The THUMOS challenge on action recognition for videos "in the wild"

Haroon R Idrees, Amir R Zamir, Yu-Gang Jiang, Alex R Gorban, Ivan R Laptev, Rahul R Sukthankar, Mubarak R Shah

The THUMOS Challenge on Action Recognition for Videos "in the Wild"[☆]

Haroon Idrees[a,*], Amir R. Zamir[b], Yu-Gang Jiang[c], Alex Gorban[e], Ivan Laptev[d], Rahul Sukthankar[e], Mubarak Shah[a]

[a] *Center for Research in Computer Vision, University of Central Florida, Orlando, USA*
[b] *Dept. of Computer Science, Stanford University, USA*
[c] *School of Computer Science, Fudan University, Shanghai, China*
[d] *INRIA Paris - Rocquencourt, France*
[e] *Google Research, USA*

Abstract

Automatically recognizing and localizing wide ranges of human actions are crucial for video understanding. Towards this goal, the THUMOS challenge was introduced in 2013 to serve as a benchmark for action recognition. Until then, video action recognition, including THUMOS challenge, had focused primarily on the classification of pre-segmented (i.e., trimmed) videos, which is an artificial task. In THUMOS 2014, we elevated action recognition to a more practical level by introducing temporally untrimmed videos. These also include 'background videos' which share similar scenes and backgrounds as action videos, but are devoid of the specific actions. The three editions of the challenge organized in 2013–2015 have made THUMOS a common benchmark for action classification and detection and the annual challenge is widely attended by teams from around the world.

In this paper we describe the THUMOS benchmark in detail and give an overview of data collection and annotation procedures. We present the evaluation protocols used to quantify results in the two THUMOS tasks of action classification and temporal action detection. We also present results of submissions to the THUMOS 2015 challenge and review the participating approaches. Additionally, we include a comprehensive empirical study evaluating the differences in action recognition between trimmed and

[☆] www.thumos.info

[*] Corresponding author
Email address: haroon@cs.ucf.edu (Haroon Idrees)

untrimmed videos, and how well methods trained on trimmed videos generalize to untrimmed videos. We conclude by proposing several directions and improvements for future THUMOS challenges.

Keywords: Action Recognition, Action Detection, Action Localization, Untrimmed Videos, THUMOS, Dataset, Benchmark, UCF101

---

1. Introduction

The action recognition community has made great progress in the last few years, driven in large part by the release of large video datasets such as UCF101 [1] and HMDB [2] in conjunction with the development of new features [3], representations [4] and learning methods [5]. Recent datasets contain challenging videos with actions from various sources such as movies [2, 6], YouTube [7], and wearable cameras [8, 9]. The performance of methods evaluated on such datasets has steadily increased over the years [3]. In line with these advances in action recognition, the THUMOS challenge was introduced to the computer vision community in 2013 with the aim to explore and evaluate new approaches for large-scale action analysis from Internet videos in a realistic setting.

The THUMOS 2013 challenge was based on the UCF101 dataset [1], which similar to most of the commonly evaluated action recognition datasets consists exclusively of manually trimmed video clips that exclude temporal clutter. The assumption of such clean and trimmed videos may be reasonable during training time since it provides methods with strongly supervised data. However, the same restriction during testing is potentially impractical and unreasonable for several reasons:

- it assumes an (unrealistic) external process to temporally segment videos into clips that precisely surround the desired action;
- it creates a test set distribution that does not match the real-world distribution since the test data is free from temporal clutter, 'background' class data notwithstanding;

- it can allow methods to inadvertently exploit side-information, such as the length of the test video clip [10], even though this information is available only due to an artifact of the evaluation methodology.

Thus, the temporally segmented clips do not reflect the real world as the actions are typically embedded in complex dynamic scenes with rich causal and spatial relations among people and objects. While elimination of temporal clutter simplifies the recognition problem, it becomes difficult to predict the performance of different methods in real applications. In literature, there have been some efforts to address the problem of action recognition in untrimmed videos. For example, temporal detection has been studied in [11, 12, 13, 14, 15, 16], while spatiotemporal localization of actions has been addressed in [17, 18, 19, 20, 21, 22]. Such works deal with substantial amount of temporal clutter from movies and sports videos. However, they typically were evaluated on only a small number of action classes and required strongly supervised training and test sets. The THUMOS'14 challenge [23] introduced thousands of untrimmed videos in validation, background and test sets for 101 action classes providing the community with the first-of-its-kind dataset for action recognition and temporal detection in realistic settings with a standardized evaluation protocol. Similarly, THUMOS'15 challenge [24] extended the THUMOS'14 dataset by including a new test set constituting 5,613 positive and background untrimmed videos.

THUMOS (Greek: $\theta\upsilon\mu\acute{o}\varsigma$) which means a *spirited contest*, consists of two principal challenges: *classification* - where the goal is to determine whether a video contains a particular action or not, and *temporal detection* - where the goal is to classify an action and find its temporal locations in each video. The THUMOS action classes are from UCF101 [1] and can be divided into five main categories: *Human-Object Interaction*, *Body-Motion Only*, *Human-Human Interaction*, *Playing Musical Instruments*, and *Sports*. All the videos are publicly available from YouTube[1], and manually annotated both for action label and temporal span.

The objectives of the THUMOS challenge are twofold: a) to serve as a benchmark

[1] http://www.youtube.com/

and enable a comparison of different approaches on the tasks of action classification and temporal detection in large-scale realistic video settings; and b) to advance the state of the art. For instance, the accuracy on UCF101 increased from 45% in 2012 to almost 90% at THUMOS'13 [25]. Similarly, the 2014 and 2015 challenges are characterized by three significant differences compared to traditional action recognition. The **first** is the introduction of background videos that share similar scenes and objects as positive videos but do not contain the target actions. This downplays the role of appearance and static information since background videos are distinguishable from action videos primarily based on the motion. Associated with this is the **second** difference where the classification task is changed from a forced-choice multi-class formulation to a multi-label *binary* task, where each video can contain multiple actions. This has been enabled through the use of background videos and is not possible with other action datasets. And **third** is the introduction of untrimmed videos (Figure 1) for validation and testing as opposed to manually pre-segmented (or "trimmed") videos [26, 27, 28, 2, 1, 7] typically used in action recognition. Consequently, a testing video in THUMOS'15 can contain zero, one or multiple instances of an action (or different actions) that can occur anywhere in the given video.

One of the contributions of this paper is to extend and complement prior work with a study of action recognition in temporally untrimmed videos and show how it differs from trimmed videos using the THUMOS dataset (see Fig. 1). We address both video-level action classification and temporal detection problems and systematically evaluate and quantify the effect of temporal clutter. In particular, we evaluate the popular Improved Dense Trajectory Features (IDTF) [3] + Fisher Vectors + SVM pipeline that has dominated several action recognition benchmarks. While temporal clutter causes a drop in recognition performance, untrimmed videos also contain additional information about the context of actions. In the evaluation study, we explore action context and show improvements in action recognition performance using context information extracted from temporal neighborhoods of untrimmed videos.

The rest of the paper is organized as follows. We provide comparison with existing datasets in Sec. 2 and define challenge tasks in Sec. 3. Next, we explain the procedure used for collection and annotation of the dataset in Sec. 4, and present the evaluation

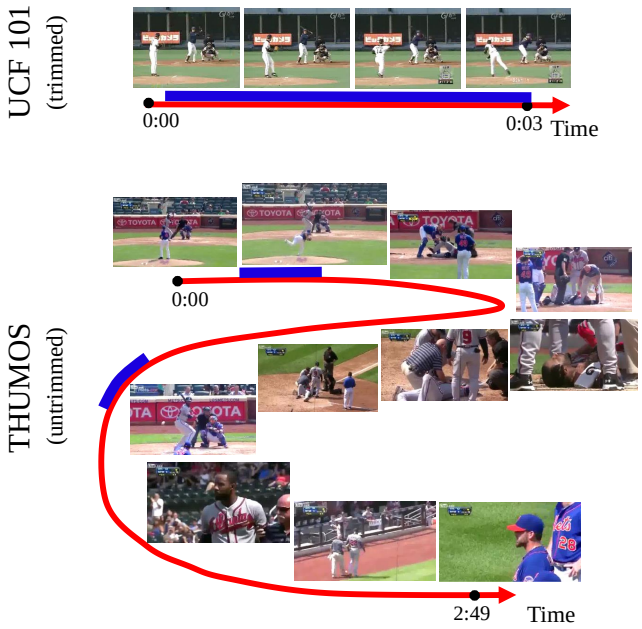


Figure 1: Illustration of contrast between a (trimmed) video clip for the 'BaseballPitch' action from the UCF101 dataset and an *untrimmed video* from the corresponding action taken from the validation set of THUMOS'15. Note that the entire temporal span of the video (shown in red) contains a variety of baseball actions with the pitch occurring multiple times (shown in blue).

protocol in Sec. 5. Since the challenge in still nascent, a longitudinal study of participants' methods would be possible after the next few years. Nonetheless, we perform a cross-sectional study of the THUMOS'15 challenge with a summary of methods presented in Sec. 6 and results reported in Sec. 7. Additionally, we study the impact of background and temporal clutter, as well as role of context for action recognition in untrimmed videos in Sec. 8. Finally, we conclude with ideas on improvements for future challenges in Sec. 9.

