature
Improved Dense Traj. (iDT) Feature
Concatenated Feature
ST-CNN + iDT

Figure 2: Concatenation of ST-CNN and iDT for human activity recognition

sample an image and extract the spatio-temporal features from the trajectories. Feature points at each grid cell are computed and tracked using Farneback optical flow (Farneback, 2003) in the other images of the video. To address scale changes, the iDT extracts dense flows at multiple image scales, where the image size increases by a scale factor of $1/\sqrt{2}$. In iDT, flow among the frames is formed by concatenating the corresponding parts of the images. This setup allows us to grab detailed motion at the specified patch. The length of a trajectory is set to be 15 frames, therefore, we record 0.5 s activities at 30 fps video.

In the feature extraction step, the iDT adopts HOG, HOF, and motion boundary histograms (MBH) as the local feature descriptors of an image patch. The size of the patch is 32 pixels, which is divided into 2 × 2 blocks. The number of HOG, HOF, and MBH dimensions are 96, 108, and 96, respectively. The size of HOG and MBH consist of 2 (x) × 2 (y) × 3 (t) × 8 (directions). HOF is described by a 2 (x) × 2 (y) × 3 (t) × 9 (directions) image quantization.

The iDT features are divided into visual words in BoW (Csurka et al., 2004) by k-means clustering. In our implementation, the iDT features are clustered into $4,000$ visual words. In this vectorization, each activity video is represented with the BoW vector containing the frequency of the visual words in activity videos.

In daily living activity recognition, fine-grained categorization is necessary for high-level performance. In the case of fine-grained activities, minor differences frequently occur among human activities. This makes visual distinction difficult using existing feature descriptors. According to Kataoka *et al.* (Kataoka et al., 2014a), their approach categorizes fine-grained activities such as *cut* and *cut slices* in a cooking dataset. The approach captures co-occurrence feature descriptors in the framework of iDT to distinguish subtle changes in human activity areas. Co-occurrence histograms of oriented gradients (CoHOG) (Watanabe et al., 2009) and Extended CoHOG (ECoHOG) (Kataoka et al., 2014b) are applied to vectorize co-occurrence features into BoW vectors.

CoHOG (Watanabe et al., 2009): CoHOG is designed to accumulate the co-occurrences of pairs. Counting co-occurrences of the image gradients at different locations and in differently sized neighborhoods reduces false positives. The co-occurrence histogram is computed as follows:

$$g(x,y) = \arctan \frac{f_y(x,y)}{f_x(x,y)} \tag{1}$$

$$f_x(x,y) = I(x+1,y) - I(x-1,y) \tag{2}$$

$$f_y(x,y) = I(x,y+1) - I(x,y-1) \tag{3}$$

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} 1, \\ \text{if}\, d(p,q) = i \\ \text{and}\, d(x+p,y+q) = j(4) \\ \\ 0 \\ \text{otherwise} \end{cases}$$

where $I(x,y)$ is the pixel value, $g(x,y)$ is the gradient orientation, $C(i,j)$ denotes the co-occurrence value of each element of the histogram, coordinates, (p,q) depict the center of the feature extraction window, coordinates $(p+x,p+y)$ denote the position of the pixel pair in the feature extraction window, and $d(p,q)$ is one of eight quantized gradient orientations.

ECoHOG (Kataoka et al., 2014b): Here, we explain the methods for edge magnitude accumulation and histogram normalization in ECoHOG. This improved feature descriptor is described below.

Human shape can be described using histograms of co-occurring gradient orientations. Here, we add to them the magnitude of the image gradients, which leads to an improved and more robust description of human shapes. The sum of edge magnitudes represents the accumulated gradient magnitude between two pixel edge magnitudes at different locations in the image block. In this way, for example, the difference between human motion and background is strengthened. ECoHOG is defined as follows:

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} \|g_1(p,q)\| + \|g_2(p+x,q+y)\| \\ \text{if}\, d(p,q) = i \\ \text{and}\, d(p+x,q+y) = j \\ \\ 0 \;\text{otherwise} \end{cases}$$

where $\|g(p,q\|$ is the gradient magnitude, and $C(i,j)$ and all the other elements are defined as in Eqs. (2)–(4).

