LSTC: Boosting Atomic Action Detection with Long-Short-Term Context

Yuxi Li[*]
yukiyxli@tencent.com
Tencent Youtu Lab
Shanghai, China

Boshen Zhang[*]
boshenzhang@tencent.com
Tencent Youtu Lab
Shanghai, China

Jian Li
swordli@tencent.com
Tencent Youtu Lab
Shanghai, China

Yabiao Wang
caseywang@tencent.com
Tencent Youtu Lab
Shanghai, China

Weiyao Lin[†]
wylin@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Chengjie Wang
jasoncjwang@tencent.com
Tencent Youtu Lab
Shanghai, China

Jilin Li
jerolinli@tencent.com
Tencent Youtu Lab
Shanghai, China

Feiyue Huang
garyhuang@tencent.com
Tencent Youtu Lab
Shanghai, China

ABSTRACT

In this paper, we place the atomic action detection problem into a Long-Short Term Context (LSTC) to analyze how the temporal reliance among video signals affect the action detection results. To do this, we decompose the action recognition pipeline into short-term and long-term reliance, in terms of the hypothesis that the two kinds of context are conditionally independent given the objective action instance. Within our design, a local aggregation branch is utilized to gather dense and informative short-term cues, while a high order long-term inference branch is designed to reason the objective action class from high-order interaction between actor and other person or person pairs. Both branches independently predict the context-specific actions and the results are merged in the end. We demonstrate that both temporal grains are beneficial to atomic action recognition. On the mainstream benchmarks of atomic action detection, our design can bring significant performance gain from the existing state-of-the-art pipeline.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

KEYWORDS

video understanding, action detection, long-short-term context

ACM Reference Format:
Yuxi Li, Boshen Zhang, Jian Li, Yabiao Wang, Weiyao Lin, Chengjie Wang, Jilin Li, and Feiyue Huang. 2021. LSTC: Boosting Atomic Action Detection with Long-Short-Term Context. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475374

1 INTRODUCTION

It has been widely studied on how to correctly recognize the human actions from videos, most of the previous efforts [2, 7, 22, 25] resort to temporal information modeling since in some traditional benchmarks [2, 23], different action instances show distinct visual motion patterns across time. However, things are different when it comes to the **atomic** action problem [10, 16]. Atomic action detection aims at localizing persons of extremely subtle behaviors (e.g. *staring at something* or *calling the phone*), where the visual motion information is limited. Under this scenario, visual context plays crucial role because most action instances involve different kinds of interactions, either among actors or between human and objects. Therefore, it is a key factor towards high-quality detection to capture contextual interaction cues in spatiotemporal domain.

There are some attempts to incorporate the spatiotemporal interaction with atomic action detection pipelines [8, 30, 31]. Nevertheless, they ignore the discrepancy between different contextual interactions. Within a short temporal scope, the subject of an action can interact with any space around it, either certain background objects or subjects of another action instance. On the other hand, with a longer temporal distance, the subjects of different action instances are more likely to interact with other subjects, in this case, the background or objects information is less informative to help reason the action types. Figure 1 shows an example of such inference process relying on both scopes of context. In this example, some action types are easy to be inferred at a glance of the short clip (e.g. *sitting* and *writing*), since the background provides cues to support the judgment (the chairs and pen). However, we can not reason that the man is talking with others if not inquiring the action instances from video clips before and after the current

[*] Both authors contributed equally to this research.
[†] Correspondence author.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475374

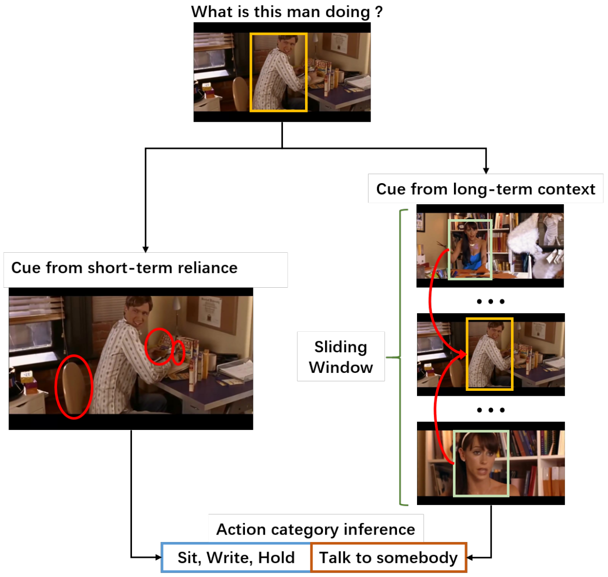

What is this man doing ?
Cue from short-term reliance
Cue from long-term context
Sliding Window
Action category inference
Sit, Write, Hold
Talk to somebody

Figure 1: A example of independent temporal inference of different scope. The left branch shows short-term reliance on surrounding. The right branch illustrates long-term cues from action subjects contained by other clips.

frame. This observation indicates that both short-term and long-term context are crucial to completely recognize the action of a person under complex interaction scenarios. Further, interactions within different time scope are usually independent of each other and can provide complementary cues for the recognition of current action. In the example illustrated in Figure 1, we can see the short-term cue and long-term counterpart are helpful to infer distinct sets of action types, the ignorance of one kind of context cue does not affect the reasoning from other cue. This indicates that context-based atomic action detection can be decoupled into independent inference process of different temporal scope.

