Temporal Segmentation of Fine-grained Semantic Action: A Motion-Centered Figure Skating Dataset

Shenglan Liu[1], **Aibin zhang**[1,*], **Yunheng Li**[1,*],
Jian Zhou[1], **Li Xu**[2], **Zhuben Dong**[1], **Renhao Zhang**[1]
[1] Dalian University of Technology, Dalian, Liaoning, 116024 China
[2] Alibaba Group
liusl@dlut.edu.cn, renwei.xl@alibaba-inc.com

Abstract

Temporal Action Segmentation (TAS) has achieved great success in many fields such as exercise rehabilitation, movie editing, etc. Currently, task-driven TAS is a central topic in human action analysis. However, motion-centered TAS, as an important topic, is little researched due to unavailable datasets. In order to explore more models and practical applications of motion-centered TAS, we introduce a Motion-Centered Figure Skating (MCFS) dataset in this paper. Compared with existing temporal action segmentation datasets, the MCFS dataset is fine-grained semantic, specialized and motion-centered. Besides, RGB-based and Skeleton-based features are provided in the MCFS dataset. Experimental results show that existing state-of-the-art methods are difficult to achieve excellent segmentation results (including accuracy, edit and F1 score) in the MCFS dataset. This indicates that MCFS is a challenging dataset for motion-centered TAS. The latest dataset can be downloaded at https://shenglanliu.github.io/mcfs-dataset/.

Introduction

Temporal action segmentation (TAS) has been widely used in sports competitions (Urban and Russell 2003), exercise rehabilitation (Lin and Kulić 2013), movie editing (Magliano and Zacks 2011) and other fields. Technically, TAS has been extended to many new topics, such as video action localization (Lee, Uh, and Byun 2020) and moment retrieval (Zhang et al. 2019) etc. In recent years, TAS has made remarkable progress on task-based video, especially in designing new temporal convolution networks (TCN) (e.g. Encoder-Decoder TCN (ED-TCN) (Lea et al. 2017), Multi-Stage Temporal Convolutional Network (MS-TCN) (Farha and Gall 2019) and Self-Supervised Temporal Domain Adaptation (SSTDA) (Chen et al. 2020a)) which achieve higher performance on cooking task datasets (e.g. GTEA (Fathi, Ren, and Rehg 2011), 50Salads (Stein and Mckenna 2013) and Breakfast (Kuehne, Arslan, and Serre 2014), etc.).

However, it can be found that the existing datasets have three limitations for TAS research, which can be mainly summarized as follows.

Coarse-grained semantics. The coarse-grained TAS is relatively easy for the existing models. However, it is difficult to meet the related applications of fine-grained semantics (Sun et al. 2015; Piergiovanni and Ryoo 2018) which is more challenging for frame-level action classification.

Spatial characteristics. In most TAS datasets, scene, tool and object (even more important than the action itself sometimes) play very important roles in human action recognition. However, we should pay more attention to the action in many practical applications (Bhattacharya et al. 2020; Li et al. 2019). Besides, the task-driven datasets cannot show the full human body in an expected manner. Therefore, it is difficult to extract more modal features to perform TAS tasks.

Temporal characteristics. Generally, the action content categories for task-driven TAS datasets are simple, and the speed difference in distinct actions is too small. The little speed variance is difficult to cause frame-level feature changes, which is less challenging for TAS tasks.

The issues above limit the broader research of TAS models. In order to exploit new methods on the task of motion-centered TAS, this paper proposed a new dataset named MCFS. MCFS is composed of 271 single figure skating performance videos. The videos are taken from the 17.3 hours competition of the 2017-2019 World Figure Skating Championships. Each clip is 30 frames per second, with a resolution of 1080×720 and a length of 162s to 285s. All actions are annotated with the semantic labels on three levels (see Fig. 1). The camera focuses on the skater to ensure that he (she) appears in every frame during the action. Compared with the existing datasets, MCFS has five remarkable advantages which are listed as follows.

Multi-level fine-grained semantics. All the annotations are carried at three levels, namely set, subset and element in this paper. Fine-grained semantics means that similar actions may have different labels because of motion-centered characteristics in figure skating (e.g. Lutz and Flip jumps are similar in motion aspect, but are two different jumps.). Such a semantic hierarchy provides a distinct structure for comprehending coarse-grained and fine-grained operations.

Multi-modal action features. Previous datasets only offer features based on RGB video content such as Flow and I3D (Carreira and Zisserman 2017).etc., while MCFS provides extra Skeleton (Cao et al. 2017) feature which provides

[*] Equal contribution.