The brightness of an image changes with respect to the light sources. The feature histogram should be normalized to be sufficiently robust for human detection under various lighting conditions. The range of normalization is 64 dimensions, that is, the dimension of the co-occurrence histogram. The equation for normalization is given as:

$$C^{'}_{x,y}(i,j) = \frac{C_{x,y}(i,j)}{\sum_{i'=1}^{8}\sum_{j'=1}^{8} C_{x,y}(i',j')}, \tag{6}$$

where C and $C^{'}$ denote histograms with and without normalization, respectively.

Vectorization: BoW is an effective vectorization approach for not only object categorization but also for activity recognition. The framework is based on feature vector quantization from a large number of features extracted from image sequences. However, co-occurrence features tend to need high-dimensional space. A low-dimensional feature is generally easier to divide into the right class. Kataoka *et al.* (Kataoka et al., 2014a) applied principal component analysis to compress this high-dimensional space (1,152 dims) into a 70-dimension vector. Finally, the low-dimensional vector is used to create a BoW vector in classification. The size of the BoW vector is based on the original iDT paper (Wang et al., 2013) as 4,000 dimensions. The BoW vector is used to carry out learning and recognition steps.

3.2 ST-CNN

We propose temporal concatenation features for CNN that are simple but more effective features for space–time motion analysis. Jain *et al.* (Jain et al., 2014) extracted CNN features at each frame. We also apply CNN features; however, space–time information should be employed in activity recognition. Here, we concatenate CNN features with temporal direction, as shown in Figure 2. We basically apply VGG Net (Simonyan and Zisserman, 2014), which is a deeper neural net model of 16 and 19 layers than the 8-layer AlexNet. Recently, researchers have claimed the depth of the neural net is the most important factor for classification performance. At the same time, feature representation has been sophisticated enough for object classification and detection. The 16-layer VGG Net is applied in this study.

(5)

3.3 Activity Definition based on ICF

The International Classification of Functioning, Disability, and Health (ICF) was proposed in 2001 ((WHO), 2001). The ICF extended the International Classification of Impairments, Disabilities, and

Handicaps (ICIDH) to apply to all people. Moreover, the ICF defined certain activities in daily life, from which we selected the activities for our framework. The activities and part of their definitions are given below.

- **d166 Reading:** performing activities involved in the comprehension and interpretation of written languages.
- **d4103 Sitting:** getting into and out of a seated position and changing body position from sitting down to any other position. **d4104 Standing:** getting into and out of a standing position or changing body position from standing to any other position.
- **d4105 Bending:** tilting the back downwards or to the side, at the torso.
- **d4452 Reaching:** using the hands and arms to extend outwards and touch or grasp something.
- **d450 Walking:** moving along a surface on foot, step by step, so that one foot is always on the ground.
- **d550 Eating:** indicating the need for and carrying out the coordinated tasks and actions of eating food that has been served, bringing it to the mouth, and consuming it in culturally acceptable ways.
- **d560 Drinking:** indicating the need for and taking hold of a drink, bringing it to the mouth, and consuming the drink in culturally acceptable ways.

In this work, we define three prediction activities, namely d166 Reading, d4452 Reaching (including using a PC and other activities), and having a meal (including d550 Eating and d560 Drinking). The three activities target indoor scene activities because they tend to occur as long-term activities compared with activities such as d4104 Standing and d4105 Bending.

4 ACTIVITY PREDICTION

To predict the next activity, an activity database is analyzed using a data mining algorithm. We explain the procedure that uses a naive Bayes classifier and a database of daily living. We investigate what the system understands and whether it can predict human activities.