With the inspiration above, we build a context-based probability graph model to analyze the **reliance** and **independence** between different variables in atomic action detection problem. Guided by the analysis, we design an atomic action detection pipeline with independent long-short-term reasoning procedure. In short-term reasoning process, we try to aggregate local information from dense context representation with a pixel-wise aggregation mechanism since the action subjects can interact with any specific spatial or temporal fragment within a short time slot. On the other hand, in a longer temporal scope, a discrete actor-wise feature selection and refinement mechanism is designed to gather informative context from a long-term feature bank [30]. Different from the long-term operation in previous works [30, 31] which only models person-to-person interaction and refered as a first-order attention model, we resort to a decoupled second order attention module to exploit person-to-pair relationship with feasible complexity. In this way we ensure that the long-term inference process can independently predict specific class given the context and we demonstrate that such design outperforms the feature-level operation in [30]. We conduct experiments on two benchmarks focusing on atomic action recognition, AVA [10] and HiEve [16], demonstrating that our two-scope inference pipeline can bring significant improvement over existing state-of-the-art methods. In a nutshell, the contribution of this paper can be summarized as:

- The atomic action problem is decoupled as a long-short-term inference task and we propose a parallel reasoning pipeline to solve the problem in both temporal grain.
- A local context aggregation branch is designed to capture helpful information from dense spatiotemporal feature within a short-term scope.
- A high-order long-term attention modeling process is incorporated with our framework to boost its actor-wise long-term reasoning ability.

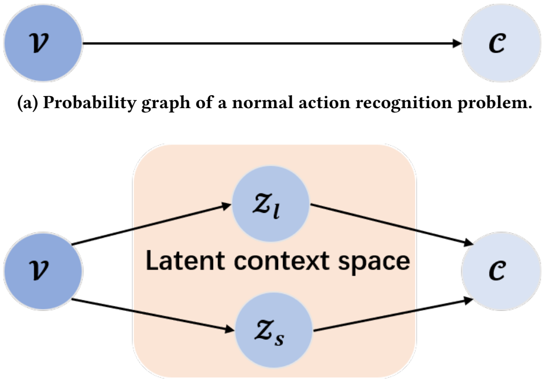
2 RELATED WORKS

2.1 Action recognition

Deep learning technique has pushed the advance of video action recognition to a large extent. Recent works have been trying to design effective CNN architectures for action recognition in short clips [2, 7, 22, 25, 28, 33]. Two-stream ConvNets [22, 28] are designed with spatial and temporal branches to capture the complementary appearance information from still frames and visual motion patterns. 3D-CNN [2, 7, 25] directly model the spatial and temporal information with 3D convolution kernels by inflating the ImageNet [3] pre-trained model and training on input clip of different sampling rate. In addition, there are lines of works resorting to the correlation mechanism at different feature dimension and scale to encode high-level relationship features [27, 29] for accurate action classification. Nevertheless, most of the existing works mentioned above focus on properly capturing the motion within a short temporal span and are more suitable for normal action recognition than subtle atomic-level recognition.

2.2 Atomic action detection

Atomic action detection aims at localizing and recognizing subtle human actions simultaneously in video clips, where the visual motion is not salient and the action subjects are involved in different types of interactions. A large-scale dataset AVA [10] is constructed focusing on solving such challenging problem, accompanied with an actor-centric two-stream model as the baseline. Some recent works follow the object detection paradigm by first localizing person bounding box with pre-trained human detector [20] and utilizing the ROI features from deep features to predict action class [7, 24]. Besides, a short-term graph model [35] is built upon the spatiotemporal location of objects to enhance the performance on complex action discrimination. Finally, there are also some works applying the attention mechanism [26] to capture either short-term dense interaction [8] or discrete long-term relation [30]. However, they solely consider relation under a single time scope, while our work take both long-term and short-term interaction into account and elaborately process them at different level.



(a) Probability graph of a normal action recognition problem.

(b) The probability graph of long-short-term modeling in atomic action detection.

Figure 2: Probability graph model of previous action methods and our long-short-term analysis.

2.3 Long-term context reasoning

Context information plays an important role in highly self-correlated data, in most cases the semantics can be directly inferred from the contextual cues. This property has been widely studied in the field language processing, [1] first proposed a nueral language model to model the concurrent likelihood of certain objective words and their surrounding text. The BERT model [4] further combined this property with attention-based selection scheme for the self-training of language model. There are also some trials in vision problems to incorporate contextual modeling with action recognition [9, 17, 21, 34], but mostly focus on the spatial and short-term context, while accurate atomic action detection requires both short-term and long-term cues in spatiotemporal domain.

3 LONG-SHORT-TERM CONTEXT

3.1 A probability view of context information

Before introducing our framework, we first give a probability formulation of atomic action detection problem. By our definition, $\mathcal{V}$ is the actor-wise representation of an input short video clip and C is the variable indicating the action categories contained by the clip. In traditional action recognition paradigm, the distribution of action class is directly dependent on input clips as $\mathcal{P}(C|\mathcal{V})$ (illustrated as Figure 2a). Nevertheless, we analyze the problem with a more complex model with the hypothesis that the action type distribution is not directly conditioned on $\mathcal{V}$ but indirectly related to it via two intermediate latent variables $\mathcal{Z}_l$ and $\mathcal{Z}_s$, where $\mathcal{Z}_l$ denotes the long-term contextual information and $\mathcal{Z}_s$ is short-term reliance variable. Both latent variables are conditioned on the input clip and together determine the distribution of action class. Figure 2b depicts the probability graph of the reliance between variables defined above.