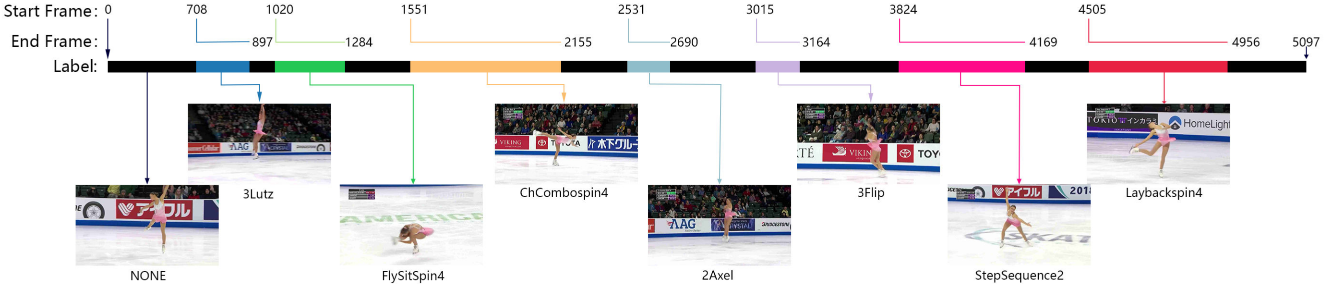


Figure 1: A video in MCFS. The labels of this video belong to the subset-level.

new opportunities and has significant to TAS methodological research.

Motion-centered human actions. All actions are independent of scenes and objects in MCFS dataset (i.e. most classes of actions are dominantly biased for skaters' pose.)

Large variance of action speed & duration. In the MCFS, the action content is complicated, and the speed difference in distinct actions is too large. For instance, one jumping action is completed within about 2s, by contrast, the longest step could reach 72s. The large speed variance always makes the large action duration variance of different actions, which can be regarded as a great challenge to frame-based TAS task.

Specialization. All videos in the MCFS are high-resolution records taken from the World Figure Skating Championships. Moreover, professional quality control is carried out on the full sequence of video annotations to guarantee the correctness, reliability and consistency of annotations.

According to the characteristics of MCFS, a series of empirical studies are conducted to discover the challenges of motion-centered TAS. Specifically, we first tested various TAS techniques and observed the performance of these methods is far from satisfactory in high-speed motion TAS. In order to provide assistance for future research, we also reviewed some modeling options, such as input data patterns. We found that for fine-grained TAS task, 1) motion information [1] plays a very important role, rather than depending on the scene and object of the video content. 2) The fine-grained categories are more likely to be used to increase the frame-based misjudgments ("burr" phenomenon) in clip action decision, which might become a new challenge for the existing TAS models. 3) The input modal of TAS model is very important, and new modal of the input (e.g. Skeleton) will exploit the research of TAS a new branch (e.g. GNN-based (Scarselli et al. 2008) TAS approaches).

Taken together, the work has contributed to the study of TAS task can be listed as the following two aspects:

(1) The MCFS dataset we collected is the first challenging dataset for TAS task with large action speed, duration variance, and complex motion-centered actions. It can be used to provide high quality and fine-grained annotations of full sequence, special annotations can be divided into three semantic levels, namely set, subset and element.

(2) We make in-depth study on MCFS, explored optional multi-modal features as input data of TAS model, and reveal the major challenges of future research and potential applications for high-vary speed motion tasks.

Related Work

Methods for TAS

Unsupervised Learning Approaches. For unsupervised TAS task, the major technique is to exploit discriminative information by clustering of spatio-temporal features. Such models introduce temporal consistency into the TAS methods by using LSTMs (Bhatnagar et al. 2017) or generalized mallows model (Sener and Yao 2018). Kukleva et al. (Kukleva et al. 2019) utilized both frame-wise clustering and video-wise clustering to model bag-of-words representation of each video. Besides, in order to use the contextual event in videos fully, Garcia et al. (Garcia del Molino, Lim, and Tan 2018) proposed an LSTM-based generative network when solving TAS task. For dynamics TAS task, Aakur et al. (Aakur and Sarkar 2019) proposed a self-supervised and predictive learning framework by utilizing features of adjacent frames as loss function. As another efficient dynamics TAS approach without training or clustering, MWS only involves curvature of action features in a neighborhood to realize segmentation's locations of a clip.

Weakly Supervised Approaches. The key idea of the weakly supervised TAS task is to mitigate the dependence of direct labeling by using indirect supervision manner to achieve highly TAS performance. For order-level weakly supervised TAS task, Ding et al. (Ding and Xu 2018) proposed a temporal autoencoder to predict frame-by-frame labels, and combined soft boundary assignment to iteratively optimize the segmentation results. To further explore the temporal structure, Kuehne et al. (Kuehne, Arslan, and Serre 2014) used "task graph" for order description and developed a hierarchical model based on HMMs for task-driven TAS. For online TAS, Richard et al. (Richard et al. 2018) used Viterbi-based loss offer a new deep model to achieve the frame-wise TAS goal. Recently, an order-free TAS method (Richard,

[1] Mainly refers to the spatial position and the time change of action sequence. In addition, it includes certain statistical characteristics of actions, such as: the variance of the action duration, and the variance of the action speed.

