Recognition of Transitional Action for Short-Term Action Prediction using Discriminative Temporal CNN Feature

Hirokatsu Kataoka[1]
hirokatsu.kataoka@aist.go.jp
Yudai Miyashita[2]
undermmusic@gmail.com
Masaki Hayashi[34]
m.hayashi@liquidinc.asia
Kenji Iwata[1]
kenji.iwata@aist.go.jp
Yutaka Satoh[1]
yu.satou@aist.go.jp

[1] National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan

[2] Tokyo Denki University
Adachi-ku, Tokyo, Japan

[3] Liquid Inc.
Tokyo, Japan

[4] Keio University
Yokohama, Kanagawa, Japan

Abstract

Herein, we address **transitional actions** class as a class between actions. Transitional actions should be useful for producing **short-term action predictions** while an action is transitive. However, transitional action recognition is difficult because actions and transitional actions partially overlap each other. To deal with this issue, we propose a **subtle motion descriptor (SMD)** that identifies the sensitive differences between actions and transitional actions. The two primary contributions in this paper are as follows: (i) defining transitional actions for short-term action predictions that permit earlier predictions than early action recognition, and (ii) utilizing convolutional neural network (CNN) based SMD to present a clear distinction between actions and transitional actions.Using three different datasets, we will show that our proposed approach produces better results than do other state-of-the-art models. The experimental results clearly show the recognition performance effectiveness of our proposed model, as well as its ability to comprehend temporal motion in transitional actions.

1 Introduction

Transitional actions belong to a class between actions for **short-term action prediction** (see Figure 1). Early action recognition is necessary for producing action predictions in the early frames of an objective action. Earlier prediction in the initial frames of an objective action is desirable for early action recognition problems, but the solutions depend on the action itself. On one hand, within the setting of a short-term action prediction, understanding a pending human action change is more natural if we have a firm grasp on transitional actions. For example, sudden motion changes that are recognized as abnormal or dangerous should be detected early in traffic and surveillance scenes. In a traffic scene, short-term

Pages 12.1-12.12
DOI: https://dx.doi.org/10.5244/C.30.12

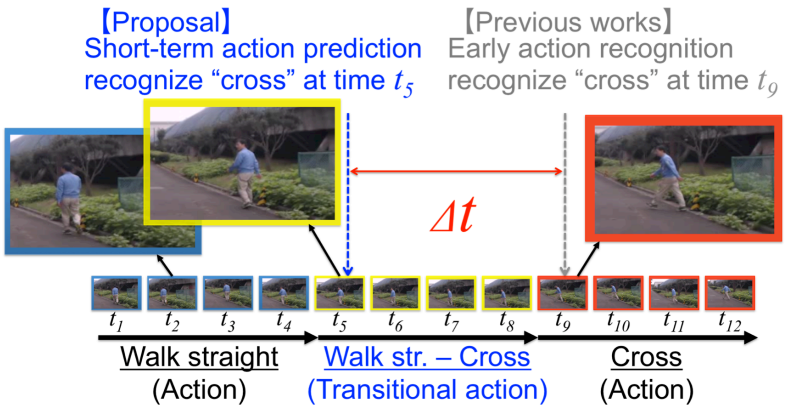


Figure 1: Recognition of transitional actions for short-term action prediction: Our proposal adds a transitional action class *Walk straight - cross* between *Walk straight* and *cross*. Identification of transitional actions allow us to understand the next activity at time t_5 before an early action recognition approach at time t_9.

action predictions are particularly crucial for avoiding accidents between humans and vehicles. Figure 1 shows sequential actions that include *Walk straight*, *Walk straight - cross*, and *cross*. Where *Walk straight* and *cross* are conventional action definitions, our proposal adds a transitional action between actions (here *Walk straight - cross*) in order to provide a better action approach to predictions. Our proposed short-term predictions achieve earlier prediction than so-called early activity recognition, since they can recognize a dangerous *cross* action while it is transitional. Additionally, a prediction approach is likely to be unstable in terms of performance rate. Here, our proposal enables us to replace a recognition-like approach with a more accurate prediction, and we can then assign direct approaches that will be better equipped to handle transitional action classes. However, since transitional actions include subtle differences, they are still difficult to grasp, even when using a state model such as the hidden Markov model (HMM) [23] and the conditional random field (CRF) [16] method. State models tend to focus on a probability distribution for a prior to understand action classes. Intuitively, the recognition difficulty arising from action and transitional action is that they tend to partially overlap each other. We believe that the use of a subtle motion descriptor (SMD) will allow us to identify sensitive differences between actions and transitional actions.