Since the action type C is not directly related to the input videos, we make attempt to model the joint distribution of tuple $(\mathcal{Z}_l, \mathcal{Z}_s, C)$ conditioned on input as $\mathcal{P}(\mathcal{Z}_l, \mathcal{Z}_s, C|\mathcal{V})$. From Figure 2b, it is easy to draw following conditional independent relationship

$$C \perp \mathcal{V}|(\mathcal{Z}_l, \mathcal{Z}_s) \quad \mathcal{Z}_l \perp \mathcal{Z}_s|\mathcal{V} \tag{1}$$

Hence the joint probability distribution can be decoupled as

$$\begin{aligned}\mathcal{P}(\mathcal{Z}_l, \mathcal{Z}_s, C|\mathcal{V}) =& \mathcal{P}(\mathcal{Z}_l, \mathcal{Z}_s|\mathcal{V})\mathcal{P}(C|\mathcal{Z}_l, \mathcal{Z}_s, \mathcal{V}) \\ =& \mathcal{P}(\mathcal{Z}_l, \mathcal{Z}_s|\mathcal{V})\mathcal{P}(C|\mathcal{Z}_l, \mathcal{Z}_s) \\ =& \mathcal{P}(\mathcal{Z}_l|\mathcal{V})\mathcal{P}(\mathcal{Z}_s|\mathcal{V})\mathcal{P}(C|\mathcal{Z}_l, \mathcal{Z}_s)\end{aligned} \tag{2}$$

In Equation (2) we decouple the joint probability into three terms. $\mathcal{P}(\mathcal{Z}_l|\mathcal{V})$ and $\mathcal{P}(\mathcal{Z}_s|\mathcal{V})$ respectively model the long-term and short-term reliance between video clips and their context, while $\mathcal{P}(C|\mathcal{Z}_l, \mathcal{Z}_s)$ can be regarded as a joint discrimination function to determine the action class according to long-short-term context.

With the discussion above, we design our pipeline with a paradigm of parallel processing and late fusion. The framework is illustrated in Figure 3. In this figure, we utilize the 3D deep neural network to extract feature and take the deep features pooled from detected actor boxes as our actor-wise representation $\mathcal{V}$, where detected actor boxes are obtained via a person detector [20] applied on the center frame of input clip (noted as the "key frame" in the rest of this paper). Next one short-term context branch is designed to aggregate helpful short-term features from the dense spatiotemporal features with the guidance of $\mathcal{V}$. Meanwhile, a long-term context processing branch attempts to mine high-level actor-wise interaction with $\mathcal{V}$ from a longer temporal scope. Finally, the output context $\mathcal{Z}_l$ and $\mathcal{Z}_s$ are fused together to determine final action category C. More detail will be introduced in following sections.

3.2 Short-term local aggregation

Although the person-wise deep feature from ROI-pooling operation [20] can extract helpful information for action detection, more accurate atomic action recognition relies on the surrounding cues of actors. Therefore, we make attempt to aggregate local spatiotemporal information from input clip to build more descriminative short-term action type representation $\mathcal{Z}_s$, which is conditioned on $\mathcal{V}$.

To be specific, the output 3D feature from backbone network is denoted as $\mathcal{X} \in \mathcal{R}^{HWT\times d}$, where H, W, T are the spatial and temporal dimension of the clip after downsampling in the backbone and d is the feature dimension. Suppose the output tensor $\mathcal{V}$ is of size $N \times d$ after the ROI operation, where N is the number of detected actors from the key frame, we take this actor-centric feature as query to generate a spatiotemporal attention map from $\mathcal{X}$ as

$$\mathcal{A}(\mathcal{X}, \mathcal{V}) = softmax\left(\mathcal{V}W_{\mathcal{A}}\mathcal{X}^T\right) \tag{3}$$

where $\mathcal{A}(\mathcal{X}, \mathcal{V}) \in \mathcal{R}^{N\times HTW}$ indicates the dense reliance between actors and its spatiotemporal surrounding and $W_{\mathcal{A}} \in \mathcal{R}^{d\times d}$ is a learnable matrix to indicate the importance of correlation between each dimension pair. The softmax operation is applied as an normalization function over the spatiotemporal dimension of original 3D features. With this attention map, we can aggregate the spatiotemporal context from surrounding feature $\mathcal{X}$ as a person-guided feature $\mathcal{V}_s$

$$\mathcal{V}_s = \mathcal{A}(\mathcal{X}, \mathcal{V})\phi(\mathcal{X}; W_{\mathcal{V}}) \tag{4}$$

