Huifen XIA, Yongzhao ZHAN, Honglin LIU, Xiaopeng REN, 2024. Enhancing action discrimination via category-specific frame clustering for weakly-supervised temporal action localization. *Frontiers of Information Technology & Electronic Engineering*, 25(6):809-823. https://doi.org/10.1631/FITEE.2300024

Enhancing action discrimination via category-specific frame clustering for weakly-supervised temporal action localization

Key words: Weakly supervised; Temporal action localization; Single-frame annotation; Category-specific; Action discrimination

Corresponding author: Yongzhao ZHAN
E-mail: yzzhan@ujs.edu.cn
ORCID: https://orcid.org/0000-0001-7475-2895

Motivation

- Existing weakly-supervised temporal action localization (W-TAL) methods of single-frame annotation use only video snippet sequences to model action or background and ignore the full utilization of the action discrimination of annotated frames, which makes the class activation sequence for TAL insufficiently robust. This may be the main cause of inaccurate and imprecise detection of action instances.
- We find that the annotated frames of the same category are the discriminative action frames, which have distinctive appearance characteristics or clear action patterns.

Main idea

- A novel method to enhance action discrimination via category-specific frame clustering is proposed for W-TAL. Specifically, the *K*-means clustering algorithm is employed to aggregate the annotated distinctive frames of the same category, and then category-specific feature representation is obtained. It can provide complementary guidance to enhance action discrimination for video snippet sequence modeling.
- A convex combination fusion mechanism is presented, which ensures the consistency of action discrimination between the annotated frames and video snippet sequence. Then, we can obtain a more robust class activation sequence for precise action classification and localization.