2. Related Datasets

Early datasets on action recognition in videos, such as **KTH** [26] and **Weizmann** [27], employed actors performing a small set of scripted actions under controlled conditions. The next series of datasets, such as **CMU** [29] and **MSR Actions** [30], introduced

scripted actions performed against challenging dynamic backgrounds. Later datasets, such as **HOHA** [31] and **Hollywood-2** [6] moved to relatively more realistic video footage from Hollywood movies and broadcast television channels, respectively. Many of these datasets provided spatiotemporal annotations for action instances in relatively short untrimmed videos. However, this level of annotation became impractical once the research community demanded larger datasets. Most of the modern datasets are collected from realistic sources, have more classes and have more temporal clutter. For instance, the **Human Motion DataBase (HMDB)** [2] dataset released in 2011 contains 51 action categories, each containing at least 101 samples for a total $\sim$6800 action instances.

The UCF series of datasets started with **UCF Sports** [28] in 2008, which comprised of movie clips captured by professional filming crew, and offered videos with camera motion and dynamic backgrounds. The next in the series were **UCF11** [7] and **UCF50** [32], released in 2009 and 2011, respectively. Both datasets consisted of trimmed clips from a variety of sources ranging from digitized movies to YouTube. The **UCF101** dataset [1] is a superset of the previous UCF11 [7] and UCF50 [32] datasets and was released in 2012. It contains 13320 video clips of 101 action classes (Appendix A). The actions are divided into 5 categories: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, Sports, as shown in Figure 2. The clips of one action class are divided into 25 groups which contain 4–7 clips each. The clips in one group share some common features, such as the background or actors. The videos have a resolution of 320$\times$240, with a total duration of $\sim$27hrs. The training data of the THUMOS challenge uses the trimmed clips of UCF101, however, the datasets for THUMOS'14 and THUMOS'15 additionally include untrimmed positive and background videos for validation and test sets.

The **Sports-1M** [33] dataset, released in 2014, contains more than 1 million untrimmed videos from almost 487 classes with about 1000–3000 videos per action class. The dataset is divided into the following categories: Aquatic Sports, Team Sports, Winter Sports, Ball Sports, Combat Sports, Sports with Animals, and taxonomy becomes fine-grained at the lower levels. While the dataset is large in the number of videos, it focuses only on sports actions and is weakly annotated (only at the video level) with au-

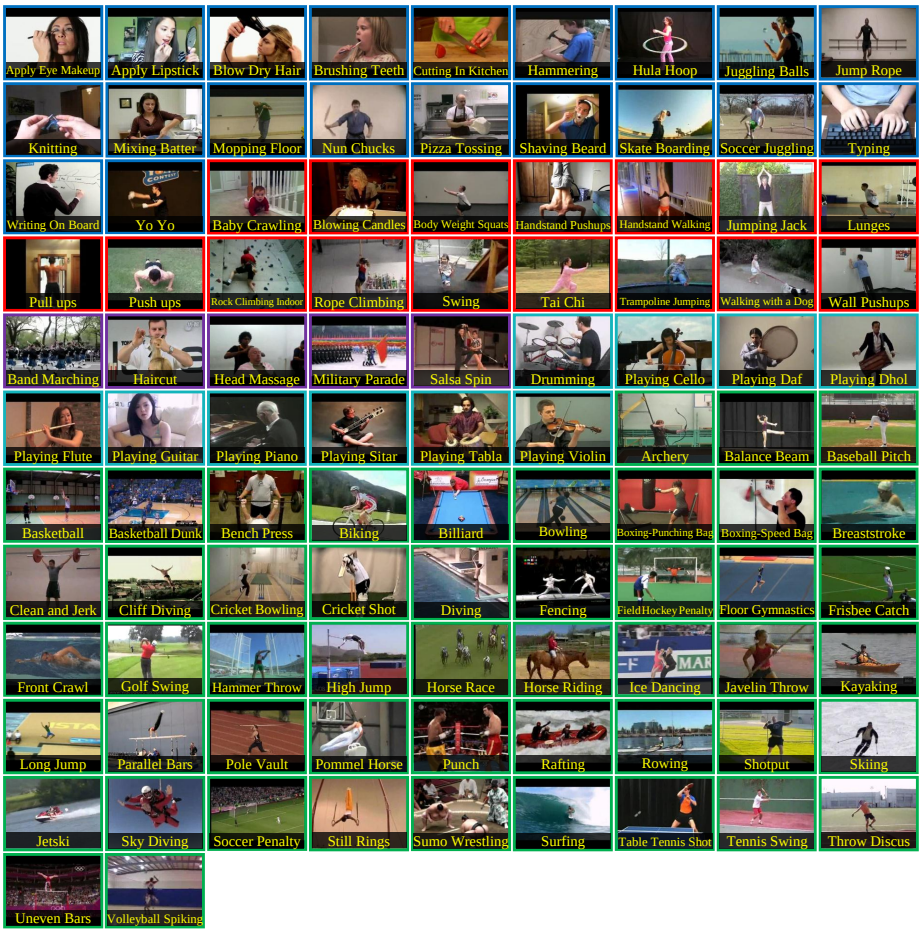


Figure 2: The figure shows the sample frames of the actions from UCF101 dataset [1]. The color of frame borders specifies the action type to which they belong: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, Sports (c.f. Appendix A).

tomatically generated – and thus potentially noisy – labels. By contrast, the THUMOS dataset includes videos that have been carefully annotated. Furthermore, THUMOS includes negative background videos for each action class in both the validation and test sets, making the action recognition task more difficult.

"TREC[2] *Video Retrieval Evaluation" (TRECVID)*[3] is a series of competitions and workshops conducted by National Institute of Standards and Technology (NIST) with the aim to stimulate research in automatic segmentation, indexing, and content-based retrieval of digital videos. Since the first competition in 2003, it now consists of several independent tasks. The datasets for each task have been typically extended each year, and are only available to the participants who register for the competition. There are two set of tasks in TRECVID that are related to THUMOS challenge. One of the task is *Semantic Indexing* (SIN) and the associated *Localization* (LOC) which focus on the detection and localization in video shots or clips. The dataset consists of **Internet Archive Creative Commons (IACC)** [34] collected by NIST with 15300 videos for a total of $\sim$1200 hours. Only short clips or shots are annotated for 500 object, scene and action concepts for training. During testing, the highest scoring shots from all participants are gathered, and used for generating ground truth. Since only a subset of test data is annotated, inferred Average Precision is used for evaluation (infAP) [35] of each concept. For 2015, only 30 concepts were evaluated for detection and 10 for spatio-temporal localization. It is important to remember that unlike untrimmed videos in THUMOS, the spatio-temporal localization in SIN task is performed on pre-defined trimmed shots.

Another TRECVID task, *Multimedia Event Detection*, requires the methods to provide a confidence score for each video from a collection as to whether the video contains the event. The collection is complemented with event kits that include a textual description of the event and information about related concepts that are likely to occur in each event. An associated task, *Multimedia Event Recounting*, has the objective of stating key evidence, in the form of text with pointers to detected concepts, that led

[2]TREC stands for "Text REtrieval Conference"

[3]http://www-nlpir.nist.gov/projects/trecvid/

a Multimedia Event Detection (MED) method to decide that a multimedia clip contains an instance of a specific event. There were 20 pre-specified events for the main task, and Mean Average Precision and inferred MAP were used as metrics for event detection. The evaluation for recounting is performed after results are returned by participants where judges evaluate the key evidences for correctness. The dataset consists of **Heterogeneous Audio Visual Internet (HAVIC)** Corpus [36] collected by the Linguistic Data Consortium. For 40 events, it has $\sim$290 hrs of training videos. The testing is performed on a separate set with 200,000 videos ($\sim$8000 hrs). The THUMOS challenge focuses on actions, which are less complex and more atomic than MED events, and are primarily affected by motion of actors. Furthermore, the action concepts in the Multimedia Event Recounting task are primarily driven by events rather than the actions themselves. Thus, miss-detections of actions are not penalized in evaluation as long as the evidence presented by a system is sufficient for detection of an event.

ActivityNet [37] is a recent dataset for recognition of human activities. It was released in 2015, two years after THUMOS, and consists of 203 activity classes with an average of 137 untrimmed videos per class. The classes are linked through a taxonomy consisting of parent-child relationships. Different from ActivityNet, THUMOS contains a large number of background videos making the problem of action recognition more realistic. For training the classifiers, the negative videos not only come from positive samples of other actions but the background videos associated with an action as well. Thus, it becomes crucial for the classifier and detector to accurately model the motion since similarity in scene in action and background videos significantly reduces the utility of appearance features. The background videos in THUMOS also aid in studying and quantifying the role of stationary and non-action context for action recognition (Sec. 8). Table 1 summarizes different characteristics of various action recognition datasets.

3. The THUMOS Challenge Tasks

This section gives an overview of the THUMOS *classification* and *temporal detection* tasks. We also describe their evolution since the first THUMOS held in 2013.

Table 1: Summary of major action recognition datasets.