4.1 Activity Database Structure

Figure 3 shows an example of the database structure for the daily living dataset. From the activity recognition, we obtain spatio-temporal activity tags that are accumulated in an activity database. We predict the *next activity* by using three input attributes, *time of day*, *previous activity* and *current activity*. In the example in Figure 3, daytime (*time of day*), walking (*previous activity*), and sitting (*current activity*) are the input into the naive Bayes classifier, “using a PC” is recognized as the *next activity* by the activity predictor.

4.2 Naive Bayes Classifier

The Bayes classifier outputs a *next activity* θ by analyzing the daily living dataset. A *next activity* θ can be calculated from *time of day* x_1, *previous activity* x_2 and *current activity* x_3. A naive Bayes classifier is a simple Bayesian model that considers independence among the attributes. The naive Bayes classifier equation for activity prediction is given below.

$$P(\theta|x_i) = argmax_{\theta} \frac{\prod_i P(x_i|\theta)P(\theta)}{\Sigma_{N_\theta} \prod_i P(x_i|\theta)P(\theta)} \quad (7)$$

$$\propto argmax_{\theta} \prod_i P(x_i|\theta)P(\theta), \quad (8)$$

where N_θ $(= 3)$ is the number of predicted activities. Here, we define three activities that include reading, reaching, and having a meal. The naive Bayes classifier is frequently employed in the data mining community because of its simple learning method and high accuracy.

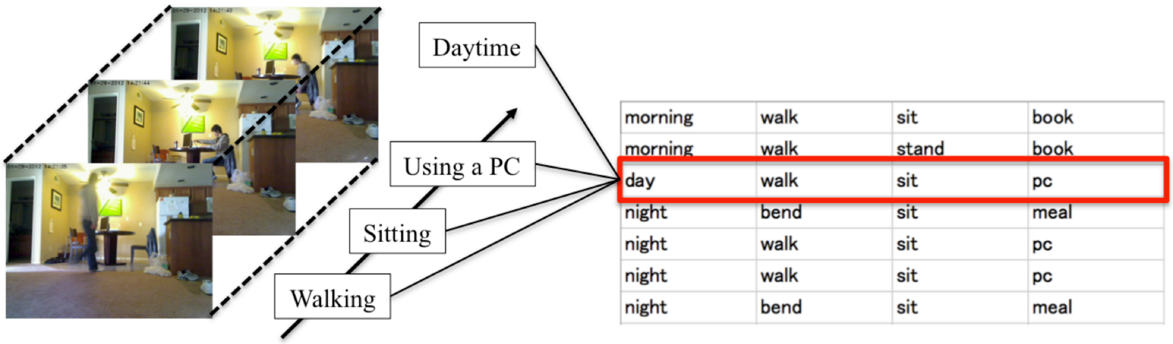
5 EXPERIMENTS

In this section, we present details of an activity recognition and prediction experiment on a daily living dataset. Figure 3 shows a part of the daily living dataset, illustrating the flow of motion for the activities walk–sit–use a PC. This section consists of the dataset description as well as the performance results for the activity recognition experiment and the prediction experiment.

5.1 Daily Living Dataset

We captured more than 20 h of video in an indoor room. The dataset consists of 640 × 480 pixels video at 30 fps.

Four attributes, “*time of day*, “*previous action*, “*current action*, and “*next activity*” are stored in the database (Figure 3). The system predicts the “*next activity*” from the other three attributes. Attribute “*time of day*” takes on the value of “morning, “daytime, or “night; “*previous activity*” and “*current activity*” are



morning	walk	sit	book
morning	walk	stand	book
day	walk	sit	pc
night	bend	sit	meal
night	walk	sit	pc
night	walk	sit	pc
night	bend	sit	meal

Figure 3: Example of a daily scene and database accumulation

extracted by using the activity recognition technique explained in Section 3. In this scene, the activity recognition system classifies four different activities: “bend, “sit, “stand, and “walk. The target activities for “*next activity*” are “reading a book, “having a meal, and “using a PC” after sitting.