Dataset	Duration	People	Segments	Task	Classes	RGB?	Skeleton?	Fine-grained?	Year
GTEA	0.57h	4	922	CA	11	√	×	×	2011
MPII	9.8h	12	5609	CA	-	√	√	√(Semantics)	2012
50Salads	5.3h	27	966	CA	19	√	×	×	2013
JIGSAWS	2.6h	-	1703	SA	3	√	×	×	2014
Breakfast	77h	52	8456	CA	48	√	×	√(Temporally)	2014
Ikea-FA	3.9h	32	-	MS	5	√	×	×	2017
EPIC-KITCHENS	55h	32	39596	CA	5	√	×	×	2018
MCFS(ours)	17.3h	186	11656	FS	130	√	√	√(Semantics)	2021

Table 1: Comparisons of attributes existing datasets. *CA: cooking activities, SA: surgical activities, AF: assembling furniture, FS: figure skating

Kuehne, and Gall 2018) based on probabilistic model for set-level weakly supervised TAS is proposed.

Fully Supervised Approaches. Fully supervised TAS aims to segment the video into semantically consistent "blocks". A large amount of related works have explored for supervised TAS tasks. Most supervised TAS models adopt autoencoder architecture for preserving temporal consistence between input and output. For example, Lea et al. (Lea et al. 2017) proposed a temporal convolutional network for TAS, which utilized dilated convolutions to improve the process of pooling and upsampling. In (Li et al. 2020), Farha et al. proposed a multi-stage structure combining smoothing loss for TAS tasks, which also involved autoencoder network. Lei (Lei and Todorovic 2018) developed temporal deformable residual network using deformable temporal convolutions to enhance the TAS performance. Yet these methods suffer from long training time and unsatisfactory segmentation accuracy, which might be explained by the model architecture.

TAS-related Datasets

TAS-related datasets include TAS and action localization. Action localization aims to localize the temporal intervals of query actions, for example FineGym (Shao et al. 2020), while TAS intend to divide a video into independent actions at frame-level. We focus on TAS in this paper. For the early datasets, GTEA (Fathi, Ren, and Rehg 2011) and 50Salads (Stein and Mckenna 2013), which are based on coarse-grained cooking tasks only, have less than 20 categories of actions (11 and 19 categories, separately), while the surgical activity dataset of JIGSAWS (Gao et al. 2014) only consists of 3 categories. The existing methods can achieve well performance based on these datasets limited by the number of categories and video duration. Later, many datasets have been improved in terms of video duration, action categories, body motion, and fine-grained semantics (including temporal fine-grained units and semantic fine-grained class). All the above improvements make the TAS-related datasets more challenging and practical. For TAS with body motion task, Ikea-FA (Toyer et al. 2017) and MPII (Schiele et al. 2012) datasets realize the upper (occlusive) body motion characteristics, due to the particularity of furniture assembly and cooking tasks. Recently, most datasets are focusing on finer determination of action boundaries, especially for temporal fine-grained action units. Breakfast (Kuehne, Arslan, and Serre 2014) constructs an order graph and units description; EPIC-KITCHEN (Damen et al. 2018), which introduces visual object detection to form a temporal fine-grained action unit, is a large-scale cooking TAS dataset. Actually, the above two tasks can be regarded as a procedure segmentation task in cooking, which is proposed in the Youcook2 (Zhou, Xu, and Corso 2018) dataset. Youcook2 provides not only a temporal location, but also descriptions of the actions in a sentence. In addition, tool-object fine-grained semantic class is offered in the MPII dataset. However, most of the existing datasets are based on tool-objection content in TAS-related tasks, and it is impractical to extract Skeleton features without full-body motion. Besides, the lack of characters, fine-grained semantics and categories in the existing datasets also limits the development of the TAS task. Table 1 shows the development of TAS-related datasets in the past decade. These datasets, where the action segmentations are more based on hands, tool and objects, are mainly task-driven. These reasons hindered the development of the TAS methods based on human motion. MCFS will make up for the shortcomings of the existing TAS datasets, and promote the discovery of new problems in the TAS tasks. We believe MCFS will be a new challenging dataset for motion-centered TAS.

The MCFS Dataset

MCFS aims to be a motion-centered dataset for TAS task, which can better exploit new TAS models. In this section, we introduce the challenges of category definition, data annotation and quality control in MCFS, seperately. Moreover, the detailed construction process including data preparation and annotation details are introduced. Finally, we stated that MCFS exhibits more characteristics of statistics and physical motion, which, competitive with the existing datasets.

Key Challenges

There are a series of challenges during the data collection procedure, due to the top-level professionalism and complexity of figure skating. Firstly, as a highly professional sport, it is impractical to define categories manually for figure skating. Fortunately, the exactitude of labeling data can be ensured under the guide of the official technical documents of figure skating. Secondly, for the annotators, it is difficult to determine the category and boundary of actions, since the action is fast and highly similar to other actions within the same subset. To address the issue, the annotators are trained with necessary specialized knowledge by professionals in figure skating. Besides, this work is guided by the labels in the original video sequences (The upper left label in Fig. 2)). Based on the above criterions, thus, the labeling and division for our MCFS would be more convincing. However, we still enforce a series of measures to ensure high-quality dataset label, as described in next part.

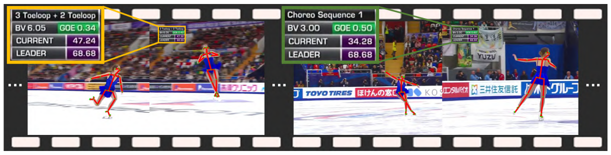


Figure 2: Labels variation in the original video sequences. The upper left label in the original video sequences will change when the action changes.