Dataset	Number of Hours	Number of Actions	Background Type	Camera Motion	Released Year	Source	Background Videos	Untrimmed Videos	Evaluation Setup	Labels Per Video	Annotation Level
Weizmann [27]	0.06	9	Static	No	2005	Staged	No	No	Multi-class	Single	Label
UCF Sports [28]	0.27	10	Dynamic	Yes	2008	TV, Movies	No	No	Multi-class	Single	Label
IXMAS [38]	0.51	11	Static	No	2006	Staged	No	No	Multi-class	Single	Label
Olympic [39]	1.84	16	Dynamic	Yes	2010	YouTube	No	No	Multi-class	Single	Label
HOHA [31]	2.24	12	Dynamic	Yes	2009	Movies	No	No	Multi-class	Single	Label
UCF11 [7]	2.82	11	Dynamic	Yes	2009	YouTube	No	No	Multi-class	Single	Label
KTH [26]	3.22	6	Static	Slight	2004	Staged	No	No	Multi-class	Single	Label
HMDB51 [2]	5.92	51	Dynamic	Yes	2011	Movies, YouTube	No	No	Multi-class	Single	Label
UCF50 [32]	13.80	50	Dynamic	Yes	2010	YouTube	No	No	Multi-class	Single	Label
UCF101 [1]	26.67	101	Dynamic	Yes	2012	YouTube	No	No	Multi-class	Single	Label, Spatio-Temporal
ActivityNet (v1.2) [37]	305.55	200	Dynamic	Yes	2015	YouTube	No	Yes	Binary Detection	Multiple	Label, Temporal
THUMOS'14 [23]	254.00	101	Dynamic	Yes	2014	YouTube	Yes	Yes	Binary Detection	Multiple	Label, Temporal
THUMOS'15 [24]	430.00	101	Dynamic	Yes	2015	YouTube	Yes	Yes	Binary Detection	Multiple	Label, Temporal

3.1. Classification

The task of action classification consists of predicting (for each video) the presence or absence of each of the 101 action classes from the UCF101 dataset. This is a *binary classification* task per action, as the actions are not mutually exclusive — a given action may occur once, multiple times or never in a testing video. This is in contrast to the typical forced-choice multi-class task whose goal is to assign a class label to a given video from a set of pre-defined classes. For the classification task, the participants are expected to provide real-valued confidences for each test video for all the 101 actions. A low confidence for a particular action means either the video contains some other action or none of the 101 actions. The participants are required to report results on all the videos, and omitting videos from evaluation results in lower performance.

The classification task of the 2013 challenge only consisted of videos from UCF101. The dataset was divided into three pre-defined splits and participants reported results using three-fold cross-validation, i.e., training on two folds and testing on the third. However, since 2014 the dataset has been extended with untrimmed validation, background and test videos. The participants can only use UCF101, validation and background sets to train, validate and fine-tune their models and then report results on the withheld test set. Participants are not permitted to perform any manual annotation at their end.

3.2. Temporal Detection

For the temporal detection task participants are expected to provide temporal intervals and corresponding confidence values for all detected instances of 20 pre-selected action classes. The task of classification is embedded within the temporal detection which makes it comparatively more difficult. For example, an instance of an action that is correctly localized in time but is assigned with an incorrect class label will be treated as an incorrect detection. For this task, participants are required to report results for 20 action classes in all the test videos. For the detection tasks, similar to classification, participants are not permitted to perform additional manual annotations.

The first THUMOS challenge in 2013 included spatio-temporal localization for 24 action categories instead of temporal detection. The spatio-temporal annotations for 24 actions were provided in the trimmed videos of UCF101. The temporal detection resembles spatio-temporal localization with the difference that the spatial location of the detections is not incorporated in the evaluation. Besides the significant reduction in annotation effort, adopting temporal detection over spatio-temporal localization in later years of the THUMOS challenge was driven by two factors. First, temporal detection is computationally more tractable, particularly in long untrimmed videos. Second, in many practical scenarios, the temporal aspect is more important than the spatial, e.g., a user may want to seek directly to the portion of the video that includes the given action and may not benefit from a bounding box localizing the action within each frame. For these reasons, the 2014 and 2015 challenges only included a temporal detection task, with both the training and test set containing temporal annotations in untrimmed videos for the 20 actions.

4. The THUMOS Dataset

This section provides an overview of the data collection and annotation procedures. In addition, we also provide various statistics related to the THUMOS'15 dataset.

4.1. Video Collection Procedure

The Internet videos for the THUMOS competitions were drawn from public videos on YouTube, which made it possible to find a large number of videos for any given

topic — but a large fraction of videos may not contain visible instances of the desired action. We employed a series of manual filtering stages to ensure the set of videos for each action contains only the relevant videos.

Positive Videos: The YouTube Data API[4] allows video search through Freebase[5] topics. Every YouTube video has several Freebase topics associated with it that are assigned based on annotations provided by the video creator, as well based on some high level video features. We defined a set of Freebase topics corresponding to the action labels. However, a Freebase topic which ideally corresponds to an action either returns too few videos or too general to be useful. Therefore, we manually augmented topic ids with a set of search keywords. Keywords combined with Freebase topics yielded a reasonable set of potential videos for each action.

An issue with YouTube videos in context of our task is that highly rated or frequently viewed videos may include "viral" videos or compilations, so we had to exclude these by explicitly blacklisting keywords "-awesome", "-crazy", "-compilation", etc. Furthermore, as the dataset is extended each year by collecting new videos, we exclude all YouTube videos and channels whose videos were used in previous THUMOS competitions to avoid adding videos that might be similar to those from previous years.

Background Videos: Collecting *useful* background videos is more involved than searching for positive videos. Simply adding videos from unrelated categories does not help since such videos are visually dissimilar to those in the positive set. The best background videos are those that share the *context* of a given action (i.e., include similar scenes, actors and objects) without actually showing instances of the given action being performed. For instance, for the 'PlayingPiano' class, a video showing a piano in which the piano is not being played is a valid background video. It is also important that background videos for one action class do not contain positive instances of other actions. Therefore, for this task we grouped all action types into super classes. Several actions occur in similar settings: e.g., 'BalanceBeam', 'FloorGymnastics', 'Par-

[4] https://developers.google.com/youtube/v3/

[5] https://developers.google.com/youtube/v3/guides/searching_by_topic

allelBars', etc. are all likely to occur indoors in Olympic gymnastic venues; whereas 'HammerThrow', 'HighJump', 'HighJump', etc., occur outdoors in track and field arenas. To find such videos, we supplemented the search with the following queries which resulted in background videos without any instance of that action:

- **X + 'for sale'**: for actions that involve an instrument, e.g., piano for sale ('PlayingPiano'), yoyo for sale ('YoYo').
- **X + venue**: for actions that involve a particular location or venue, e.g. baseball stadium or Coors Field ('BaseballPitch'), climbing tower ('RockClimbing'), bathroom ('BrushingTeeth').
- **Co-occurring events**: for sports related actions, e.g., cheer leading or dance, e.g., waist twirling dance -hoop -contra ('HulaHoop').
- **X + brands**: for actions involving branded objects e.g., L'oreal eye makeup ('ApplyEyeMakeup').
- **X + 'drill' or 'workout'**: for some sports actions, e.g., shotput drill ('ShotPut').
- **X + 'review' or 'how to choose'**: for products, e.g., lipstick overview ('ApplyLipstick').
- **General Freebase topics**: excluding class names e.g., circus gymnastics ('StillRings'), computer ('Typing'), macramé ('Knitting').
- **Object names**: for actions involving object e.g., 'piano -playing' ('PlayingPiano'), bat ('CricketShot').
- **Different object / action combination**: mechanical bull ride ('PommelHorse'), Invisible drum ('PlayingTabla'), running with dog ('WalkingDog'), yoga standing pose ('Lunges').

The video collection procedure builds lists of putative positive and background videos for each action class. The *YouTube id*, *channel id*, and *title* of each video are saved in the list. Next, the videos go through an annotation stage, followed by downloading and final verification.

4.2. Annotation and Verification Procedure

The video collection procedure provides a set of potential positive and background videos for each of the 101 action classes. For positive videos, the annotators were asked to first go through the videos of a particular action class in UCF101, and then annotate the videos from the list as either *positive* or *irrelevant*. The videos for a particular action were presented to the annotator in a batch of four (for User Interface efficiency reasons), which were played simultaneously from YouTube. As soon as the annotator found a positive and valid instance of the action class being annotated, s/he marked it as positive. A video may contain an instance of an action, but was marked as *irrelevant* if it satisfied any of the following criteria:

- **Slow Motion**: The video contains action that has been performed in slow motion or in an unrealistic way, and looks different from the instances of an action class in UCF101 dataset.
- **Sped Up**: The action is being performed faster than usual.
- **Occlusions / Partial Visibility**: There is text or any other object significantly occluding the actor.
- **Motion Blur**: Video is blurry or camera is shaking to the extent that the action cannot be seen properly.
- **Clutter / Incorrect Background**: Action is performed in an environment where it is partially visible e.g., a 'GolfSwing' action recorded from a camera directly behind the audience, therefore they are blocking the field-of-view, or if it has an atypical backdrop, e.g., somebody performing 'PushUps' on the moon.
- **Unrealistic Instances**: The action does not seem realistic. For example, an instructional video on how to perform a 'PushUp' might have a person performing the action much slower than usual. The person might also stop half-way while performing the action to explain, or performs an action in an unusual way, not seen in the UCF101 dataset.
- **Animation**: Any animated examples of the action of interest, e.g. a character from a video game performing the action or from a cartoon, etc.
- **Fake Action**: The action does not seem realistic or is poorly performed.