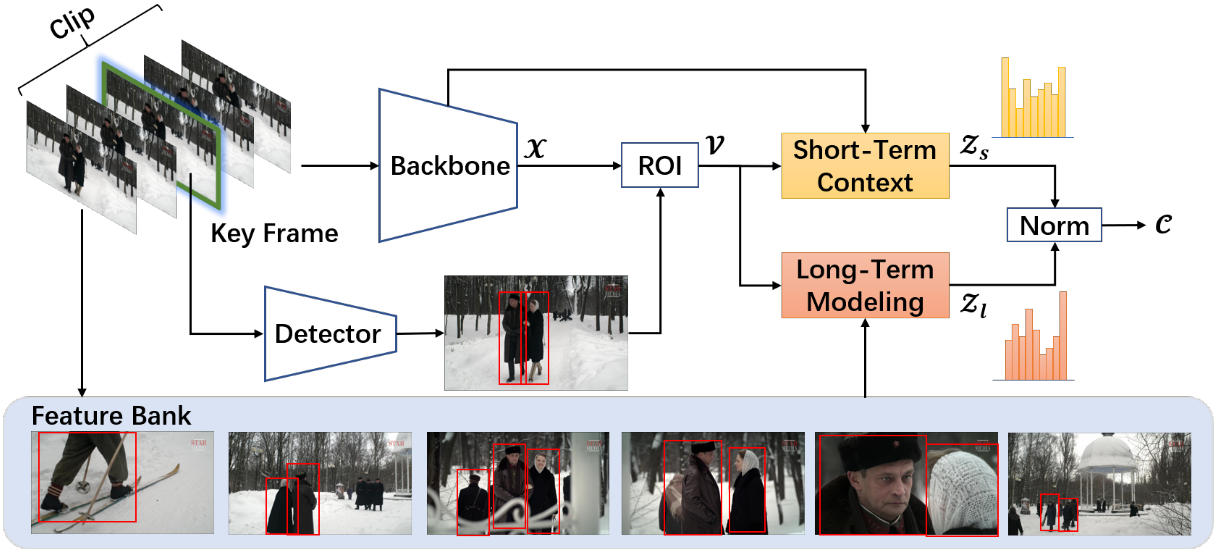


Figure 3: Overview of our pipeline with long-term and short-term context module.

where $\phi(\cdot)$ is a simple linear mapping function parameterized by $W_{\mathcal{V}}$. Since both the actor-centric feature $\mathcal{V}$ and its corresponding surrounding descriptor $\mathcal{V}_s$ encode the action-specific short-term context, we combine them together to get our final short-term local aggregation output as

$$\mathcal{Z}_s = h(FFN([\mathcal{V}, \mathcal{V}_s]; W_f); W_s) \tag{5}$$

where $[\cdot, \cdot]$ denotes the concatenation operation along the channel dimension, *FFN* is short for feed forward network, which is composed of a multi-layer perceptron and W_f is its learnable parameters. Finally the merged feature is processed by a short-term discriminator function $h(\cdot)$ parameterized by W_s and mapped to the short-term context latent space $\mathcal{Z}_s \in \mathcal{R}^c$, where c is the number of classes.

3.3 High order long-term interaction modeling

In addition to the surrounding objects within the short input clip, the long-term interaction with other actors from different video shot can also provide helpful cues for action detection in current clip. To modeling such long-term reliance, attention-based approach is widely used in previous works focusing on video analysis [18, 30, 31], in [29] such attention mechanism is summarized as Equation (6) and implemented with a NonLocal block [29].

$$z_i^{1st} = \frac{1}{\mathcal{T}(v_i)} \sum_{j \in C_i} s\left(v_i, v_j\right) g(v_j) \tag{6}$$

where C_i is the context space of data sample v_i, $s(\cdot, \cdot)$ is a pairwise similarity metric, $g(\cdot)$ is a mapping function, and $\mathcal{T}(v_i) = \sum_{j \in C_i} s\left(v_i, v_j\right)$ is normalization term. Equation (6) can be refered as a **first-order** attention model since it resorts to the correlation to single instance in context space. However, in a more comprehensive scope, the semantic reliance of an object in videos can be not only relative to pairwise relationship, in contrast, the co-occurrence of other instances in context space can also provide important cues for semantic inference [19]. Therefore in our design, we try to introduce a more representative **second-order** attention model to exploit the correlation between single-person and co-occurrence pairs

$$z_i^{2nd} = \frac{1}{\mathcal{T}'(x_i)} \sum_{(j,k) \in C_i \times C_i} s'\left(v_i, v_j, v_k\right) g'(v_j, v_k) \tag{7}$$

where $\mathcal{T}'(v_i) = \sum_{(j,k) \in C_i \times C_i} s'\left(v_i, v_j, v_k\right)$, however, calculating the second-order attention in Equation (7) results $O(|C_i|^2)$ complexity for each person, which is infeasible when context space is large. Inspired by the attempts in tensor decoupling [11], we approximate such attention with decoupling to achieve $O(|C_i|)$ complexity

$$\begin{aligned} z_i^{2nd} &\approx \frac{1}{\mathcal{T}'(x_i)} \sum_{(j,k) \in C_i \times C_i} s_1'\left(v_i, v_j\right) s_2'\left(v_i, v_k\right) g_1'(v_j) g_2'(v_k) \\ &= z_{i,1}^{1st} z_{i,2}^{1st} \\ z_{i,l}^{1st} &= \frac{1}{\sum_j s_l'(v_i, v_j)} \sum_{j \in C_i} s_l'\left(v_i, v_j\right) g_l'(v_j) \quad l = 1, 2 \end{aligned} \tag{8}$$

Via Equation (8), we decouple the second-order attention into the form of multiplication between output of two first-order NonLocal blocks. Combining the attention in different order z_i^{1st}, z_i^{2nd} we can build our long-term interaction module as Figure 4, where we follow the experience of multi-head and cascaded attention [26]. To be specific, sequentially we cascade K Reader Units (RU) to recurrently extract and refine long-term reliance to get output long-term representation $\mathcal{Z}_l$. Within each Reader Unit, we apply M parallel NonLocal pairs and utilize learnable weight parameters β_m to aggregate the results to form approximate second-order attention. In implementation, we take the long-term feature bank [30] as the context space for each person from the input clip.