Dataset Construction

Data Preparation. In MCFS, we collect 38 official videos of 186 skaters from 2017 to 2019 World Figure Skating Championships. The complexity of the data distribution in the same set can be ensured with sufficient skaters. Each video sequence is of high resolution and FPS to preserve the integrity of actions. Besides, the duplicate video are removed through manual checking. In addition, We provide I3D, Skeleton for subsequent experiments.

Annotation Collection. In MCFS, we apply three level semantics annotations and collect 4 sets, 22 subsets and 130 elements at each annotation level, respectively. The MCFS structure is shown in Fig. 3. For example, the set "Spin" can be expressed as Spin = {ChCombospin, CamelSpin, Laybackspin, Sitspin, ChSitSpin, ChCamelspin} with 6 subsets. The elements in each subset will be further annotated with defined element labels. Such a semantic hierarchy provides a distinct structure for comprehending coarse-grained and fine-grained operations. The following requirements should be observed throughout the annotation process. First, it is necessary to refer to the real-time labels, which are provided in the original videos, to determine the start frame and the end frame. Meanwhile, all incomplete video clips will be removed in this process. Second, according to the element-level action category, the official manual and categorization structure are referred to classify them into subsets and sets.

Annotation Tool. As the segments are variant in length and content, the workload to annotate the MCFS with a conventional annotation tool will be crushing. In order to improve the annotation efficiency, we develop a new tool to preview the two frames before and after the current frame. In addition, with this tool, the start and end frames can be selected directly while updating the category manually.

Quality Control. We annotate all the frame-level fine-grained action categories and temporal segmentation boundaries in MCFS. To assure the quality of the MCFS dataset, a series of control mechanisms are adopted, including: 1) Train annotators with professional knowledge. 2) Provide reference documents and sample videos. 3) Test the annotator's labeling level strictly before formal annotation. 4) Review the annotated videos.

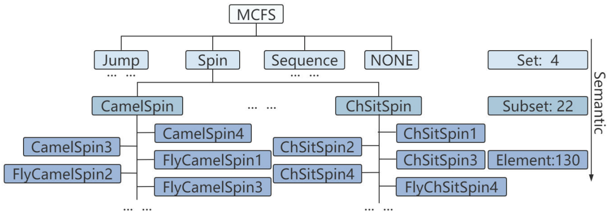
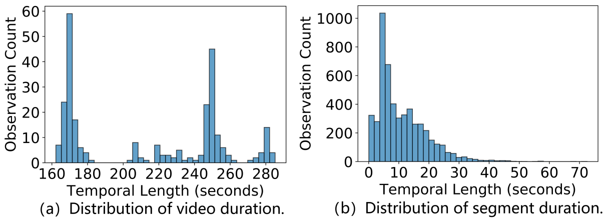


Figure 3: A three level semantics annotations and collect 4 sets (e.g. Spin), 22 subsets (e.g. CamelSpin) and 130 elements (e.g. CamelSpin3) at each annotation level.

Statistic

The MCFS dataset consists of 271 samples captured from 38 competition videos which include more than 1.73 million frames. All the annotations are carried at three levels, and the number of categories at each level is 4, 22 and 130 respectively. There are 93 out of 130 elements has at least two samples that present the natural heavy-tail distribution. Except the "NONE" category, we annotate 2,995 effective clips in all samples. The distribution of video duration is shown in Fig. 4 (a). The total video length is 15.9 hours with an average duration of 212s per video. All the videos remain untrimmed and can be up to 300s. The distribution of segment durations is shown in Fig. 4 (b) with mean and standard deviation of 9.4s and 8.3s, respectively. The longest segment lasts 72s and the shortest one lasts 1s. The large range of sample duration is a challenge to the TAS task.



(a) Distribution of video duration.

(b) Distribution of segment duration.

Figure 4: MCFS-22 dataset duration statistics.

Dataset Properties

Motion-centered Human Actions. For most TAS datasets, many factors such as hand, tool and object can affect the results. For example, the action and scene are same in 50Salads dataset for "cut-tomato" and "cut-cucumber". Discriminant information only depends on the object in the hand. However, all samples have a relatively consistent scene in the MCFS dataset. The discriminant of the action category and boundary information only depends on the human body action should be realized to meet the challenge of modeling new TAS methods.

Multi-modal Action Features. We extract not only Flow and I3D features but also two Skeleton features (2D locations of 18 and 25 major body joints) based on RGB video in the MCFS dataset. Fig. 2 shows the Skeleton features. Skeleton features offers a new opportunity for the combination of GCN (Huang, Sugano, and Sato 2020) and TAS methods. It may also develop a new direction for the multi-modal of temporal human action segmentation. We hope the MCFS dataset can promote the research of machine learning for temporal action segmentation of human action.