- **Long Video**: Video is longer than 10 minutes.
- **Compilation**: Video is compiled using multiple videos.
- **Slide Show of Images**: The video contains a slide show of images, but no video of the action of interest.
- **First Person Video**: The video is recorded from an egocentric perspective by the same person who is performing the action i.e. actions viewed from a wearable camera.
- **Not Related**: The video neither contains any instance of the action of interest nor the background for that action.

The positive videos are also annotated with secondary actions, ones which occur or co-occur with the primary action in a video. Some of the actions are subset of others, for instance, 'BasketballDunk' implies 'Basketball', 'HorseRace' implies 'HorseRiding', and 'CliffDiving' implies 'Diving'. Similarly, there are several actions that are usually proximal in time, such as 'CricketBowling' and 'CricketShot', as well as videos involving playing of musical instruments that can have multiple secondary actions. In contrast to positive videos, the task of annotating background videos is somewhat more difficult as each background should not contain instances of any of the 101 action classes. To achieve this, each annotator was asked to review at most 34 actions at a time, and ensure none of those occurred in the background video being annotated. Thus, each background video was annotated by three different annotators for three distinct subsets of 101 action classes. Once the annotation is finished for positive and background videos, all of them are verified by a different set of annotators both for consistency and accuracy.

4.3. Temporal Annotations

Action boundaries (unlike objects) are generally vague and subjective. This makes the evaluation less concrete as human experts define the action boundaries differently from each other. The same is true for different methods whose output can vary among each other. However, we observed that the 101 action classes can be divided into two categories: the *instantaneous* actions which have short time span and can be well-

localized in time e.g., 'BasketballDunk', 'GolfSwing'; and *cyclic* actions that are repetitive in nature, e.g. 'Biking', 'HairCut', 'PlayingGuitar'. To select the action classes for the temporal detection task, we handpicked the instantaneous ones[6] with well-defined temporal boundaries (c.f. Appendix A).

Besides only focusing on instantaneous actions for the temporal detection, we also take additional measures to ensure that evaluation for this task is objective. First, we annotated action intervals consistently with the temporal segmentation of corresponding actions in the UCF101 dataset. Second, we also marked some action instances as ambiguous in cases of partial visibility, incomplete execution or strong deviation in the style. Third, we use a liberal Intersection-Over-Union threshold (small, 10%) to quantify the performance on this task, since actual actions are only a small fraction of the entire videos. Lastly, we ensured that evaluation at multiple IOU thresholds keeps the rankings unaffected.

For the 20 instantaneous actions selected for the task of temporal detection, we annotated their temporal boundaries in untrimmed videos. Each instance of these action classes is annotated with the start and end time in all videos in the Validation and Test sets. The labels include any of the 20 actions or *'ambiguous'*. To ensure consistency, the annotation has been made by one annotator in two passes over the data, and then verified by another annotator. The annotation has been performed using the Viper[7] tool. Action annotation for a few example videos is illustrated in Figure 3. In these and other examples each video typically contains instances of one action category only. Exceptions include 'CricketBowling' and 'CricketShot' actions which often co-occur within the same video.

[6] BaseballPitch (07), BasketballDunk (09), Billiards (12), CleanAndJerk (21), CliffDiving (22), CricketBowling (23), CricketShot (24), Diving (26), FrisbeeCatch (31), GolfSwing (33), HammerThrow (36), HighJump (40), JavelinThrow (45), LongJump (51), PoleVault (68), Shotput (79), SoccerPenalty (85), TennisSwing (92), ThrowDiscus (93), VolleyballSpiking (97).

[7] http://viper-toolkit.sourceforge.net/products/gt/

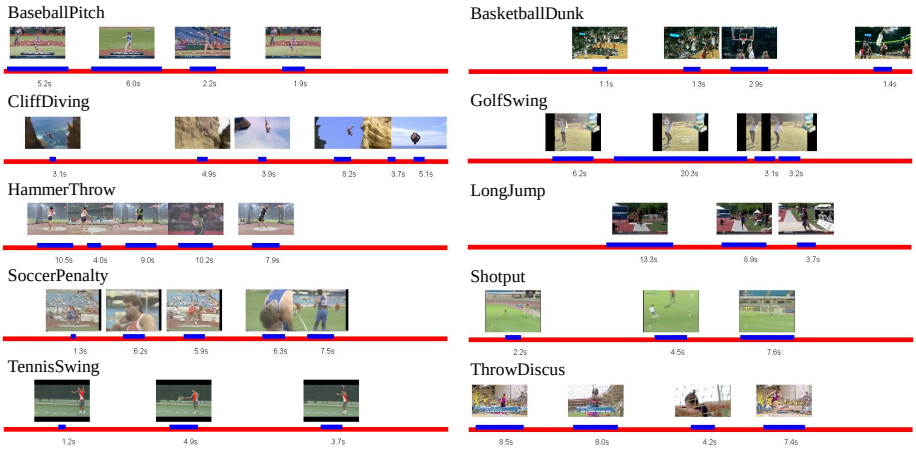


Figure 3: Illustration of temporal annotation (shown in blue) for eight video samples from the Validation set of THUMOS'15 dataset.

4.4. Attributes

Besides the video and clip level annotations provided with the THUMOS dataset, we also provided semantic relationships between the 101 action classes and several attributes. Each action class is associated with one or more of these attributes, as summarized in Table 2. Although video-level annotations for the attributes are not provided, such semantic knowledge can be incorporated while training and testing action categories.

4.5. Dataset Statistics

We summarize the statistics of THUMOS'15 benchmark dataset below:

- Validation set: 2,104 untrimmed videos with temporal annotations of actions. This set contains on average 20 videos for each of the 101 classes found in the UCF101 dataset.
- Background set: 2,980 relevant videos that are guaranteed not to contain any instances of the 101 actions.
- Test set: 5,613 untrimmed videos with temporal annotations for 20 classes.

Body Motion	***Body Parts Visible***	***Object***	***Indoor***
Flipping	*Head Closeup*	*Ball Like*	*Pool*
Walking	*Face Closeup*	*Big Ball Like*	*Office*
Running	*Upper Body*	*Stick Like*	*Court*
Riding	*Lower Body*	*Rope Like*	*Gym*
Up down	*Full Body*	*Sharp*	*Home*
Pulling	*One Hand*	*Circular*	*Track*
Lifting	*Two Hands*	*Cylinderical*	
Pushing	*One*	*Musical Instrument*	***Outdoor***
Diving	*Two*	*Portable Musical ...*	*Grass*
Jumping Up	*Many*	*...Instrument*	*Water*
Jumping Forward		*Animal*	*Ocean/Lake*
Jumping Over ...	***Body Parts Used***	*Boat Like*	*Court*
...Obstacle	*Head*		*Sky*
Spinning	*Hands*	***Posture***	*Street/Road*
Climbing Up	*Arms*	*Sitting*	*Track*
Horizontal	*Legs*	*Sitting In Front Of…*	
Vertical Up	*Foot*	*… Table Like Object*	
Vertical Down		*Standing*	
Bending		*Lying*	
		Handstand	
Body Part Articulation-Arm			
One Arm Motion	*Two Arms Bent*	*Legs Open To The Side*	*Facing Down*
Two Arms Motion	*One Arm Stretched*	*One Leg Bent*	*Facing Up*
Synchronized Arm Motion	*Two Arms Stretched*	*Two Legs Bent*	*Facing Front*
Alternate Arm Motion	*One Arm Swinging*	*One Leg Stretched*	*Facing Sideways*
One Arm Raised Over Head	*Two Arms Swinging*	*Two Legs Stretched*	*Straight Position*
Two Arms Raised Over Head	*Synchronized Leg Motion*	*Throw Release Motion*	*Tilted Position*
One Arm Raised Chest Level	*Alternate Leg Motion*	*Synchronized Hand Motion*	*Down Forward Motion*
Two Arms Raised Chest Level	*Fold Unfold Motion*	*One Hand Closed*	*Twist Motion*
One Arm Open To The Side	*Up Down Motion*	*Two Hands Closed*	*Bent Position*
Two Arms Open To The Side	*Up Forward Motion*	*One Hand Grab*	*Straight Up Position*
One Arm Down	*Side Stretch Motion*	*Two Hands Grab*	*Touching Ground*
Two Arms Down	*One Leg Raise*	*One Hand Open*	*In Air*
One Arm Bent	*Two Legs Raise*	*Two Hands Open*	

Table 2: Attributes for the 101 action classes.

The THUMOS'15, which is an extension of THUMOS'14 dataset, was designed to provide a realistic action recognition scenario. Unlike UCF101 [1], the videos in the set were not temporally segmented to contain only the actions of interest. Therefore, in most of the videos the action only takes a small percentage of time when compared to the length of the video in which it occurs (see Fig. 4) (the only notable exceptions are videos of cyclic actions). The use of variable length videos, each containing different numbers of actions of different lengths makes it less likely that a system could inadvertently exploit side-information [10], such as action length during the classification task. The mean clip length for UCF101 is 7.21 seconds, which is about 80% more than the average action length in the THUMOS'15 dataset.