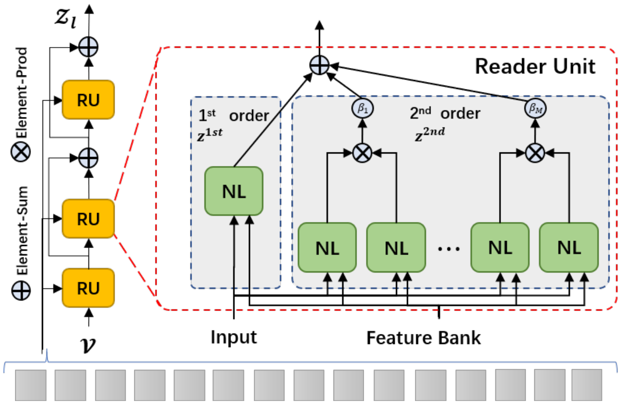


Figure 4: Detailed structure of long-term modeling process. Where "RU" is short for the reader unit and "NL" is short for NonLocal block.

3.4 Joint discrimination

In the design above, either $\mathcal{Z}_s$ or $\mathcal{Z}_l$ can be regarded as action estimation in the latent space, while in a global sense, the action category of specific person is determined by both long-term and short-term context, therefore we simply incorporate these two types of context of two temporal scope with a late-fusion mechanism

$$C = Norm(\mathcal{Z}_l + \mathcal{Z}_s) \tag{9}$$

where *Norm* is a normalization function to map the context variable to class probability space $[0, 1]^c$. During the training stage, the output distribution is supervised by the given action labels via cross entropy loss function.

4 EXPERIMENTS

4.1 Experimental settings

4.1.1 Datasets. - **Atomic Visual Action (AVA)** [10] is the first dataset constructed for spatiotemporal detection of subtle atomic actions including 80 categories of actions in total. This dataset consists of 235 long movie sequences for training and other 64 for validation. The videos are annotated with boxes on each frame at a frequency of 1FPS, each actor is associated with one or more action labels.

- **Human in Events (HiEve)** [16] is a recently public benchmark towards comprehensive analysis on surveillance video in events. Different from AVA, it contains more crowded daily and emergency scenarios. The dataset includes 14 different action classes and consists of 32 video sequences (19 for training and other 13 for testing). Each sequence last for about 2 minutes on average, these videos are annotated with bounding boxes and action labels every 20 frames.

4.1.2 Evaluation protocol. - **Frame-mAP**. For both benchmark, we report the frame-level mAP value as the object detection tasks [5, 15] for performance evaluation, if an action proposal has overlap larger than δ with a ground-truth box and has the same action label, it is regarded as positive proposal. In AVA, δ is set as 0.5, following [10], we only evaluate the 60 most common action classes. In HiEve, we follow [16] to compute mAP score under $\delta = [0.5, 0.6, 0.75]$ and calculate the average score as final performance. Besides, we also report weighted-mAP in HiEve, which focuses more on complex and crowd scenes in videos and assign larger evaluation weight to frame with large crowd index.

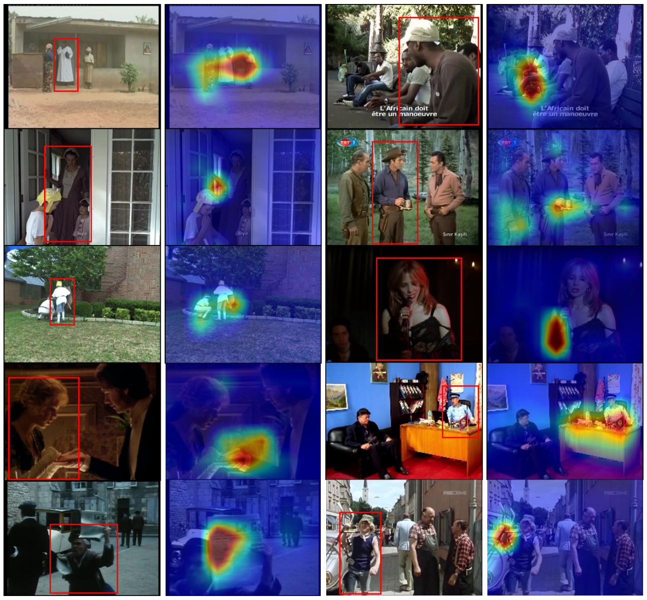


Figure 5: Attention heatmap from short-context module. The left column indicates the detected person in red box, the right column shows its corresponding spatiotemporal attention. (best viewed in color)

4.1.3 Implementation detail. - **Person detector.** For AVA, We apply the person detector from [30] to detect actors on key frames, which is a modified version of FPN-Faster-RCNN [14, 20] and pretrained on COCO Keypoints [15], the detector is then finetuned on AVA. For HiEve, we directly use the public detected results from the multi-object tracking track of HiEve challenge[1].