Large Variance of Action Speed and Duration. For sport datasets like MCFS, the different actions are with large action speed and duration variance. For example, jump is generally completed in 2-3s, but the longest sequence can be more than 70s. We have calculated the nearest neighbor variance of I3D features for 21 frames in four datasets as shown in Fig. 5. It can be clearly seen that compared to the other three kitchen datasets, there are dramatic changes between different actions in the MCFS dataset. This brings great challenges to the division of boundaries.

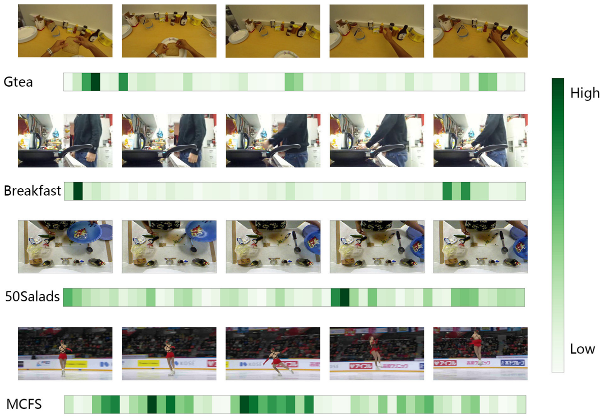


Figure 5: The nearest neighbor variance results of GTEA, 50Salads, Breakfast and MCFS-22 dataset.

High Similarity of Category. In the MCFS dataset, two samples of different categories may only have few different frames called key frames. For example, in a single jump, “Lutz” and “Flip” performs are basically the same, except for the differences inside and outside the ice skate blade. For continuous jump like “3Lutz_3Loop” and “3Lutz_3Toeloop”, the first jump is exactly the same, while the difference only depends on the subsequent jump. Such subtle differences can easily make the model misjudge the category and segmentation point of actions. Meanwhile, because similar frames may appear in different actions, multi-semantics frames become another inevitable problem.

Experiments

In this section, experimental setup is first introduced for TAS task. Then, we report the experimental results on benchmark datasets (such as 50Salads, GTEA and Breakfast) and list the baseline results of state-of-the-art TAS methods by leveraging MCFS dataset. In addition, the characteristics of MCFS is discussed based on the experimental results.

Experimental Setup

Data. MCFS is randomly split into 189 and 82 videos for training and testing, respectively. Then, we utilize 5-fold cross validation to assess generalization of the models. MCFS-4, 22 and 130 share the same splits, but are annotated by the different three hierarchical semantic labels (set, subset, element), respectively, which have been introduced in section “the MCFS Dataset”.

I3D Feature Based on RGB. For each frame, a 2048 dimensional feature vector of I3D, whose final feature vector for each frame is obtained by concatenating the vectors form both RGB and flow streams which results in a 2048 dimensional vector for each frame, is pretrained on Kinetics (Kay et al. 2017). Specifically, temporal window for I3D of a frame consists of 20 temporal nearest neighbored frames of current frame (altogether 21 frames). More details can be referred to reference (Carreira and Zisserman 2017).

Skeleton Feature. On the MCFS, we use the 2D pose estimation results from the OpenPose (Cao et al. 2017) toolbox which outputs 18 joints and 25 joints. In addition, these joints of Skeleton feature are normalized by dividing two spatial direction coordinates of joints by corresponding frame size respectively, and then centralized by the waist joint (center joint). All our experiments utilize the Skeleton feature of 25 joints.

Dataset	F1@{10,25,50}			Edit	Acc
50Salads					
Bi-LSTM	62.6	58.3	47.0	55.6	55.7
ED-TCN	68.0	63.9	52.6	59.8	55.7
MS-TCN	76.3	74.0	64.5	67.9	80.7
SSTDA	83.0	81.5	73.8	75.8	83.2
GTEA					
Bi-LSTM	66.5	59.0	43.6	-	55.5
ED-TCN	72.2	69.3	56.0	-	64.0
MS-TCN	85.8	83.4	69.8	79.0	76.3
SSTDA	90.0	89.1	78.0	86.2	79.8
Breakfast					
Bi-LSTM	33.4	21.9	13.6	35.8	56.6
ED-TCN	48.6	43.1	27.7	38.6	67.3
MS-TCN	52.6	48.1	37.9	61.7	66.3
SSTDA	75.0	69.1	55.2	73.7	70.2

Table 2: Comparison with the state-of-the-art on 50Salads, GTEA, and the Breakfast dataset (All data obtained from (Farha and Gall 2019) and (Chen et al. 2020a)).