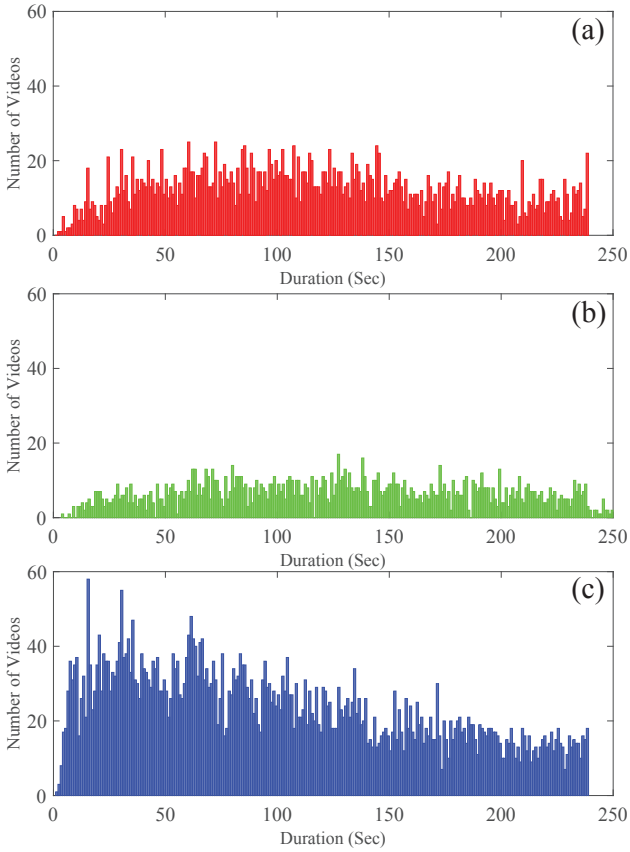


Figure 4: Histogram of video lengths in THUMOS'15 (a) Background, (b) Validation and (c) Test set, respectively. We excluded videos from the Validation set which were over 250 seconds long.

Statistics of the temporal annotation for the 20 action classes in the Validation set is presented in Table 3. As can be seen, the average length of such actions is $\sim$4.6 seconds while their temporal intervals occupy $\sim$28% of corresponding videos. The relatively large number of action instances and the low ratio of action length indicate the difficulty of the THUMOS temporal detection task.

Action ID	07	09	12	21	22	23	24	26	31	33	36	40	45	51	68	79	85	92	93	97	All
Instances	71	791	187	140	360	316	351	887	151	67	441	406	361	305	519	214	114	210	208	266	6365
Length (secs)	3.1	1.8	2.7	11.9	3.1	1.7	1.4	3.1	3.0	8.6	7.5	5.1	6.6	7.4	7.2	5.5	3.2	2.2	5.0	2.6	4.6
Ratio	12.2	24.0	14.5	47.8	27.1	13.5	11.6	29.7	38.2	30.3	40.8	31.3	24.6	31.5	40.2	33.7	19.2	22.6	39.3	28.5	28.0

Table 3: Statistics of temporal annotation for 20 action classes in the validation set of THUMOS'15 dataset. For each class the rows indicate (i) the number of actions instances, (ii) the average length of action intervals in seconds and (iii) the ratio of action length with respect to the length of the video.

5. Submission and Evaluation

5.1. Action Recognition

For action recognition, each system is expected to output a real-valued score indicating the confidence of the predicted presence in a video. Due to the untrimmed nature of the videos, a significant part of a test video may not include any particular action, and multiple instances may occur at different time-stamps within the video. Similarly, the video may not contain any of the actions, for which the expected confidence for each action is zero.

Each team was allowed to submit the results of at most five runs. The run with the best performance is selected as the primary run of the submission and is used to rank the teams. Each run has to be saved in a separate text file with 102 columns[8], where the first column contains the name of the test video, and rest of the columns contain confidences for the 101 actions. Essentially, each row shows the results of one test video, and each column contains the confidence score of presence of the corresponding action class anywhere in the video. The confidence scores must be between 0 and 1. A larger confidence value indicates greater confidence to detect the action of interest in a test video.

We use **Interpolated Average Precision (AP)** as the official measure for evaluating the results on each action class. Given a descending-score-rank of videos for the test action class c, the AP(c) is computed as:

$$AP(c) = \frac{\sum_{k=1}^{n}(\text{Prec}(k) \times \text{rel}(k))}{\sum_{k=1}^{n} \text{rel}(k)}, \quad (1)$$

[8]Sample output for Classification: http://goo.gl/sNQQBh

where n is the total number videos, Prec(k) is the precision at cut-off k of the list, rel(k) is an indicator function equaling to 1 if the video ranked k is a true positive, and to zero otherwise. The denominator is the total number of true positives in the list. Mean Average Precision (mAP) is then used to evaluate the performance of one run over all action classes.

5.2. Temporal Detection

Temporal detection is evaluated for twenty classes of instantaneous actions[6] in all test videos. The system is expected to output a real-valued score indicating the confidence of the prediction, as well as the starting and ending time for the given action[9]. For this task, each team is allowed to submit at most 5 runs. The run with the best performance is selected as the primary run of the submission and is used to rank across teams. Each run must be saved in a separate text file with the following format, where each row represents one detection output by the system:

[video name] [starting time] [ending time] [class label] [confidence score]

Each row has five fields representing a single detection. A detector can fire multiple times in a test video (reported using multiple rows in the submission file). The time must be in seconds with one decimal point precision. The confidence score should be between 0 and 1.

For evaluation, detected time intervals of a given class are sorted in the order of decreasing detector confidence and matched to ground truth intervals using Intersection over Union (IoU, also known as Jaccard) similarity measure. Detections with IoU above a given threshold are declared as true positives. To penalize multiple detections of the same action, at most one detection is assigned to each annotated action and the remaining detections are declared as false positives. Annotations with no matching detections are declared as false negatives. Given labels and confidence values for detections, the detector performance for an action class is evaluated by Average Precision

[9] Sample output for Temporal Detection: http://goo.gl/SWZbBM

(AP). The mean AP value for twenty action classes (mAP) provides the final performance measure for a method. To account for somewhat subjective definition of action boundaries, the evaluation is reported for different values of IoU threshold (10%, 20%, 30%, 40%, and 50%). Action intervals marked as ambiguous are excluded from the evaluation, hence, all detections having non-zero overlap with ambiguous intervals are ignored.

6. Methods

This section presents methods used by participants for both tasks at the THUMOS'15 challenge. A comprehensive survey of techniques and their evolution across years is beyond the scope of this paper, and will be made after several more challenges in the future.

6.1. Classification

In this subsection we briefly summarize the classification methods of the 11 teams. Table 4 summarizes the major feature extraction and fusion methods. Most teams adopted two kinds of features, deep learning based features and the improved Dense Trajectories (iDT) [3].

Deep learning features extracted by Convolutional Neural Networks (CNN) have been popular in many visual recognition tasks. By considering different network architectures and feature pooling methods, the resulting CNN features may vary greatly. For network architectures, VGGNet [51], GoogleNet [52], ClarifaiNet [53] and 3D ConvNets (C3D) [54] were used. In particular, VGGNet was used by most teams, and GoogleNet was used by three teams (UTS&CMU, CUHK&SIAT, UvA). Each of the remaining two networks was used by only one team (CUHK&SIAT used ClarifaiNet, and MSM used C3D), which are therefore excluded from the table due to space limitations. In addition, the recent two-stream CNN approach [5], which explores both spatial stream (static frames) and temporal stream (optical flows), was adopted by the CUHK&SIAT team.

For the CNN based models, typically the outputs of 6^{th}, 7^{th} or 8^{th} fully connected layers (FC6, FC7, FC8) are used as features. A few teams also explored a recent

Team	Deep Features: Structures & Encoding						Traditional Features			Fusion Methods			
	VGGNet	GoogleNet	FC 6,7,8	LCD	VLAD	Mean/Max Pool	iDT	MFCC	ASR	Average	Logistic Regression	Weighted	Geometric Mean
UTS & CMU [40]	•	•	•	•	•	–	•	•	•	–	•	–	–
MSR Asia (MSM) [41]	•	–	•	–	–	•	•	•	–	•	–	–	–
Zhejiang U. [42]	•	–	•	•	•	–	•	–	–	•	–	–	–
INRIA LEAR [43]	•	–	•	•	•	•	•	–	–	•	–	–	–
CUHK & SIAT [44]	•	•	•	–	–	•	•	–	–	•	–	–	–
U. Amsterdam [45]	–	•	•	–	•	–	•	–	–	•	–	–	–
Tianjin U. [46]	•	–	–	•	•	–	•	–	–	–	–	•	–
USC & THU [47]	•	–	•	–	–	•	•	–	–	–	–	–	•
U. Tokyo [48]	•	–	•	–	•	–	•	–	–	•	–	–	–
ADSC, NUS & UIUC [49]	•	–	•	–	–	•	•	–	–	•	–	–	–
UTSA [50]	•	–	•	–	–	•	–	–	–	–	–	–	–

Table 4: The major feature extraction and fusion methods of all the teams. Here, the symbols • and – represent the presence or absence of a feature or technique, respectively.

method called latent concept descriptors (LCD) [55]. In addition, as the CNN features are computed on video frames, a pooling scheme is needed to convert the frame-level feature into a video-level representation. For this, most teams adopted the Vectors of Locally Aggregated Descriptors (VLAD) [56] and the conventional mean/max pooling.