In this paper, we address the recognition of transitional action for short-term action prediction. We also propose a discriminative temporal convolutional neural network (CNN) feature that can be used for recognizing transitional actions in order to overcome the difficulty of indistinguishable feature classification in transitional actions. To accomplish this, we employ an SMD that captures subtle differences between consecutive frames. Our paper contains two main contributions: *(i) the definition of transitional action for short-term action prediction that achieves earlier prediction than early action recognition, and (ii) identifying CNN-based SMD to create a clear distinctions between action and transitional action.* Experimental results show that our proposed descriptor is effective for recognizing transitional actions. Additionally, we will consider a couple of tunings for discriminative temporal CNN, such as late fusion with multi-channel input and parameter optimization. The per-frame feature is based on the Oxford Visual Geometry Group Network (VGGNet) [27] model, which is a deeper CNN architecture.

2 Related works

The first noteworthy work in action recognition is space-time interest points (STIP) [18]. The STIP extends Harris corner detector to time t domain. The STIP is improved in [19], [20] and [4] by using expanded feature representation. However, the best approach for activity recognition to date is arguably dense trajectories (DT) [29], which is based on describing the trajectories of tracked densely sampled feature points. Using these trajectories, the following spatio-temporal features are applied: trajectory histograms of oriented gradients (HOG) [2], histograms of optical flow (HOF) [19], and motion boundary histograms (MBH) [3].

Dense sampling approaches for activity recognition have also been proposed in [8, 12, 28] after the introduction of the first DT. They have incremented DT, for example, by introducing eliminating extra-flow [8], and integrating a higher-order descriptor into the conventional features for fine-grained action recognition [12]. Additionally, Wang *et al.* proposed improved DT (IDT) [28] by executing camera motion estimation, detection-based noise canceling, and adding a Fisher vector [22]. The more recent works have reported achieving state-of-the-art performance with the concatenation of CNN features and IDT in the THUMOS Challenge [6, 9, 34]. Jain *et al.* employed a per-frame CNN feature from layers 6, 7, and 8 with AlexNet [15]. Zhu *et al.* extended both the [34] representations with multi-scale temporal sampling in IDT [17] and video representation in a CNN feature [33]. The combination of IDT and CNN synergistically improve recognition performance.

Recently, CNN features with temporal representations have been proposed [1, 10, 11, 21, 25, 26, 30, 33]. In the first work [1], Baccouche *et al.* combined CNN features from nine sequential frames into recurrent neural networks (RNN) [5]. Their approach attempted to extend convolutional features into evolutional time-series feature with RNN. Ji *et al.* applied three-dimensional (3D) convolution that extracts features from sequential patches at the same patch locations. Such 3D convolution effectively records temporal information from multiple frames at the same time. Karpathy *et al.* improved CNN model from input of multi-resolution image sequences in a slow fusion manner [11]. The most recent work, Ryoo *et al.* clearly outperforms the IDT+CNN with their pooled time series (PoT) that continuously accumulates frame differences between two frames [25]. The feature is simple, but effective, for grasping continuous action sequences. The feature type that should be implemented, however, would improve the representation so that it would adequately fit the transitional action recognition. It is difficult to achieve short-term prediction using the PoT because it describes features from a whole image sequence.

To facilitate recognition of human actions, Ryoo proposed a method for recognizing activities from the early portions of those activities [24]. The integral bag-of-words framework continuously accumulates each frame feature into a histogram that represents time-series human motion for early action recognition. Huang *et al.* has tried to implement kernel-based reinforcement learning for early human-interaction recognition [7]. Their approach predicts an interaction from a person's action with a foreground map and an HOG feature. In the most recent work, Koppula *et al.* proposed a framework for anticipating human actions from robot vision [14]. To accomplish this, they applied an anticipatory temporal conditional random field (ATCRF) that includes object affordance and human sub-activity and their connections in RGBD input. Although these early recognition frameworks are effective, they produce predictions in the early frames of an action. In contrast, our proposal expands the potential of early action recognition by means of transitional action recognition.