5.2 Activity Recognition Experiment

To investigate the effectiveness of the ST-CNN and iDT approach on the daily living dataset, we implemented several recognition strategies. The comparison approaches include per-frame CNN and improved iDT features such as co-occurrence features (Kataoka et al., 2014a). Moreover, we performed activity classification with all feature descriptors. Scenes were taken from the daily living dataset to show the effectiveness of the iDT approach for complicated activities.

Table 1 shows the accuracy of the iDT features with six descriptors and various CNN features. In the daily living dataset, four activities (one for both eating and drinking) are targeted for prediction, and we also need to recognize four activities: bend, sit, stand, and walk. The results show that the co-occurrence features (CoHOG and ECoHOG) outperform the other iDT features (HOG, HOF, MBHx, and MBHy) with respect to classification. Moreover, the method that integrated all iDT features is a more accurate feature descriptor. We conclude that the CNN feature is an effective approach for human activity classification. Table 1 indicates the ST-CNN is slightly better than other single features in iDT and CNN. The descriptor that integrated all features with iDT and CNN achieved the best performance rate.

Table 1: Accuracy of the ST-CNN and iDT approach on the daily living dataset.

Feature	%
iDT(HOG) (Wang et al., 2013)	65.7
iDT(HOF) (Wang et al., 2013)	60.2
iDT(MBHx) (Wang et al., 2013)	67.6
iDT(MBHy) (Wang et al., 2013)	62.3
iDT(CoHOG) (Kataoka et al., 2014a)	77.3
iDT(ECoHOG) (Kataoka et al., 2014a)	78.7
iDT(All features)	84.9
CNN (Simonyan and Zisserman, 2014)	99.6
ST-CNN	99.7
ST-CNN+iDT	**99.8**

5.3 Activity Prediction Experiment

We also carried out an activity prediction experiment on the daily living dataset. Intention (*next activity*) was estimated using the three attributes, *time of day*, *previous activity*, and *current activity*. We set the probability threshold of the naive Bayes classifier for deciding the *next activity* at 80%. However, we can calculate two or more next activity candidates and it is possible to rank these activities using the Bayesian framework. The activity recognition method (using the “integrated iDT feature in the activity recognition experiment) achieved high performance in real-time. The feature integration is effective for daily activity recognition. The dataset includes eight activities for recognition and prediction. (d166 Reading, d4103 Sitting, d4104 Standing, d4105 Bending, d4452 Reaching (including using a PC), d450 Walking, and d550/d560 Eating/Drinking, i.e., having a meal). Figure 5 shows the results of activity prediction for the daily scenes. The system predicted the future activities in advance. In this example, the series of activities walking–bending–sitting–having a meal for the daily scene and walking–reaching–reading for the laboratory scene were used. Our proposed method estimated the activity "having a meal" after sitting (Figure 5 row 1) and "reading a book (reaching)" having taken a book (Figure 5 row 2). In the case of Figure 5 row 3, the Bayesian network calculated the following results for each attribute: "*time of day*" = "night, "*previous activity*" = "walking, "*current activity*" = "taking. In this case, the performance results for analysis are 16% (read), 0% (PC) and 84% (meal). According to these percentages, the system output "having a meal (drinking)" using the attributes. The activity "using a PC" is not displayed in Figure 5 because its probability was 0%. The daily scene dataset includes "reading a book, "having a meal," and "using a PC" as the *next activity*. Table 2 shows the accuracy of activity prediction on the daily living dataset. The possible activities for prediction were d166 Reading, d4452 Reaching (including using a PC), and d550/d560 Eating/Drinking (having a meal). In total, we achieved an 81.0% performance accuracy for activity prediction. Thus, activity recognition and data mining allow a human's next activity to be predicted. Furthermore, the prediction system runs at 5.26×10^{-8} s because activity prediction using the naive Bayes classifier can be executed using only three multiplications.