- **Baseline model.** In our experiments, we take the SlowFast [7] network with ResNet [13] backbone and Inflated-3D network (I3D) [29] as our base network for spatiotemporal feature extraction. All these base model in our experiments are first pretrained on the Kinetics dataset [2] and then finetuned on the target benchmark. In our implementation, we append an additional 1 × 1 × 1 convolution after the network trunk to reduce the channel size by half. For the ROI operation, we separately apply average and max pooling along the temporal dimension of the output 3D feature to get two 2D spatial feature, then take ROIAlign [12] on each spatial feature, the pooled feature vectors are summed together as person-specific feature. In a baseline setting, this feature is directly sent to an action classifier.

- **Training detail.** We implement our framework with the PySlowFast platform [6]. We select each annotated frame as the key frame, and uniformly sample 32 frames centered on this frame with a sampling rate of 2×, this results in a short input clip with temporal endurance of around 2 seconds. The training process is splitted into two stage, first we train a base model without long-term context

[1] http://humaninevents.org

backbone	component		mAP@0.5
	$\mathcal{Z}_s$	$\mathcal{Z}_l$	
I3D			15.8
	✓		20.7
	✓	✓	**22.4**
Res50			24.6
	✓		26.1
	✓	✓	**28.7**
Res101-NL			27.2
	✓		27.7
	✓	✓	**30.3**

Table 1: Ablation study on the directly effect of long-short-term context on AVA v2.2 validation set.

modeling, and extract the ROI feature into long-term feature bank, then in the second stage, we train the network with both long-term and short-term context. During training, we take both the annotated bounding boxes and detected person boxes with confidence score larger than 0.9 as box proposals to extract person-wise ROI feature to bridge to localization gap between person detector and groundtruth. The network is trained on 8 GPUs with a total batch size of 64 clips. The learning rate is initialized with a warmup value of 1.25×10^{-4} and linearly increases to 0.1 in 5 epochs, then the learning rate is degraded by a factor of 10 every 5 epochs. The network parameters are optimized by SGD algorithm with a weight decay ratio of 10^{-7}. Similar to [7], we apply random flip and crop to augment the data.

4.2 Ablation studies

We conduct our ablation studies on AVA to analysis the effect of each component in our framework. If not specified, we take v2.2 for both training and evaluation in default.

(1). How does each context component directly affect the final results ? To investigate the effectiveness of long-term and short-term context, we conduct experiments with baseline models of different backbones. We start from a baseline model with a simple ROI operation followed by a classier, and then test on our trained model from the first training stage without long-term context, finally report results on full model. The corresponding performance is listed in Table 1. From the table we observe that both long-term interaction modeling and short-term local aggregation can bring improvement in overall detection score regardless of the backbone we used, where our full model can achieves at most **+6.6** higher performance than the simple baseline. We find that the attention-based local aggregation is less helpful on the SlowFast with Res101-NL backbone than I3D and SlowFast-Res50, this can be due to the fact that the inserted NonLocal block already aggregate some global and surrounding information from the clip in the feature extraction stage. In contrast, we find the short-term improvment is more salient when the backbone is weaker.

In Figure 7, we visualize the per-class AP value in terms of the output from SlowFast-Res50 backbone with different settings. We find the short-term context is more helpful to recognize actions with semantics highly related to the scene or surrounding objects (e.g. *listen to* and *climp*). On the other hand, the long-term interaction modeling is benifical especially for actions involving group activities or with long endurance (e.g. *dance* and *swim*). Besides, we visualize the person-guided spatiotemporal attention map in Figure 5, we can see, our short-context aggregator can properly catch some critical part from the short-term clip for corresponding action recognition.

context source		mAP@0.5	Increment
short-term	long-term		
Res50	Res50	28.7	-
Res50	Res101-NL	30.0	+ 1.3
Res101-NL	Res50	29.4	+ 0.7
Res101-NL	Res101-NL	**30.3**	**+ 1.6**

Table 2: Investigation on different combination of long-short-term context sources.

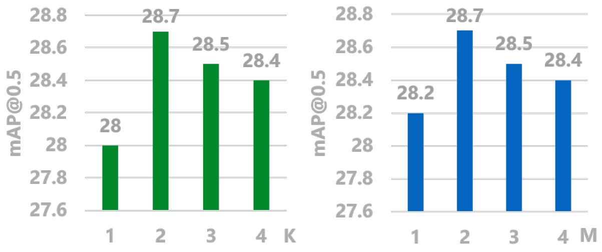


Figure 6: Invstigation on hyperparameter K and M.

(2). How does the parameter K and M affect the performance ? Next we investigate the optimal configuration (K for Reader Unit number and M for number of second attention head) for long-term modeling with the SlowFast-res50 backbone. The results are reported in Figure 6. We see that both parameters achieves optimal results at the value of 2, when K&M increase, the results are degraded. This is probably due to more complicated model structure harmful to model training.

(3). What if the long and short context are from different backbones ? In our normal setting, the long-short-term context is extracted from the same backbone, in Table 2, we investigate the effect of different context source combination. In detail, we adopt the Res50 and Res101-NL as base models, taking one for long-term feature bank extraction and the other as short-term backbone to process input video clip. The results are compared with our full model with either backbone. From Table 2 we observe that when the quality of either context is improved, there will be obvious enhancement in recognition results. Further, we see improving long-term context brings more obvious benefits (+1.6), demonstrating that our designed high-order attention can effectively harness the long-term reliance.