Figure 2: A transitional action class is defined as a class between action classes. Transitional actions should be recognized in order to predict human actions while an action is transitive because a symptom of the next action appears in the transitional action. For example, *Put down item* must be predicted at the *Walking - Put down item* state, which is a transitional action. Similarly, *Fetch book* and *Reading* are predicted in *Put down item - Fetch book* and *Fetch book - Reading*, respectively.

3 Transitional action class

Transitional action is defined as the transition class between actions. Figure 2 shows image sequences that include actions and transitional actions (with meshed blue rectangle). Transitional actions should be used to predict human actions while an action is transitive because symptoms of upcoming actions appear in transitional actions. For example, in Figure 2, *Put down item* must be predicted at the *Walking - Put down item* which is a transitional action. Similarly, *Fetch book* and *Reading* are predicted in *Put down item - fetch book* and *Fetch book - Reading*, respectively. However, it is difficult to divide the transitional action class since each class contains neutral elements of previous and upcoming actions. The difficult task is dividing a transitional action from two neighboring actions (e.g., *Fetch book - Reading*, *Fetch book*, and *Reading*, as shown in Figure 2).

Therefore, we must identify the subtle changes in transitional actions using the subdivided primitive motions (described in Section 4). In our experimental section, we investigate the effectiveness of various recognition approaches, including conventional approaches, for the problem of transitional action recognition.

4 Discriminative temporal CNN feature with SMD

Figure 3 shows the flowchart of our proposed temporal CNN representation. Our proposed SMD is used to identify the sensitive differences between action and transitional action. While our proposal is mostly based on [25], we add zero-mean thresholding to the pooled motion feature ($x_i^{\Delta V^{0+}}$ and $x_i^{\Delta V^{0-}}$ in addition to the $x_i^{\Delta V^{+}}$ and $x_i^{\Delta V^{-}}$ in Figure 3). Other temporal CNN features include adding multi-channel input, optimizing zero-mean thresholding ($\pm TH$) with stochastic gradient decent (SGD), and late fusion of the two-stream CNN feature vectors. VGGNet, pre-trained with ImageNet, is assigned as the neural net architecture [27].

An RGB (I) and differential image (I^{diff}) are input into a neural net. The differential image is calculated as follows:

$$I_t^{diff}(x,y) \quad = \quad I_{t+1}(x,y) - I_t(x,y) \tag{1}$$

where x and y are the pixel locations and t is the temporal parameter. The I^{diff} contains two-frame differential motions that efficiently represents moving parts on human body area.