Framework

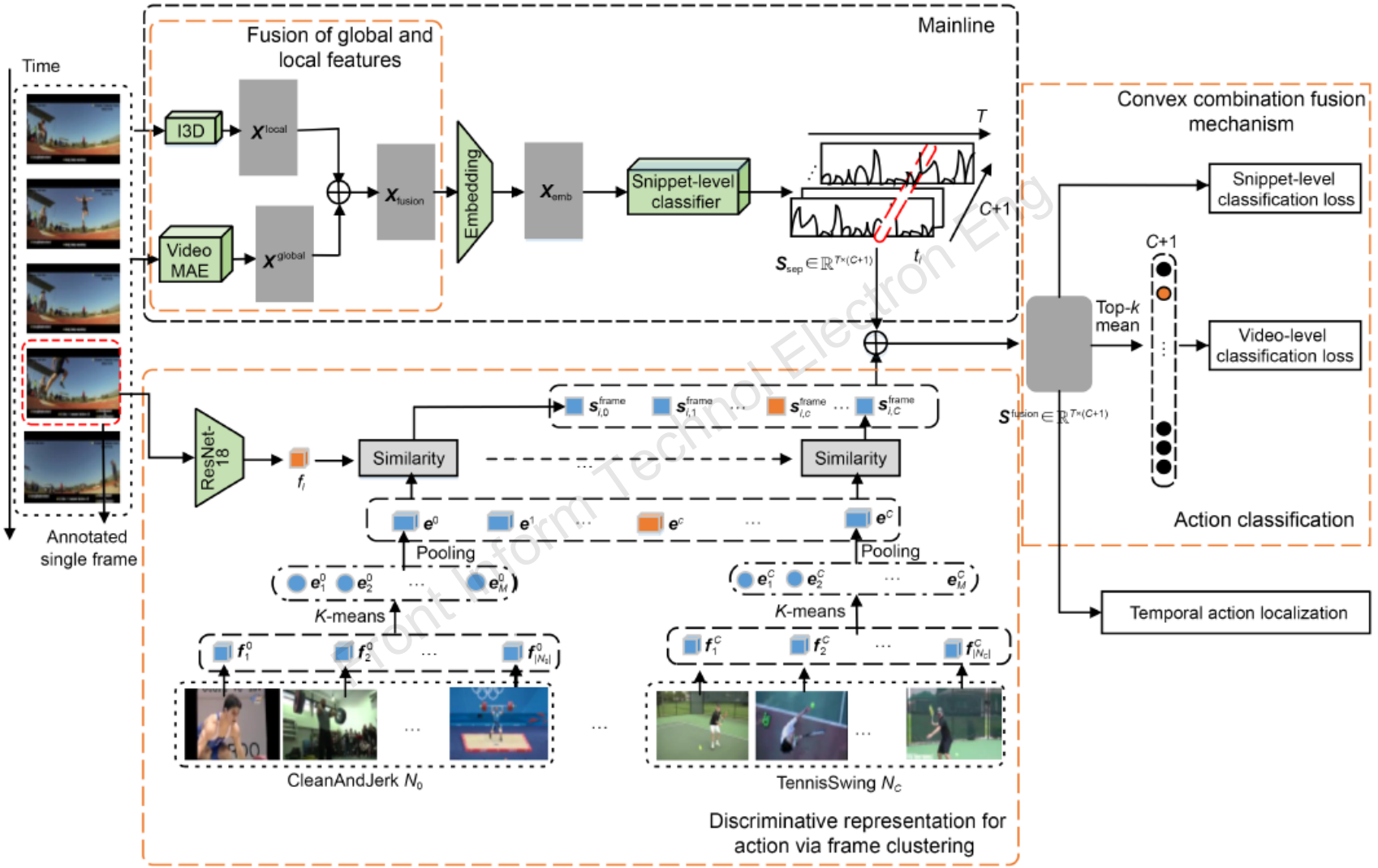

Fig. 1 Overall framework of our proposed method. It consists of the mainline, discriminative representation for action via frame clustering, and convex combination fusion mechanism of annotated frames and video snippet sequences. A robust class activation sequence is generated for precise action classification and localization

Method

- Specifically, we divide the annotated frames of the same category into *M* typical characteristics by the *K*-means clustering algorithm, where *M* is a hyper-parameter.
- Then, we take the average pooling on the *M* typical characteristics, which are used to represent the feature representation of each category.
- On this basis, the TAL task is formulated as comparing a frame with an exemplar of each category. By calculating the similarities, we obtain the class activation scores.
- As a result, modeling action discrimination enhancement can provide complementary guidance information to the video snippet sequence of the mainline and make the TAL algorithm robust to noise within the local video snippets.

Conclusions

A novel method to enhance action discrimination via category-specific frame clustering is proposed for W-TAL. Specifically, we make full use of annotated frames in the video to enhance the action discriminative representation. Since all the annotated frames in the video are discriminative, we cluster the representative annotated frames from the same category to obtain category-specific representations. The single-frame-level class activation score is generated by calculating the similarities between the frame and various categories. Then, a convex combination fusion mechanism between the annotated frames and video snippet sequences is presented to ensure the consistency of action discrimination for generating a robust class activation sequence.

Conclusions

Experiments conducted on three popular datasets validate that single-frame image modeling can provide complementary guidance information to the video snippet sequence, and our method outperforms state-of-the-art methods. Our method can be effectively applied to untrimmed/trimmed videos, which have the same action categories and similar scenes. When the scene changes greatly and the action categories are different, the model needs to be retrained before it is applied.

Huifen XIA received her MS degree in system engineering from Jiangsu University in 2008 and is currently pursuing a PhD in this area at the School of Computer Science and Communication Engineering in Jiangsu University, China. She has high interests in machine learning and deep learning. Her research interests include computer vision, video understanding, and especially video temporal action localization.

Yongzhao ZHAN received the BS degree from Fuzhou University, China, in 1984, the MS degree from Jiangsu University, China, in 1990, and the PhD degree from Nanjing University, China, in 2000, all in computer science. He is currently a professor with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include big data, multimedia, and the Internet of Vehicles. He was a recipient of the Science and Technology Progress Award from the Government of Zhenjiang, in 2006, and from the Government of Jiangsu, in 2013.