(4). Comparison with the Feature Bank Operator (FBO) [30]. Some other methods to deal with long-term features are proposed, within which the FBO [30] is the most similar counterpart to ours, which attempts to fuse the long-term information in feature representation with a first-order attention model. We compare this operation with our high-order modeling part on different backbones. To do this, we take the backbone with pure short-term context as a

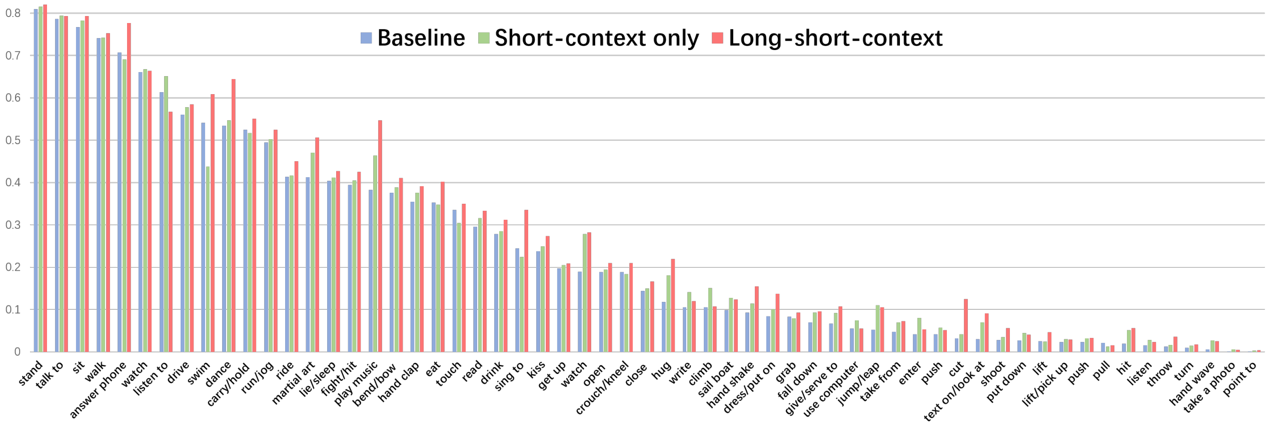


Figure 7: AP@0.5 value for each action class on AVA v2.2 validation set in the descending order of baseline performance, obtained from models with SlowFast network and Res50 backbone. (best viewed in color)

backbone	setting	long-term order	mAP@0.5
Res50	short-term only	zero order	26.1
	+ FBO [30]	first order	27.5
	+ LSTC (ours)	second order	**28.7**
Res101-NL	short-term only	zero order	27.7
	+ FBO [30]	first order	29.4
	+ LSTC (ours)	second order	**30.3**

Table 3: Comparison results between our long-term modeling and FBO [30] on AVA v2.2 validation set.

method	backbone	pretrain	mAP@0.5
Baseline [10]	I3D	Kinetics-400	15.6
ACRN [24]	S3D	Kinetics-400	17.4
VAT [8]	I3D	Kinetics-400	25.0
SMAD [35]	I3D	Kinetics-400	22.2
SlowFast [7]	Res50	Kinetics-400	24.2
SlowFast [7]	Res101-NL	Kinetics-600	27.3
LFB [30]	Res50-NL	Kinetics-400	25.8
LFB [30]	Res101-NL	Kinetics-400	27.7
C-RCNN [31]	Res50-NL	Kinetics-400	28.0
LSTC (ours)	Res50	Kinetics-400	28.4
LSTC (ours)	Res101-NL	Kinetics-600	**30.0**

Table 4: Comparison results with other methods on AVA v2.1 validation set.

method	backbone	pretrain	mAP@0.5
SlowFast [7]	Res50	Kinetics-400	24.9
SlowFast [7]	Res101-NL	Kinetics-600	29.2
AVSF [32]	Res50	Kinetics-400	25.9
AVSF [32]	Res101-NL	Kinetics-400	28.6
LSTC (ours)	I3D	Kinetics-400	22.4
LSTC (ours)	Res50	Kinetics-400	28.7
LSTC (ours)	Res101-NL	Kinetics-600	**30.3**

Table 5: Comparison results with other methods on AVA v2.2 validation set. The SlowFast results are reported in the official code repository [6].

method	backbone	mAP	w-mAP
RPN+I3D [16]	I3D	8.3	6.8
VAT [8]	I3D	7.0	7.3
SlowFast [7]	Res50	7.4	5.3
LSTC (ours)	Res50	**8.9**	**7.4**

Table 6: Comparison results on HiEve test set. The values of mAP and w-mAP are averaged over all thresholds.

zero-order model, and append first-order (FBO) and second-order (LSTC) attention respectively. In Table 3 we list the comparison results on Res50 and Res101-NL backbone, we observe that either backbone benefit more from our long-term modeling mechanism than FBO, this indicates that our scheme can capture the long-term reliance more sufficie ntly.

4.3 Comparison with state-of-the-art methods

In Table 4 and Table 5, we list the comparison results on the standard AVA benchmarks with other methods. It can be observed that on both version of AVA validation set, our method outperforms most of other methods. Especially, with similar backbone and pretraining settings, our LSTC can outperform the SlowFast [7], LFB [30] and C-RCNN [31] counterpart with marginal computation cost (SlowFast processes 28.6 clips per second while ours achieves a processing rate of 27.5 clips in each second). It is also noticeable that the performance of our model with *solely short-term context* (25.6 on v2.1 and 26.1 on v2.2) is still better than SlowFast [7], AVSlowFast [32] and comparable to LFB [30]. These comparison demonstrate that our LSTC is well suitable for atomic action detection. In Table 6, we compare our scheme with other methods on recently public HiEve datasets. It is observed that our LSTC outperforms other methods in both mAP and w-mAP with similar backbone network, indicating that our approach can properly handle person-level action detection in crowd surveillance scenes.