Evaluation Metric. For evaluation, we report the frame-wise accuracy (Acc), segmental edit distance and the segmental F1 score at overlapping thresholds 10%, 25% and 50%, denoted by F1@{10, 25, 50} (Farha and Gall 2019). The F1 score can penalize over-segmentation errors while

Dataset	Modality	F1@{10,25,50}			Edit	Acc
MCFS-4						
Bi-LSTM	I3D	33.4	21.9	13.6	35.8	56.6
ED-TCN	I3D	48.6	43.1	27.7	38.6	67.3
MS-TCN	I3D	74.1	67.4	50.2	79.6	71.9
MS-TCN	Skeleton	86.8	82.6	72.1	86.9	82.0
SSTDA	I3D	75.8	69.9	52.5	82.1	71.4
SSTDA	Skeleton	88.7	84.9	74.6	89.3	82.0
MCFS-22						
Bi-LSTM	I3D	14.8	5.9	1.5	13.6	54.3
ED-TCN	I3D	32.3	25.7	11.6	25.6	58.8
MS-TCN	I3D	49.4	44.1	29.8	52.6	62.6
MS-TCN	Skeleton	74.3	69.7	59.5	74.2	75.6
SSTDA	I3D	52.7	46.3	31.1	56.3	59.1
SSTDA	Skeleton	76.7	72.2	61.2	77.5	75.7
MCFS-130						
Bi-LSTM	I3D	9.9	2.5	0.3	7.6	54.3
ED-TCN	I3D	30.2	22.7	10.6	23.1	54.5
MS-TCN	I3D	36.6	30.5	20.0	36.3	58.0
MS-TCN	Skeleton	56.4	52.2	42.5	54.5	65.7
SSTDA	I3D	42.6	37.3	24.6	44.4	55.1
SSTDA	Skeleton	63.8	60.1	49.8	63.5	65.4

Table 3: Element-level action recognition results of representative methods. Specifically, results of recognizing element categories across all sets, within a subset, and within an element.

it does not penalize minor temporal shifts between the predictions and ground truth. This is appropriate for TAS task because it is important to avoid over-segmentation errors for video summarization. As for this reason, we use the F1 score as a measure of the quality of the prediction. The detailed description of the above evaluation metrics can be referred to the related reference (Lea et al. 2017).

Baselines for Temporal Action Segmentation

In this subsection, we conduct the experiments utilizing I3D feature on 50Salads, GTEA and Breakfast datasets, and list the detailed experimental results of four TAS methods based on both TCN (including ED-TCN (Lea et al. 2017), MS-TCN (Farha and Gall 2019) and SSTDA (Chen et al. 2020a)) and the LSTM (i.e. Bi-LSTM (Graves, Fernández, and Schmidhuber 2005)) in Table 2. We show results for two modalities (I3D and Skeleton) of MCFS, as well as for the four TAS methods in Table 3 (We only select two state-of-the-art models for Skeleton.). The detailed experimental results illustrate three challenging properties of MCFS as follows.

Motion-centered. Table 2 illustrates that I3D can achieve superior performance on the benchmark datasets (50Salads, GTEA and Breakfast). Specially, most values of metrics (including accuracy, segmental edit distance and F1 score) of SSTDA model are over 70% (only F1@25,50 on Breakfast is below 70%). This is because the scene and objects, which can be well characterized by I3D in a video sequence, play important roles on object-based TAS datasets such as 50Salads. For example, recognizing "cut_tomato" and "cut_cheese" is free of "cut", but is to distinguish the different characteristics between tomato and cheese. In contrast, figure skating pays no attention to scene and object. Specifically, the accuracy in Table 3 are generally much lower than the accuracy in Table 2 when using the same experimental setup. In addition, some categories of actions may be confusing because of the extremely high similarity of motion in MCFS. As shown in Fig. 6, "Toeloop" is wrongly recognized as "salchow" and "Lutz_3Toeloop". The reason for the confusion of actions is that single-jump can only be recognized by a few of key frames, while joint-jump pay attention to more key frames of the two consecutive jumps. The above results illustrate MCFS is challenging on motion-centered TAS.

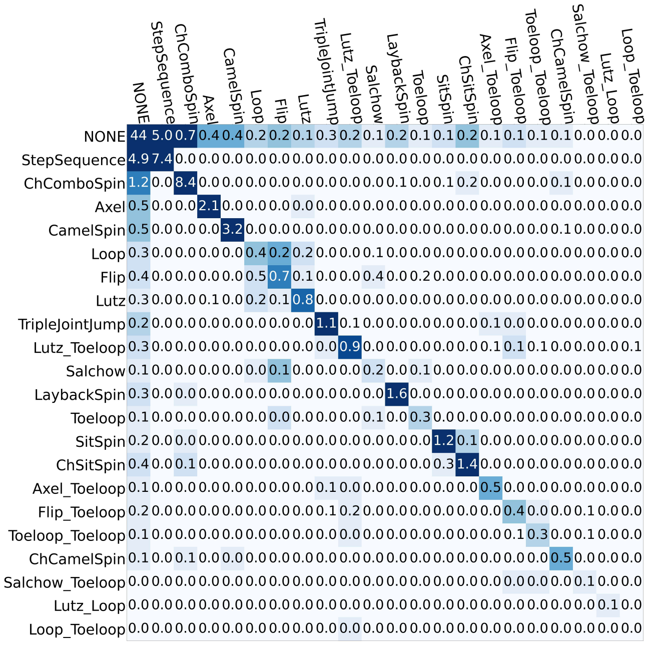


Figure 6: The confusion matrix results of MCFS-22 (Skeleton) utilizing MS-TCN.