The distribution of prediction time is shown in Figure 4. The most frequent time is around 5.0 s because our proposed approach outputs only one next activity. The temporal gap between activities is around 5.0 s in the daily living dataset. Moreover, the proposed approach generally predicts within approximately 5 s, and the maximum it requires is 15–20 s. Although the prediction time depends on the activity sequence, the proposed approach accomplishes state-of-the-art prediction. Our method performs better than (Ryoo, 2011) and is of the same standard as (Li et al., 2014) with respect to prediction time. We believe the most important point is the achievement of high accuracy prediction using a data mining approach. The proposed approach understands the predicted activity from the context in activity sequences, in contrast to the cause and effect used in (Li et al., 2014).

To include more varied situations and predict more long-term activity, we would like to add attributes based on the activity history database (e.g., situation, place, more than two activities, and various numbers of activities) as well as improve the data mining technique for activity prediction.

Table 2: Accuracy of activity prediction.

Intention	Accuracy (%)
d166:reading	73.4
d4452:reaching	82.0
d550&d560:having a meal (eating & drinking)	88.5
Total	81.0

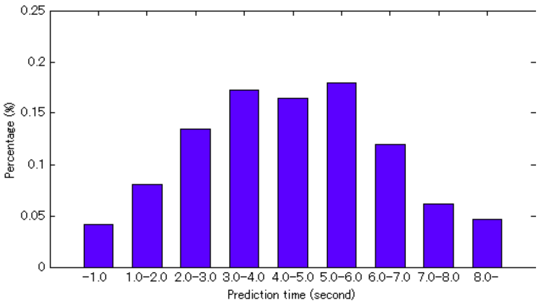


Figure 4: Distribution of prediction time

6 CONCLUSION

We proposed an activity prediction approach using activity recognition and database analysis. To recognize activities, a concatenated vector consisting of an ST-CNN and iDT was employed for more effective human activity recognition. A naive Bayes classifier effectively predicted human activity from three attributes including two previous activities and time of day. We believe the combination of computer vision and data analysis theory is beneficial to both fields.

In the future, we would like to include posture and object information in activity recognition. With a good understanding of these elements, activity recognition and prediction can be improved. Moreover, we would like to improve the approach for fine-grained activity prediction by using a large number of classification methods for activity recognition.