The iDT is probably the most powerful hand-crafted feature for video classification. It extracts four kinds of features, i.e., trajectory shape, HOG, HOF and MBH, on the spatial-temporal volumes along the extracted dense trajectories. The features are encoded with the Fisher Vector (FV) [57] to generate a video level representation. The UTS&CMU team used a variant of iDT, called enhanced iDT [58]. The UTS&CMU and the MSM teams also used auditory features MFCC and ASR.

For classification, all of the teams adopted SVM as the classifier. In addition, the

USC&Tsinghua team adopted kernel ridge regression (KRR) [59] as an alternative classifier. While the classifiers are consistent across the teams, the fusion method varies. As shown in the table, average fusion is the most popular option due to its simplicity and good generalizability, but there are other strategies like weighted fusion, logistic regression fusion, geometric mean fusion, etc.

6.2. Temporal Detection

This section summarizes the methods used for temporal detection of actions in testing videos. For the THUMOS'15 challenge, we received 5 runs from only one team. The team consists of researchers from Advanced Digital Sciences Center (ADSC), National University of Singapore (NUS), and University of Illinois Urbana-Champaign (UIUC). The temporal detection task attracted fewer participants compared to the classification task due to its higher computational requirements. Furthermore, temporal detection is a new problem that was introduced recently in THUMOS. With very few research efforts related to temporal detection in the past, we believe it will gain interest of the wider community resulting in increased participation in the future.

The runs from ADSC, NUS and UIUC were obtained using the following pipeline: First, the Improved Dense Trajectory (iDT) [3] features are extracted throughout the video. For forming the Gaussian Mixture Model dictionary, only features from UCF101 are used. The video segments are encoded using Improved Fisher Vectors. The FVs were not normalized to maintain additivity of Fisher Vectors. Besides the motion features, scene features were extracted from VGG-19 deep net model [60]. In particular, features were made from the last 4096-d rectified linear layer.

Since different actions have different lengths, the team used a pyramid of score distributions as features. For each frame, they used nine windows of 10, 20, ..., 90 frames around it. The hypothesis was that the scores at the correct window length should be highest, and should vary smoothly for neighboring temporal resolutions. Next, the FV in each window are normalized to obtain Improved FV. This yields 9×101 scores, which are concatenated to form a feature vector. The action confidences are then computed using a 21-class SVM (20 actions, 1 background). Afterwards, they use median filtering on output labels for smoothness.

7. Results

In this section, we present results and analysis of the approaches from the THUMOS'15 challenge presented in the previous section.

7.1. Classification

In this subsection, we summarize and discuss the results of the classification task. We received 47 submissions from the 11 teams. Table 5 shows the overall results of all the submissions, measured by mAP. The best mAP from each team is highlighted in bold. The teams are sorted based on their highest mAP.

Rank	Team	Run1	Run2	Run3	Run4	Run5
1	UTS & CMU [40]	**0.7384**	0.7157	0.7011	0.6913	0.647
2	MSR Asia (MSM) [41]	0.6861	0.6869	0.6878	0.6886	**0.6897**
3	Zhejiang U. [42]	**0.6876**	0.6643	0.6859	0.6809	0.5625
4	INRIA LEAR [43]	**0.6814**	0.6811	0.5395	0.6739	0.6793
5	CUHK & SIAT [44]	0.4894	0.5746	**0.6803**	0.6576	0.6604
6	U. Amsterdam [45]	**0.6798**	NA	NA	NA	NA
7	Tianjin U. [46]	**0.6666**	0.6551	0.6324	0.5514	0.5357
8	USC & THU [47]	0.6354	**0.6398**	0.6346	0.5639	0.6357
9	U. of Tokyo [48]	0.6159	0.6172	**0.6174**	0.6087	0.4986
10	ADSC, NUS & UIUC [49]	0.4471	0.3451	0.4849	**0.4869**	0.3466
11	UTSA [50]	**0.3981**	NA	NA	NA	NA

Table 5: Classification Results measured by mAP (%). Each team could submit up to five runs. The teams are sorted based on their highest mAP.

As discussed earlier, most of the approaches adopted two kinds of features: iDT features and deep learning features. iDT features were used by all the top-10 teams, and deep learning features were used by all the teams. Based on the results, we make the following observations: 1) The LCD coding with the VLAD representation [55] is very effective; 2) fine-tuning the CNN models can bring further improvements; and 3) some specially designed network structures for video analysis are helpful, e.g., the

Easy Classes	AP	Difficult Classes	AP
SkyDiving	0.964	Punch	0.198
PommelHorse	0.955	ShotPut	0.216
Rowing	0.933	Lunges	0.252
Skiing	0.925	BrushingTeeth	0.265
BalanceBeam	0.905	BreastStroke	0.273
Rafting	0.902	MoppingFloor	0.286
Surfing	0.881	Haircut	0.290
FloorGymnastics	0.875	Hammering	0.315
Drumming	0.873	PushUps	0.331
Bowling	0.872	BlowDryHair	0.347

Table 6: The top 10 easy and difficult classes in THUMOS'15.

two-stream CNN [5]. Furthermore, the results also indicate that multi-modal fusion with audio clues can consistently improve the results.

7.1.1. Per-action Results

Figure 5 shows the results of each action class, where the bars depict the AP of each action and the curve represents the results of all the actions sorted in decreasing AP values. For each action, the result is obtained by averaging the results of all the submissions. We can see that the AP varies significantly across different actions, from the lowest value of 19.8% to the highest of 96.4%. The curve of sorted AP fits well with a straight line, which indicates that the numbers of actions that are easy/hard to be distinguished are evenly distributed. The mAP over all the action classes is 61.3%, which reflects an average level of recognition capability of all the teams.

While the results are promising in general, there is still room for improvement. Table 6 lists the action classes which are easy or hard to be recognized. Some classes like 'Bowling' and 'Surfing' are easy but there are many difficult ones that can confuse the classifier. For example, 'BlowDryHair' is visually very similar to 'Haircut'. More advanced techniques are needed to distinguish these classes.

Figure 6 further shows the precision-recall curves. We plot the curves for a few

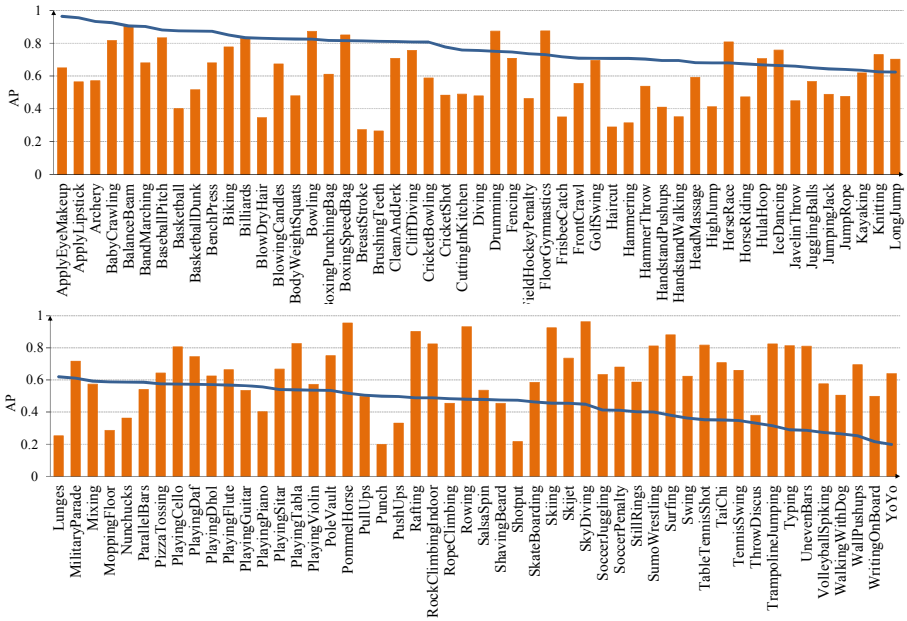
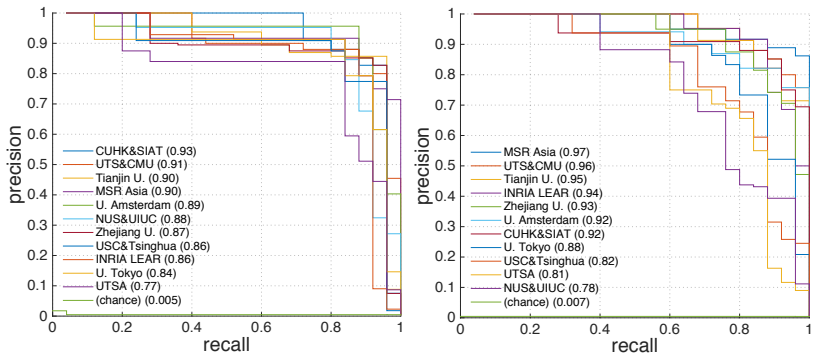


Figure 5: Per-action results, measured by AP: The bars depict the AP for each action, and the curve represents the results of all the actions sorted in decreasing AP values. For each action, we report the average AP from all the submissions.