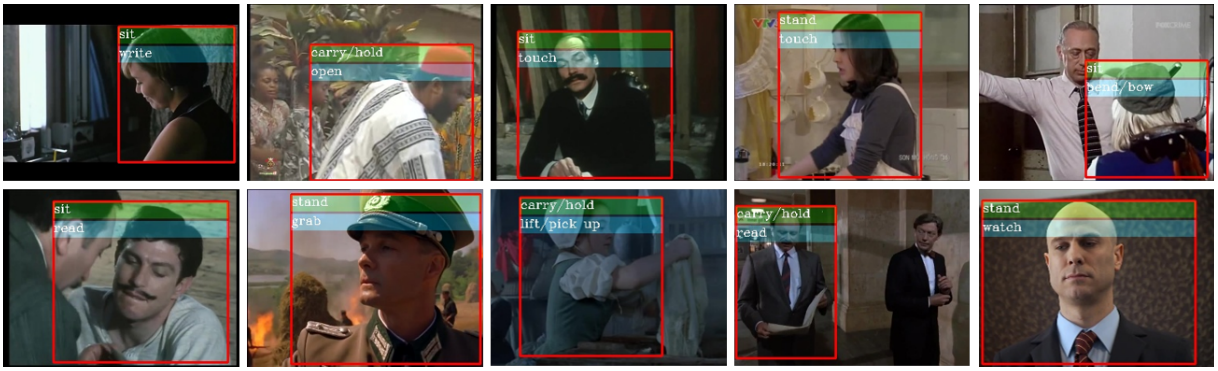


Figure 8: Qualitative results with most confident prediction via only short-term context (in green rectangle) and only long-term context (in blue rectangle, best viewed in color).

4.4 Qualitative results

In Figure 8, we investigate the independent prediction results from either short-term or long-term context. To be specific, we only pass $\mathcal{Z}_s$ or $\mathcal{Z}_l$ to normalization function and generate final classification C. We visualize the bounding boxes together with the most confident action type obtained from each temporal scope. From Figure 8, we find the long-term and short-term context can provide complementary cues for prediction, thus the most confident action type from each branch can be different in most cases. The short-term context will guide the classifier to prefer pose or short interaction, while long-term context leans to some actions not captured completely in current shot but lasting for a relatively long time span, which can be inferred from the video segments before and after current timestamp.

5 CONCLUSION

In this paper, we analyze the atomic action detection problem from the aspects of both long-term and short-term temporal span. We first build a context-based probability graph model and derive the conditional independence between two types of context. Guided by this conclusion, we decouple the task into two inference procedure with different spatiotemporal cues for atomic action recognition. For short-term context, we introduce dense attention to aggregate cues, for long-term information, second-order attention is designed to well handle long-term reliance. Results on challenging benchmarks demonstrate our LSTC is beneficial to high-quality action detection.

6 ACKNOWLEDGEMENT

This paper is supported in part by the following grants: National Key Research and Development Program of China Grant (No.2018AAA0 100400), National Natural Science Foundation of China (No. 61971277).

REFERENCES

[1] Y. Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2006. *Neural Probabilistic Language Models.* Vol. 3. 137–186.

[2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6299–6308.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision (IJCV)* 111, 1 (Jan. 2015), 98–136.

[6] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. 2020. PySlowFast. https://github.com/facebookresearch/slowfast.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6202–6211.

[8] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 244–253.

[9] Rohit Girdhar and Deva Ramanan. 2017. Attentional pooling for action recognition. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*. 34–45.

[10] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6047–6056.

[11] Jianbo Guo, Yuxi Li, Jianguo Li, and Weiyao Lin. 2018. Network Decoupling: From Regular Convolution to Separable Depthwise Convolution. In *BMVC*.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2117–2125.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*. Springer, 740–755.

[16] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Guo-Jun Qi, Rui Qian, Tao Wang, Nicu Sebe, Ning Xu, Hongkai Xiong, and Mubarak Shah. 2020. Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events. arXiv:2005.04490 [cs.CV]

[17] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. 2009. Actions in context. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2929–2936.

[18] S. W. Oh, J. Y. Lee, N. Xu, and S. J. Kim. 2020. Space-time Memory Networks for Video Object Segmentation with User Guidance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2020), 1–1.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*. 91–99.

[21] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015).

[22] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*. 568–576.

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[24] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. 2018. Actor-centric relation network. In *Proc. European Conference on Computer Vision (ECCV)*. 318–334.

[25] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. 4489–4497.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[27] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. 2020. Video Modeling With Correlation Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 20–36.

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7794–7803.

[30] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 284–293.

[31] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. 2020. Context-aware rcnn: A baseline for action detection in videos. In *European Conference on Computer Vision (ECCV)*. Springer, 440–456.

[32] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. 2020. Audiovisual SlowFast Networks for Video Recognition. arXiv:2001.08740 [cs.CV]

[33] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proc. European Conference on Computer Vision (ECCV)*. 305–321.

[34] Bangpeng Yao and Li Fei-Fei. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 17–24.

[35] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. 2019. A structured model for action detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9975–9984.