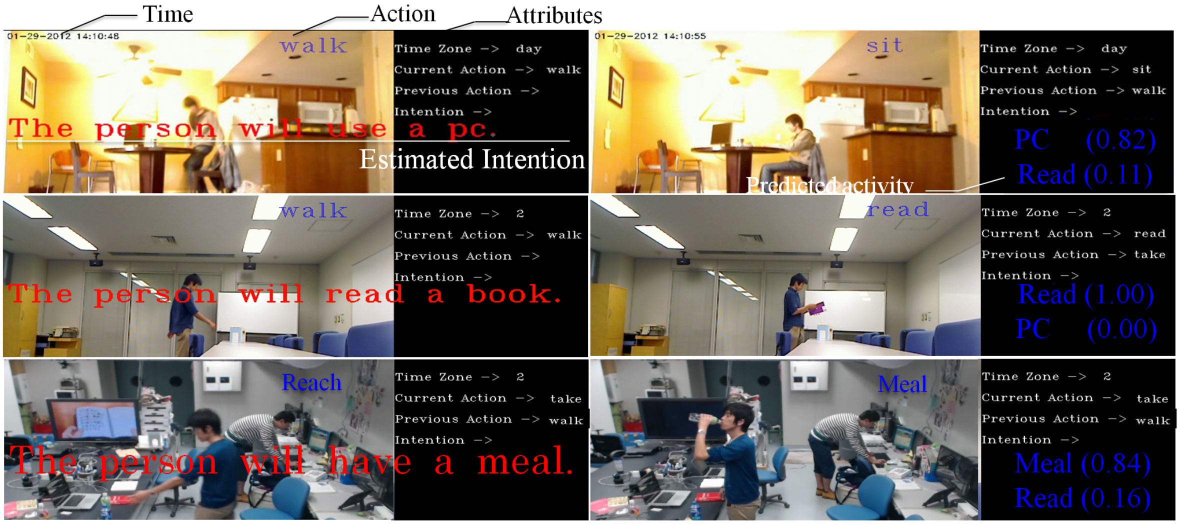


Figure 5: Activity prediction: predicting intention NextActivity = read, PC, meal from given attributes TimeOfDay = morning, day, night, and PreviousAction/CurrentAction = bend, sit, stand, walk. The calculated probabilities are read = 0.11, meal = 0.89 and PC = 0.0. Based on the ranking, “meal” is the estimated intention.

REFERENCES

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. ACM Computing Survey.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. European Conference on Computer Vision Workshop (ECCVW).

Farneback, G. (2003). Two-frame motion estimation based on polynomial expansion. Proceedings of the Scandinavian Conference on Image Analysis.

Jain, M., Gemert, J., and Snoek, C. G. M. (2014). University of amsterdam at thumos challenge2014. World Health Assembly.

Jain, M., Jegou, H., and Bouthemy, P. (2013). Better exploiting motion for better action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kataoka, H., Hashimoto, K., Iwata, K., Satoh, Y., Navab, N., Ilic, S., and Aoki, Y. (2014a). Extended co-occurrence hog with dense trajectories for fine-grained activity recognition. Asian Conference on Computer Vision (ACCV).

Kataoka, H., Tamura, K., Iwata, K., Satoh, Y., Matsui, Y., and Aoki, Y. (2014b). Extended feature descriptor and vehicle motion model with tracking-by-detection for pedestrian active safety. In IEICE Trans.

Kitani, K., Ziebart, B., J., A. B., and M., H. (2009). Activity forecasting. European Conference on Computer Vision (ECCV).

Klaser, A., Marszalek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. British Machine Vision Conference (BMVC).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. NIPS.

Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision (IJCV).

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Li, B., Camps, O., and Sznaier, M. (2012). Cross-view activity recognition using hankelets. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Li, K., Hu, J., and Fu, Y. (2014). Prediction of human activity by discovering temporal sequence patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).

Moeslund, T. B., Hilton, A., Kruger, V., and L., S. (2011). Visual analysis of humans: Looking at people. Springer.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words. British Machine Vision Conference (BMVC).

Pellegrini, S., Ess, A., Schindler, K., and Gool, L. V. (2009). You’ll never walk alone: Modeling social behavior for multi-target tracking. IEEE International Conference on Computer Vision (ICCV).

Peng, X., Qiao, Y., Peng, Q., and Qi, X. (2013). Exploring motion boundary based sampling and spatial temporal context descriptors for action recognition. In BMVC.

Perronnin, F., Sanchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. European Conference on Computer Vision (ECCV).

Raptis, M., Kokkinos, I., and Soatto, S. (2013). Discovering discriminative action parts from mid-level video representation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rohrbach, M., Amin, S., M., A., and Schiele, B. (2012). A database for fine grained activity detection of cooking activities. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ryoo, M. S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. IEEE International Conference on Computer Vision (ICCV).

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv technical report 1409.1556.

Wang, H., Klaser, A., and Schmid, C. (2013). Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision (IJCV).

Wang, H., Klaser, A., Schmid, C., and Liu, C. L. (2011). Action recognition by dense trajectories. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. IEEE International Conference on Computer Vision (ICCV).

Watanabe, T., Ito, S., and Yokoi, K. (2009). Co-occurrence histograms of oriented gradients for pedestrian detection. PSIVT.

(WHO), W. H. O. (2001). The international classification of functioning, disability and health (icf). World Health Assembly.

Zinnen, A., Blanke, U., and Schiele, B. (2009). An analysis of sensor-oriented vs. model - based activity recognition. In ISWC.

APPENDIX

This work was partially supported by JSPS KAKENHI Grant Number 24300078.