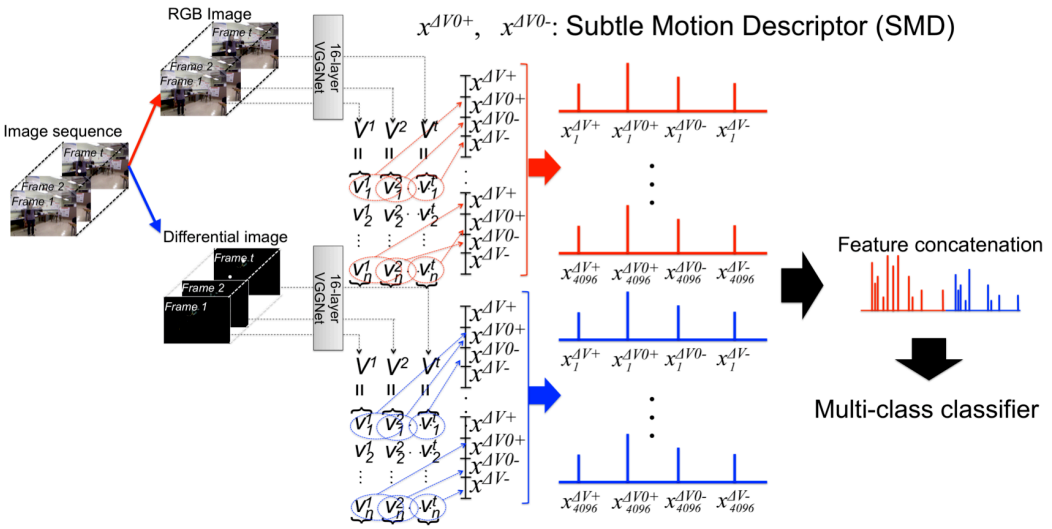


Figure 3: Discriminative temporal CNN feature with SMD for transitional action recognition: Multi-channel input from RGB and differential image is divided into two streams. At each frame, a CNN-based feature (V^t) is extracted with the first fully connected layer of 16-layer VGGNet ($N = 4{,}096$). The consecutive subtractions (ΔV^t) are pooled into four vectors, namely $x^{\Delta V^+}$,$x^{\Delta V^{0^+}}$,$x^{\Delta V^-}$,$x^{\Delta V^{0^-}}$. Here, the $x^{\Delta V^{0^+}}$ and $x^{\Delta V^{0^-}}$ are the proposed SMD. The feature concatenation of RGB and differential image streams is the final classification vector.

Moreover, the use of pixel differential space is effective for extracting CNN features since the image space is similar to the pre-trained ImageNet model. Since the CNN architecture enables per-channel convolution, the RGB channel from differential images are divided into each channel (R, G, B) in the convolution layer.

We apply VGGNet [27] in order to capture per-frame CNN features from an image sequence. The per-frame CNN feature at time t (V^t) is shown below:

$$V^t = [v_1^t, v_2^t, ..., v_N^t] \tag{2}$$

where N is the number of feature vector elements in the CNN feature. Here, we acquire $N = 4{,}096$ when the first fully connected layer is adopted. The frame subtraction ΔV is calculated with a temporal V^t at each frame:

$$\Delta V^t = [(v_1^t - v_1^{t-1}), (v_2^t - v_2^{t-1}), ..., (v_N^t - v_N^{t-1})] \tag{3}$$

The temporal feature accumulation is executed by comparing positive- or negative-number in [25]. Additionally, we insert zero-mean thresholding ($\pm TH$) classes. The temporal accumulated vectors are as follows:

$$x_i^{\Delta V^+} = \Sigma_{t=t_s}^{t_e} h_i^+(t), \quad x_i^{\Delta V^{0^+}} = \Sigma_{t=t_s}^{t_e} h_i^{0^+}(t) \tag{4}$$

$$x_i^{\Delta V^-} = \Sigma_{t=t_s}^{t_e} h_i^-(t), \quad x_i^{\Delta V^{0^-}} = \Sigma_{t=t_s}^{t_e} h_i^{0^-}(t) \tag{5}$$

where

$$\begin{cases} h_i^+(t) = |\Delta v_i^t| \ \ (\Delta v_i^t > TH) \\ h_i^{0^+}(t) = |\Delta v_i^t| \ \ (0 < \Delta v_i^t < TH) \\ h_i^{0^-}(t) = |\Delta v_i^t| \ \ (-TH < \Delta v_i^t < 0) \\ h_i^-(t) = |\Delta v_i^t| \ \ (\Delta v_i^t < -TH) \end{cases} \tag{6}$$

The inside of threshold ($|TH|$; $x_i^{\Delta V^{0^+}}$ and $x_i^{\Delta V^{0^-}}$) indicates subtle differences in the frame subtraction (ΔV) needed to fully grasp motions, including transitional action. Our pooled feature is represented as $x^t = [x^{\Delta V^+}, x^{\Delta V^{0^+}}, x^{\Delta V^-}, x^{\Delta V^{0^-}}]$. The concatenated vector has 16,384 ($= 4{,}096 \times 4$) dimensions.

A couple of feature integration methodologies are listed in [11]. Our proposal is most similar to late fusion. Here, we concatenate two vectors from RGB and differential images as $X^t = [x^t_{RGB}, x^t_{Diff}]$ that is the input of an SVM classifier.

5 Experiments

5.1 Datasets

Three datasets that includes sequential actions, NTSEL, UTKinect-Action (UT), and Watch-n-Patch (WnP), were used in an attempt to understand the transitions between human actions. These datasets are considered to be closely related to the proposal.