Temporal Information. In TAS task, it is very important to capture the time dynamics. Both TCN-based and LSTM-based methods could work well by utilizing the existing datasets (50Salads and GTEA etc.) without the complex temporal characteristics. Due to the problem of large variance of action speed and duration in MCFS, the LSTM-based methods (Bi-LSTM) will suffer gradient disappearance by a long time series inputting, while the TCN-based methods can avoid this issue and can obtain far superior performance (Table 3). Another possible reason of the above issue is the temporal weak correlation among actions in MCFS. Besides, the complex transition motion (non-regular patterns in transition motions including content, duration and location) interspersed among actions is also challenging to determine the temporal intervals of actions for TCN-based

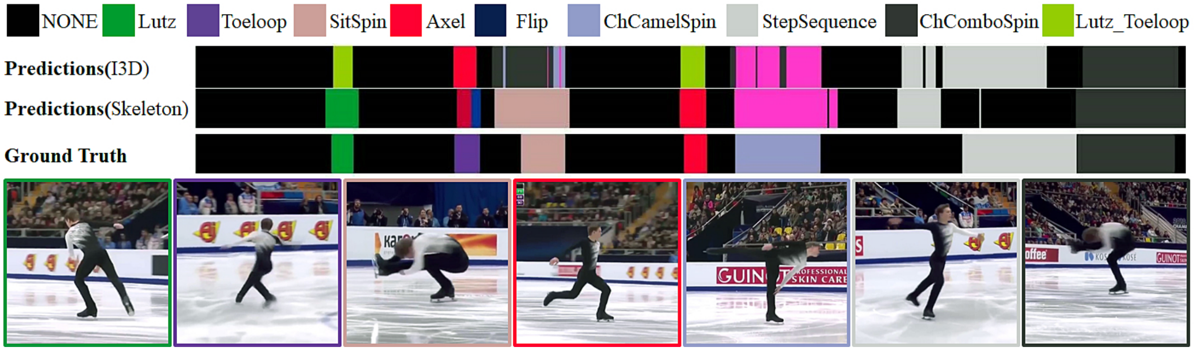


Figure 7: Qualitative results for the TAS task on MCFS.

networks.

Fine-grained Semantics Label. MCFS provides three levels of fine-grained annotations which result in confusion of different categories by similar actions. It is possible to cause many over-segmentation errors in label prediction because of fine-grained characteristics. MCFS-22 achieves excellent accuracy than 50Salads as shown in Table 2 and Table 3 by utilizing ED-TCN with the same setup. However, segmental edit distance and F1 score of MCFS-22 are much lower than that of 50Salads and GTEA. It is a serious problem that the finer semantics label will lead the more over-segmentation errors in MCFS. For example, the MCFS-130 performs worse than either MCFS-4 or MCFS-22 by any compared TAS methods in Table 3.

Skeleton Features of Action

It can be seen in Fig. 7 and Table 3, both the errors of action recognition and the errors of over-segmentation based on I3D predictions are far more than that based on Skeleton, which illustrates that MCFS depends on the human motion. In addition, OpenPose can be easily used for Skeleton extraction because of the whole body appearing of the skater in the video. The two TAS methods (MS-TCN and SSTDA) using Skeleton feature achieves better performance than that using I3D feature. For example, in MCFS-22, the performance of SSTDA using Skeleton are 24% and 30.1% higher than that using I3D on F1@0.1 and F1@0.5 respectively.

Directions for Future Works

In human action classification task, GNN based models have be developed rapidly, such as ST-GCN (Yan, Xiong, and Lin 2018), 2S-AGCN (Shi et al. 2019) and MS-G3D (Liu et al. 2020). So far as we are aware, due to the lack of Skeleton features in the existing datasets, GNN-based approach is not used in TAS task. MCFS could be utilized to exploit more excellent multi-modal and Skeleton-based models by using optic flow and Skeleton features in TAS field.

Potential Applications

The high-quality data of MCFS has offered a foundation for various applications. Besides fine-grained action segmentation tasks, it also includes some potential applications.

Video Description. While there has been increasing interest in the task (Xu et al. 2016; Wang et al. 2018) of describing video with natural language, current computer vision algorithms are still severely limited in associated language that they can recognize. We believe MCFS can be utilized for Video Description because it can build the embedding between video frames and the words.

Action Reasoning. Action reasoning (Pirsiavash, Vondrick, and Torralba 2014) is an interesting issue. For example, it is straightforward to conclude a 3Lutz-3Toeloop jump if single 3Lutz jump and 3toeloop jump have been recognized. This direction provides more empirical research ideas for model design.

Video-Text Retrieval. Cross-modal retrieval between videos and texts (Chen et al. 2020b) has attracted growing attentions. We believe that MCFS can contribute to Video-Text Retrieval, since all actions are annotated with the semantic labels on three levels in MCFS. Besides, such hierarchical structure enables methods has better generalization and improves the ability to distinguish fine-grained semantic differences.