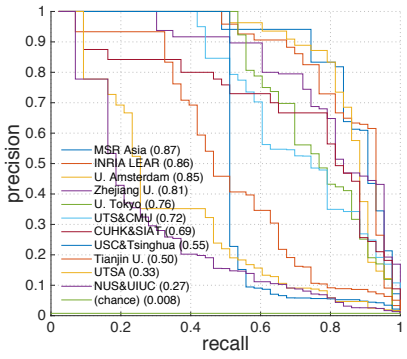
classes with high ('Bowling', 'Surfing'), medium ('CricketBowling', 'PlayingGuitar') and low ('BlowDryHair', 'Haircut') AP numbers. The team names in the legend of each figure are sorted by their AP values. Overall, the classes with higher accuracies tend to contain more unique/representative objects/scenes, while some difficult classes often share similar visual contents that are hard to be separated using state-of-the-art features (e.g., the classes 'BlowDryHair' and 'Haircut').

We also provide several representative frames from videos in Figures 7—12, respectively for the classes with precision-recall curves shown in Figure 6. The frames are selected based on the best run in THUMOS'15 (from the UTS&CMU team). For each class, we show the top-5 positive videos found by the best run in the first row, the bottom-5 positive videos in the second row, and the top-5 negative videos (false alarms) in the third row. As can be seen from the figures, the top ranked negative samples are all visually very similar to the positive ones, which demand more advanced features

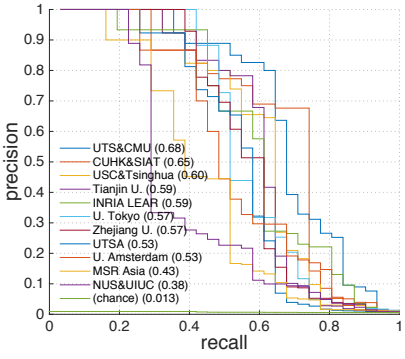


(a) Bowling

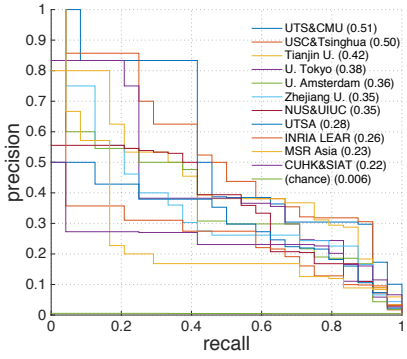
(b) Surfing



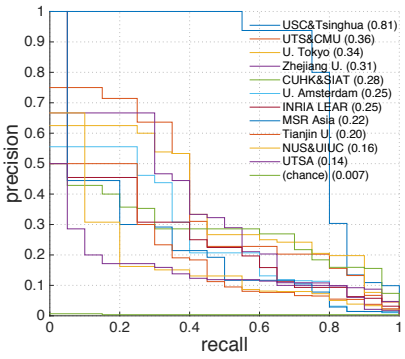
(c) CricketBowling



(d) PlayingGuitar



(e) BlowDryHair



(f) Haircut

Figure 6: Precision-recall curves of a few classes with high ('Bowling', 'Surfing'), medium ('CricketBowling', 'PlayingGuitar') and low ('BlowDryHair', 'Haircut') AP values.

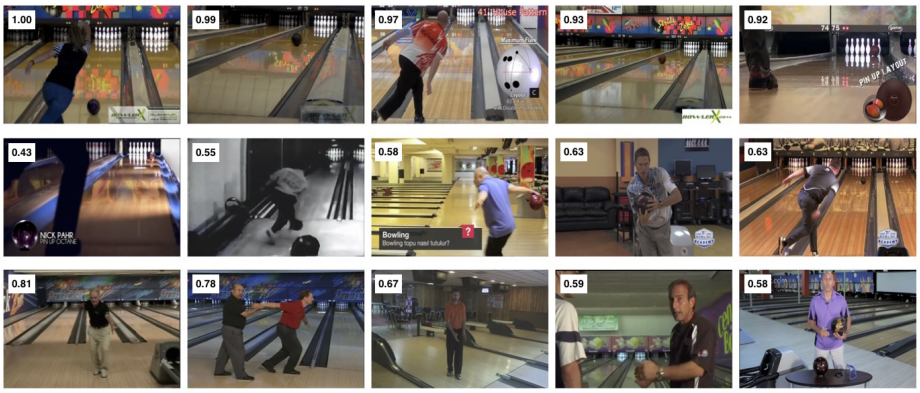


Figure 7: Video frames for class 'Bowling': First row: top-5 positive videos. Second row: bottom-5 positive videos. Third row: top-5 negative videos. Prediction scores are shown on the frames.

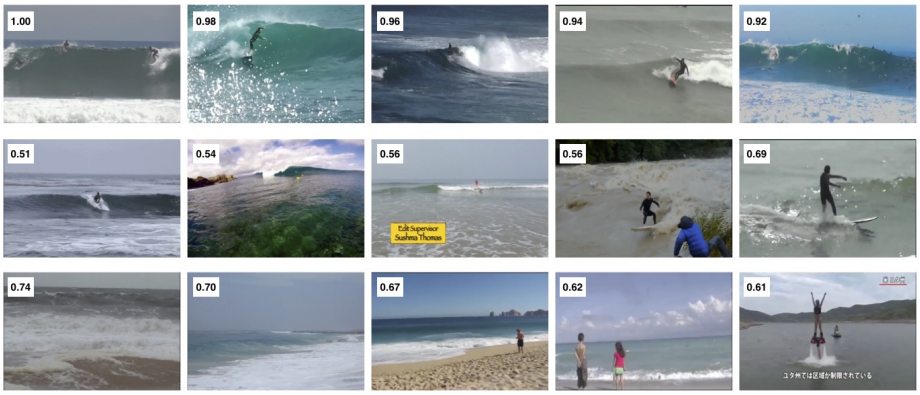


Figure 8: Video frames for class 'Surfing': First row: top-5 positive videos. Second row: bottom-5 positive videos. Third row: top-5 negative videos. Prediction scores are shown on the frames.

and classifiers to be correctly separated. We also observe that, for many classes that are easier to be recognized, they contain unique background scene settings. While for the difficult classes (e.g., 'BlowDryHair'), the actions may happen under different scene backgrounds. This indicates that current algorithms may significantly be relying on background scenes to support action recognition, not just focusing on the actions themselves.

7.1.2. Impact of Background Videos

We also evaluate the impact of background videos in Figure 13 which shows AP per-action with and without background videos in the test set. In this figure, the blue

Figure 9: Video frames for class 'CricketBowling': First row: top-5 positive videos. Second row: bottom-5 positive videos. Third row: top-5 negative videos. Prediction scores are shown on the frames.

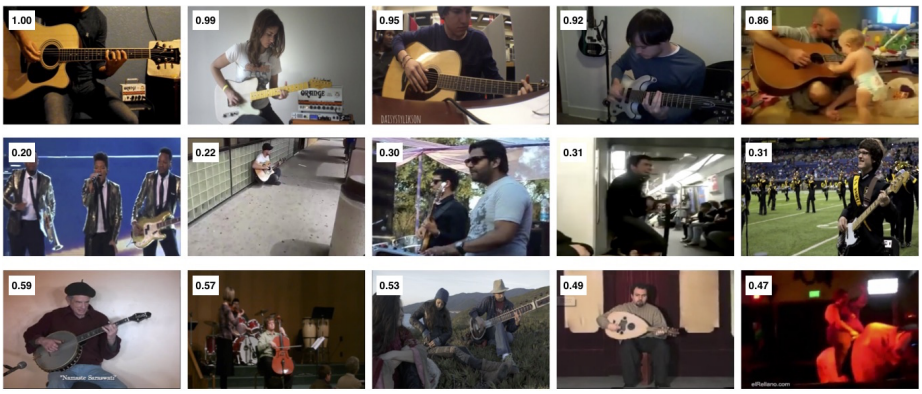


Figure 10: Video frames for class 'PlayingGuitar': First row: top-5 positive videos. Second row: bottom-5 positive videos. Third row: top-5 negative videos. Prediction scores are shown on the frames.