The **NTSEL dataset (NTSEL)** [13] contains near-miss events captured by a vehicle. We focused on gradual pedestrian changes *Walking straight-turning*, which are fine-grained activities on real roads. The four activities are *Walking, turning, crossing* and *bicycle riding*.

The **UTKinect-Action dataset (UT) [32].** contains 10 different actions: *Walk, Sit down, Stand up, Pick up, Carry, Throw, Push, Pull, Wave hands,* and *Clap hands*. We re-annotate 8 *transitional action* classes – (1) walk–sit down (# of video is 20) (2) sit down–stand up (#20) (3) stand up–pick up (#20) (4) pick up–carry (#20) (5) carry–throw (#18) (6) throw–push (#18) (7) pull–wave hands (#20) (8) wave hands–clap hands (#20) – in addition to the 10 original classes. The 10 actions are listed from *Walk* to *Wave hands* in order; therefore, we can extract eight classes by excluding *Push-Pull*, which are seldom found in the frames. Totally, 18 actions are included in the re-annotated UT dataset. Although the dataset has depth and kinematic data, our CNN architecture will only consider RGB input.

We chose the office scene from the **Watch-n-Patch dataset (WnP)** [31]. The dataset was also re-annotated and the input was limited to RGB sequences. By adding to the regular 10 classes (*Reading, Walking, Leave-office, fetch-book, Put-back-book, Put-down-item, take-item, Play-computer, Turn-on-monitor, Turn-off-monitor*), we selected 10 more transitional action classes that frequently occur from the dataset – (1) walking–put down item (# of video is 99) (2) take item–leave office (#93) (3) reading–leave office (#77) (4) turn on monitor–play computer (#57) (5) fetch book–reading (#50) (6) put down item–reading (#49) (7) play computer–turn on monitor (#41) (8) walking–reading (#38) (9) put down item–play computer (#38) (10) play computer–turn off monitor (#34)–.

In the moment of re-annotation, each frame of transitional action classes were carefully annotated between 2 regular action classes. In the both datasets, we inserted transitional action classes into a couple of last frames and blank frames. The transitional action classes and regular action classes are partially overlapped each other, but no more than 5 frames overlap. We deleted an action which contains less than 10 frames. Moreover, we deleted a small number of transitional actions in Watch-n-Patch dataset from our experimental dataset (at least 30 at each action).

Table 1: Detailed performance rate of our proposal (late fusion) and PoT [25] on three different datasets.

	% on NTSEL		% on UT		% on WnP	
	10 frm	3 frm	10 frm	3 frm	10 frm	3 frm
Proposal (Late Fusion)	**99.18**	**85.78**	**99.19**	**69.77**	**59.75**	**49.93**
PoT, CVPR2015 [25]	97.00	77.15	92.00	65.46	54.93	44.81

5.2 Settings of evaluation approaches

We began by implementing several state-of-the-art approaches, including the PoT, proposed by Ryoo *et al.* [25]. The CNN feature should be replaced by VGGNet [27], which is one of the most effective neural net architectures. The algorithm is based on PoT - all (Σ+max+Δ2) in their paper [25]. Additionally, we executed the IDT+Per-frame CNN feature (layer-6,7,8) [9] which is the winner of THUMOS Challenge 2014, as a fusion of IDT and CNN features. The other, CNN-based feature includes simple space-time concatenation (ST-CNN). ST-CNN is constructed with $V = [v_1, v_2, ..., v_T]$, where v is the output of VGGNet layer-16 and T is the time-series.

Furthermore, we prepared improved dense trajectories (IDT) with multiple descriptors. The basic IDT [28] consists of trajectory features [29], HOG [2], HOF [19] and MBH [3]. We evaluated shape- and motion-based descriptors in order to grasp their characteristics. For fine-grained action recognition, we conducted the co-occurrence feature descriptors (Co-occurrence HOG (CoHOG) and Extended CoHOG (ECoHOG)) proposed by Kataoka *et al.* [12] in an attempt to capture fine-grained differences in transitional actions. All IDT features are concatenated as IDT (all features- ECoHOG, CoHOG, HOG, HOF, MBH) in this experiment. Since the basic IDT feature accumulation is 15 frames, we set the same accumulation period for our proposal and the other CNN-based approaches.