Conclusion

In this paper, we introduce a new fine-grained dataset called MCFS for the TAS task. Hierarchical semantic structure of our dataset has been organized by professional knowledge. In addition, MCFS differs from existing TAS datasets in multiple aspects, including motion-centered human actions, large variance of action speed and duration, multi-modal action features and high category similarity. Based on the above differences, a number of comparative experiments are conducted on MCFS. The experimental results indicate it is promising and challenging for MCFS to be used in the TAS task. Besides, MCFS could be utilized to exploit more excellent multi-modal and Skeleton-based models by using optic flow and Skeleton features in TAS field. We will move on to propose more state-of-the-art TAS methods. We hope that our dataset would promote the development of action analysis and related research topics.

References

Aakur, S. N.; and Sarkar, S. 2019. A Perceptual Prediction Framework for Self Supervised Event Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bhatnagar, B. L.; Singh, S.; Arora, C.; Jawahar, C.; and CVIT, K. 2017. Unsupervised Learning of Deep Feature Representation for Clustering Egocentric Actions. In *IJCAI*, 1447–1453.

Bhattacharya, U.; Mittal, T.; Chandra, R.; Randhavane, T.; Bera, A.; and Manocha, D. 2020. STEP: Spatial Temporal Graph Convolutional Networks for Emotion Perception from Gaits. In *AAAI*, 1342–1350.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, M.-H.; Li, B.; Bao, Y.; AlRegib, G.; and Kira, Z. 2020a. Action Segmentation with Joint Self-Supervised Temporal Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9454–9463.

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020b. Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10638–10647.

Damen, D.; Doughty, H.; Maria Farinella, G.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 720–736.

Ding, L.; and Xu, C. 2018. Weakly-Supervised Action Segmentation with Iterative Soft Boundary Assignment. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Farha, Y. A.; and Gall, J. 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3575–3584.

Fathi, A.; Ren, X.; and Rehg, J. M. 2011. Learning to recognize objects in egocentric activities. In *IEEE Conference on Computer Vision & Pattern Recognition*.

Gao, Y.; Vedula, S. S.; Reiley, C. E.; Ahmidi, N.; Varadarajan, B.; Lin, H. C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D. D.; et al. 2014. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *Miccai workshop: M2cai*, volume 3, 3.

Garcia del Molino, A.; Lim, J.-H.; and Tan, A.-H. 2018. Predicting visual context for unsupervised event segmentation in continuous photo-streams. In *Proceedings of the 26th ACM international conference on Multimedia*, 10–17.

Graves, A.; Fernández, S.; and Schmidhuber, J. 2005. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In *Artificial Neural Networks: Formal Models & Their Applications-icann, International Conference, Warsaw, Poland, September*.

Huang, Y.; Sugano, Y.; and Sato, Y. 2020. Improving Action Segmentation via Graph-Based Temporal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14024–14034.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* .

Kuehne, H.; Arslan, A.; and Serre, T. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 780–787.

Kukleva, A.; Kuehne, H.; Sener, F.; and Gall, J. 2019. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12066–12074.

Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 156–165.

Lee, P.; Uh, Y.; and Byun, H. 2020. Background Suppression Network for Weakly-Supervised Temporal Action Localization. In *AAAI*, 11320–11327.

Lei, P.; and Todorovic, S. 2018. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6742–6751.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person reidentification. In *Proceedings of the IEEE International Conference on Computer Vision*, 3958–3967.

Li, S.-J.; AbuFarha, Y.; Liu, Y.; Cheng, M.-M.; and Gall, J. 2020. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Lin, J. F.-S.; and Kulić, D. 2013. Online segmentation of human motion for automated rehabilitation exercise analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22(1): 168–180.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 143–152.

Magliano, J. P.; and Zacks, J. M. 2011. The impact of continuity editing in narrative film on event segmentation. *Cognitive science* 35(8): 1489–1517.

Piergiovanni, A.; and Ryoo, M. S. 2018. Fine-grained activity recognition in baseball videos. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1740–1748.

Pirsiavash, H.; Vondrick, C.; and Torralba, A. 2014. Assessing the quality of actions. In *European Conference on Computer Vision*, 556–571. Springer.

Richard, A.; Kuehne, H.; and Gall, J. 2018. Action sets: Weakly supervised action segmentation without ordering constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5987–5996.

Richard, A.; Kuehne, H.; Iqbal, A.; and Gall, J. 2018. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7386–7395.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1): 61–80.

Schiele, B.; Andriluka, M.; Amin, S.; and Rohrbach, M. 2012. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision & Pattern Recognition*.

Sener, F.; and Yao, A. 2018. Unsupervised learning and segmentation of complex activities from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8368–8376.

Shao, D.; Zhao, Y.; Dai, B.; and Lin, D. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2616–2625.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Stein, S.; and Mckenna, S. J. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*.

Sun, C.; Shetty, S.; Sukthankar, R.; and Nevatia, R. 2015. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the 23rd ACM international conference on Multimedia*, 371–380.

Toyer, S.; Cherian, A.; Han, T.; and Gould, S. 2017. Human pose forecasting via deep markov models. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–8. IEEE.

Urban, T. L.; and Russell, R. A. 2003. Scheduling sports competitions on multiple venues. *European Journal of operational research* 148(2): 302–311.

Wang, B.; Ma, L.; Zhang, W.; and Liu, W. 2018. Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7622–7631.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1247–1257.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.