For a classifier, we used a non-linear SVM with RBF kernel and parameter C was set as 0.03. In multi-class action recognition, we employed one-versus-rest strategy. We assigned leave-one-out cross validation to calculate an accuracy and an error bar. To be more precise, we divided 5 groups at each action class and executed leave-one-group-out cross-validation.

5.3 Parameter tuning

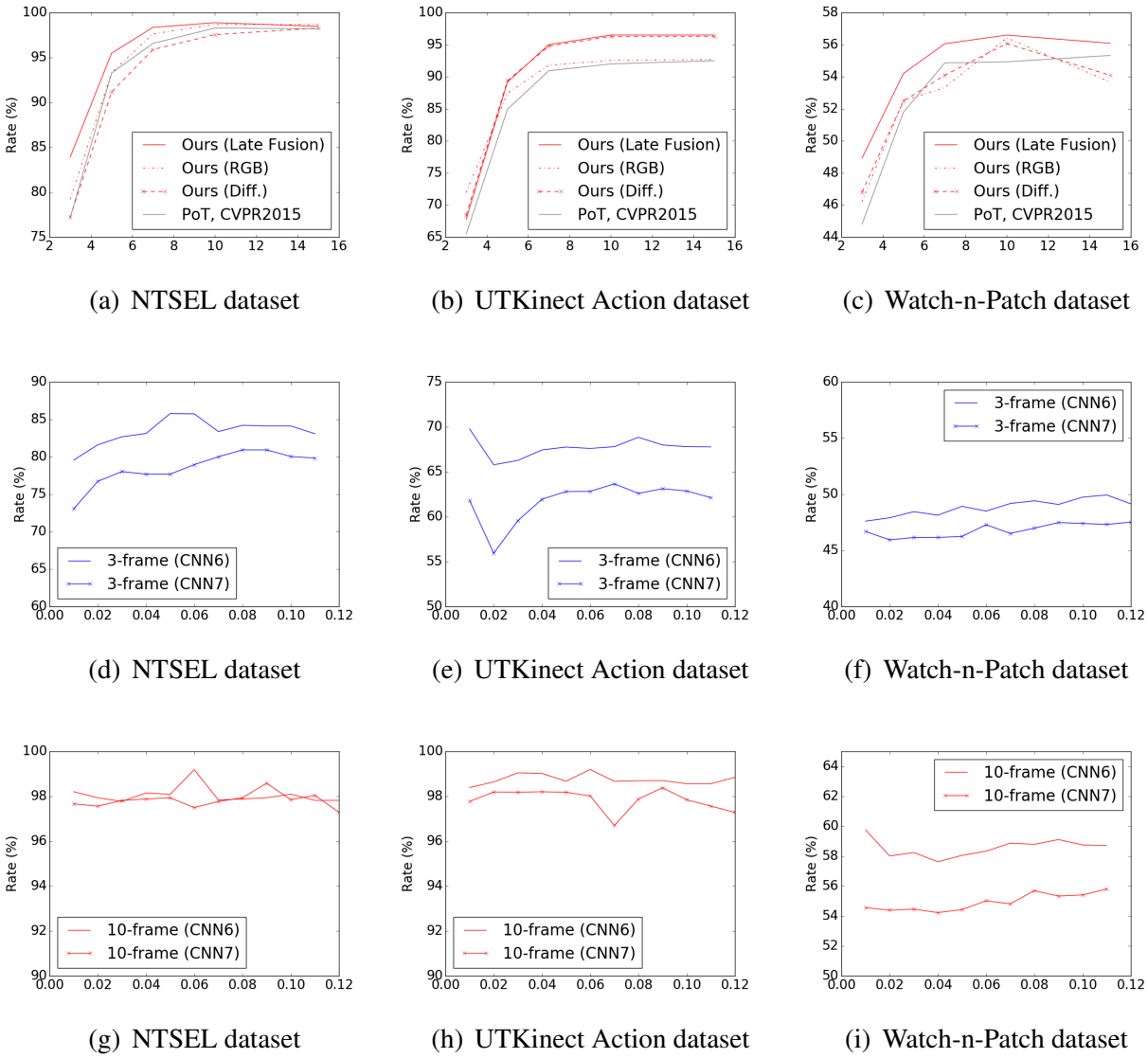
We conducted parameter tuning evaluations on the three datasets. The elements which are listed below:

Late fusion from multi-channel (Figure 4(a) - (c)). The RGB and differential image inputs are seen complementing each other from the result. We assign the late fusion from RGB and differential image.

Frame accumulation (Figure 4(a) - (c)). 3- and 10-frame accumulation are employed. Earlier recognition is better for early prediction since transitional actions directly relate to traffic accidents on the NTSEL dataset. The figures show the 10-frame accumulation reaches a high level of sufficiency at around 10 frames.

Thresholding value (Figure 4(d) - (i)). The thresholding values are depending on data. We use a thresholding value which achieves the top rate at each dataset.

Usage of fully-connected layer (Figure 4(d) - (i)). The first fully-connected layer (CNN6 in the figures) significantly performs better rate than the second fully-connected layer from the result.



(a) NTSEL dataset (b) UTKinect Action dataset (c) Watch-n-Patch dataset

(d) NTSEL dataset (e) UTKinect Action dataset (f) Watch-n-Patch dataset

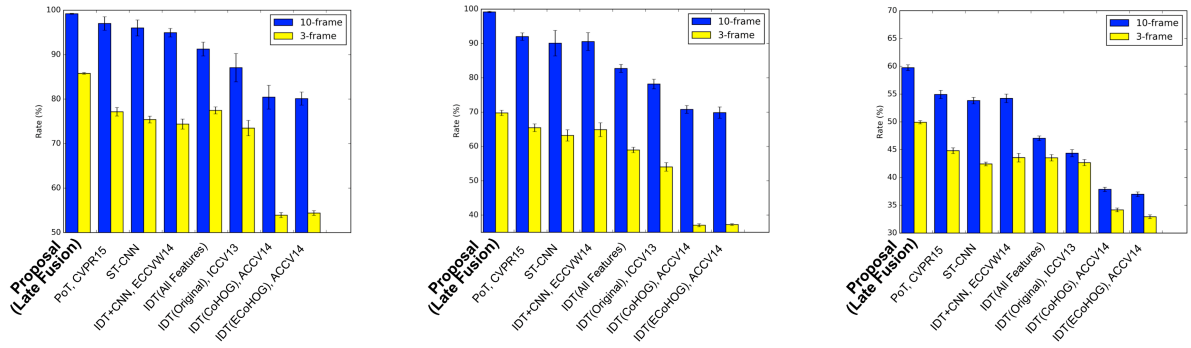
(g) NTSEL dataset (h) UTKinect Action dataset (i) Watch-n-Patch dataset

Figure 4: Parameter tuning. From right to left, the graph indicates NTSEL, UTKinect and Watch-n-Patch dataset. The top row shows frame accumulation with various channels (rgb and differential image) and their fusion, the bottom two rows display thresholding value from two different fully-connected layers (CNN6 and CNN7).

According to the graphs, our proposal shows a higher performance value than the [25]. It should also be noted that 10 or more frames of feature accumulation saturated the performance rate. Therefore, we fixed the feature accumulation at 10 frames. Our proposal shows the top rate on the three datasets used.

5.4 Comparison to state-of-the-art works

Figure 5 shows a comparison of our proposal and state-of-the-art approaches at 3 and 10 frames of feature accumulation. Our proposal records the top rate on the three different datasets contain transitional actions. Basically, the CNN-based approaches outperforms the IDT-based approaches. The IDT+CNN combination shows a better rate than other IDT-based approaches. Although the CNN is a per-frame setting, its appearance helps its performance. More sophisticated settings include our proposal and the PoT. These two approaches pool consecutive subtraction into quantized motions. Space-time tuning is effective for CNN-based action recognition. Next, we will consider our proposal (late fusion) and the PoT [25] for each dataset.



(a) NTSEL dataset (b) UTKinect Action dataset (c) Watch-n-Patch dataset

Figure 5: Comparison of our proposal and state-of-the-art approaches.

Table 1 shows the comparison of our proposal with the PoT [25]. The incrementation between the two approaches is an additional subtle motion difference, $h_i^{0^+}(t)$ and $h_i^{0^-}(t)$ in equation (6), which are the zero-mean small differences with subtraction. The considerations of three datasets are as follows:

NTSEL dataset. The performance rate is increasing +2.18% at 10 frame accumulation and +8.63% at three frame accumulation. This indicates that the proposed SMD is effective for the both settings. Here, we will focus on three frame accumulation. Since pedestrian transitional actions in traffic scenes directly relate to traffic accidents, the result is beneficial for actively promoting safety. The dataset seems to require severe work; however, our proposal achieved 85.78% when using 3-frame accumulation. The result comes from the settings of detected pedestrian rectangles and restricted actions by three actions and one transitional action. The SMD is more effective when a lower feature accumulation level is set in the NTSEL dataset.

UTKinect-Action dataset. The rate is seen increasing +7.19% at 10-frame accumulation and +4.31% at 3-frame accumulation. This dataset includes 10 actions and 8 transitional actions in a single indoor scene. The classification between actions and transitional actions, such as separation *Walk - Sit down* (transitional action), are indistinguishable from *Walk* and *Sit down* (both are actions). Our proposal also shows significant success in classifying various actions and transitional actions.

Watch-n-Patch dataset. The percentage rate increases are +4.82% at 10-frame accumulation and +5.12% at 3-frame accumulation. The performance rate is lower than other two datasets as 59.75% at 10-frame accumulation. It is obvious that transitional action recognition is difficult from the pattern recognition perspective. One reason for this is that the WnP dataset involves various office scenes and camera angles. The feature property is slightly different when using CNN-based feature. Our proposal captures differences in transitional actions in hard situations. Specifically, our proposal achieves 49.93% accuracy for short-term transitional actions prediction of 10 actions within the first three frames. This indicates that our 3-frame accumulation performance rate is outstanding.

6 Conclusion

In this paper, we proposed a definition of transitional action for short-term action prediction. The recognition of transitional action allows us to produce earlier action predictions. Moreover, we also proposed a subtle motion descriptor (SMD) to facilitate recognition of transitional actions in order to identify the sensitive differences between action and transi-

tional actions. Although transitional action classification is a difficult problem, the descriptor successfully divides a transitional action from neighboring actions. Our CNN-based SMD demonstrated the best rate of success on three different trial datasets. Even when using the shortest (3-frame) feature accumulation for recognition tuning, we confirmed outstanding results with 85.78%, 69.77%, and 49.93% on the three different datasets.

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. International Workshop on Human Behavior Understanding (HBU), 2011.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. European Conference on Computer Vision (ECCV), 2006.

[4] I. Everts, J. C. Gernert, and T. Gevers. Evaluation of color stips for human action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[5] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with lstm recurrent networks. *Journal of Machine Learning Research*, 3, 2003.

[6] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2015.

[7] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. European Conference on Computer Vision (ECCV), 2014.

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

[9] M. Jain, J. Gemert, and C. G. M. Snoek. University of amsterdam at thumos challenge2014. European Conference on Computer Vision Workshop (ECCVW), 2014.

[10] S. Ji, W. Xu, M. Yang, and Y. Kai. 3d convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 35 (1), 2013.

[11] K. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

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

[13] H. Kataoka, Y. Aoki, Y. Satoh, S Oikawa, and Y. Matsui. Fine-grained walking activity recognition via driving recorder dataset. IEEE Intelligent Transportation Systems Conference (ITSC), 2015.

[14] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic responce. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems (NIPS), 2012.

[16] McCallum A. Pereira F. Lafferty, J. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. International Conference on Machine Learning (ICML), 2001.

[17] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

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

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

[20] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[21] J. Y.-H Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

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

[23] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 257–286. Proc. of the IEEE, 1989.

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

[25] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[26] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition. Neural Information Processing Systems (NIPS), 2014.

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

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

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

[30] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[31] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[32] L. Xia, C.C. Chen, and J.K. Aggarwal. View invariant human action recognition using histograms of 3d joints. pages 20–27. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012.

[33] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[34] L. Zhu, Y. Yang, and A. G. Hauptmann. Uts-cmu at thumos 2015. CVPR2015 International Workshop and Competition on Action Recognition with a Large Number of Classes, 2015.