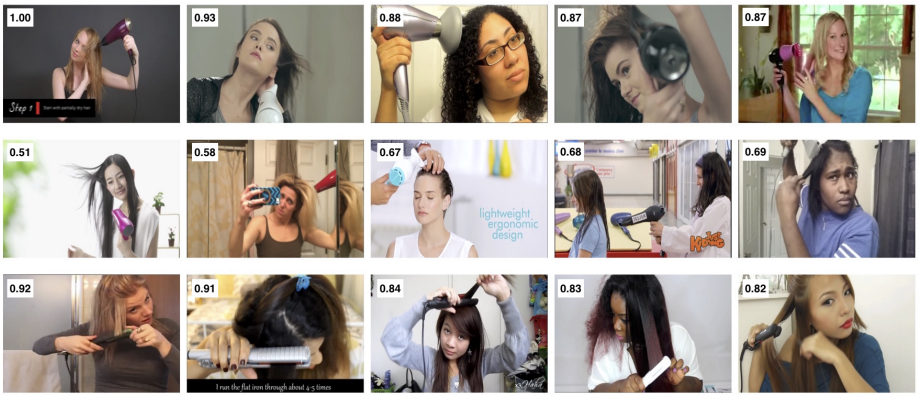


Figure 11: Video frames for class 'BlowDryHair': First row: top-5 positive videos. Second row: bottom-5 positive videos. Third row: top-5 negative videos. Prediction scores are shown on the frames.

Figure 12: Video frames for class 'Haircut': First row: top-5 positive videos. Second row: bottom-5 positive videos. Third row: top-5 negative videos. Prediction scores are shown on the frames.

histogram represents the results without background videos and the red histogram represents the official results with the background videos. Overall, the mAP after excluding the background videos is 76.3%, which is 15% higher than the results with the background videos (61.3%). This indicates that background videos have critical influence on the performance, which is easy to understand. Some classes like 'FrisbeeCatch', 'WalkingWithDog' and 'BlowDryHair' show significant performance degradation. The main reason is that the background videos contain samples that are visually (but not semantically) similar to these classes. Adding more negative samples during model training might be helpful for these classes. It would be interesting to study this in the future.

7.2. Temporal Detection

The results for the temporal detection task for THUMOS'15 are presented in Table 7. In this table, the mAP is computed at overlaps of 10%, 20%, 30%, 40% and 50%. Run1 from ADSC, NUS and UIUC has the best results compared to the other four runs, with mAP of $\sim$41% at an overlap of 10%. The difference between Run 1 and Runs 2-5 is the use of context features. Run 1 only uses iDT features, while others fuse appearance and scene features from deep networks. This is contradictory to the classification results, where fusion with appearance features in general, and features from deep networks, in particular, result in significant improvement in performance.

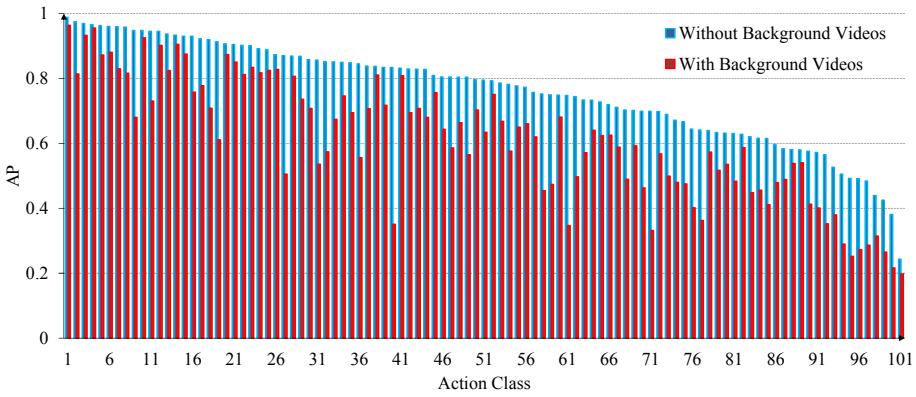


Figure 13: Effect of background videos: Blue histogram represents results without the background videos, and red histogram plots results including the background videos. Results are sorted based on the former. This figure is best viewed in color.

However, due to the nature of temporal detection task, the appearance of scene features cause a significant drop in performance. This is because for detection, it is important that the algorithm correctly detects the action, and does not produce false alarms on the rest of the positive videos. The appearance features reduce the discrimination between action segments and background within positive videos, and therefore result in drop in performance. Furthermore, ADSC, NUS and UIUC concluded that it is important to use multiple temporal scales while temporally localizing the actions. Using just a single scale (instead of 9) results in $\sim$30% drop in performance.

Figure 14 shows the per-action performance on the 20 classes. The action classes with high performance include 'HammerThrow', 'LongJump', and 'ThrowDiscus', whereas the classes with low performance include 'Billiards', 'ShotPut' and 'TennisSwing'. 'GolfSwing' and 'VolleyballSpiking' have the worse results of all. The results are correlated with the length of the actions, with short and swift actions such as 'GolfSwing' being the most difficult to localize.

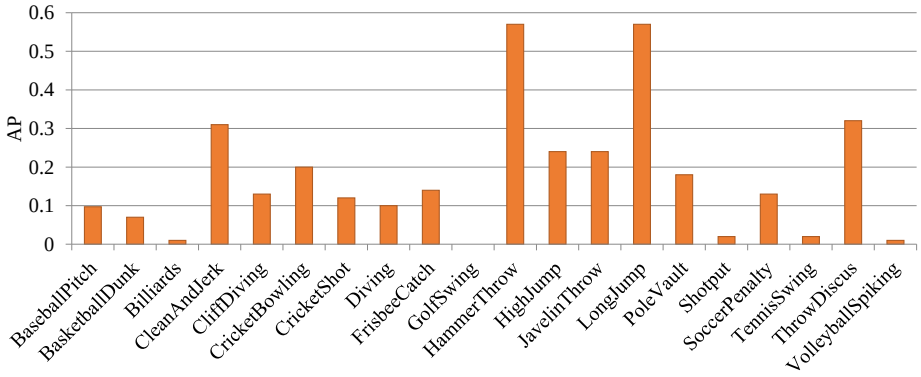


Figure 14: Per-Action Average Precision using ADSC, NUS and UIUC - Run1 on the 20 classes used for the temporal detection task.

Rank	Team-Run / Overlap	10%	20%	30%	40%	50%
1	ADSC, NUS & UIUC - Run1 [49]	**0.4086**	**0.3629**	**0.3076**	**0.2351**	**0.1830**
1	ADSC, NUS & UIUC - Run2 [49]	0.1611	0.1349	0.1072	0.0830	0.0562
1	ADSC, NUS & UIUC - Run3 [49]	0.1577	0.1346	0.1117	0.0882	0.0652
1	ADSC, NUS & UIUC - Run4 [49]	0.1386	0.1154	0.0939	0.0728	0.0510
1	ADSC, NUS & UIUC - Run5 [49]	0.1413	0.1180	0.0980	0.0773	0.0552

Table 7: Temporal Detection results measured by mAP (%). Each team can submit up to five runs. The percentages correspond to different values of overlaps.

8. Action Recognition in Untrimmed Videos

The past few decades of research on action recognition has primarily focused on trimmed videos that only contained an action of interest in each video. The lack of a dataset for untrimmed videos and preference of classification over detection task deviated the research on action recognition to focus on pre-segmented trimmed videos. Nevertheless, there have been a few approaches developed for classification [11, 12, 33, 39, 61, 62] and localization [13, 14, 29, 17, 20, 30] in untrimmed videos. However, the lack of a large-scale benchmark dataset of untrimmed videos was a pressing need that was first fulfilled in 2014 with the release of THUMOS'14. In this section,

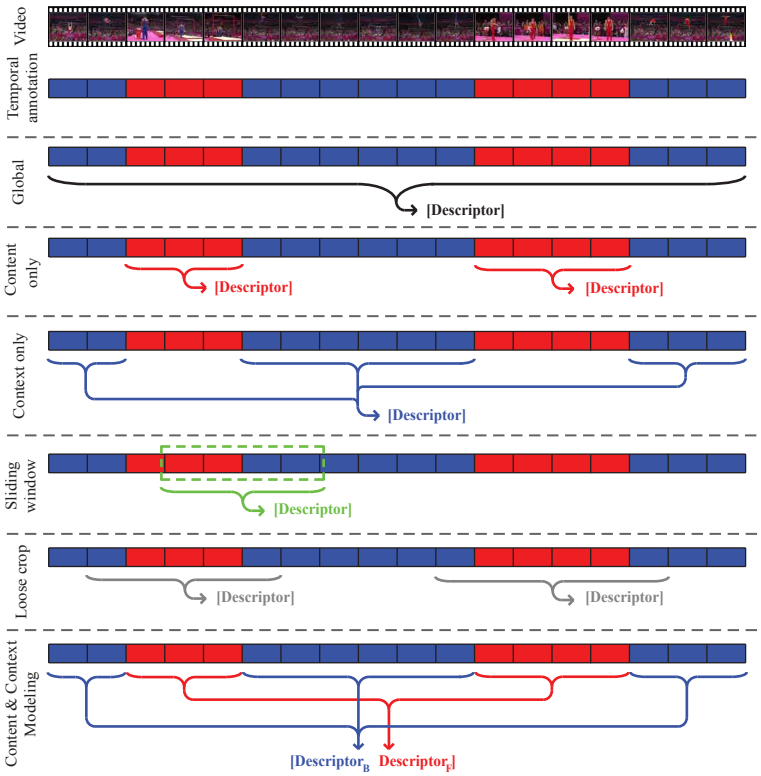


Figure 15: Video representations for action recognition in untrimmed videos. Here, red represents a positive action instance in the video whereas blue indicates the background portion.

we investigate classification performance of state-of-the-art action representations and learning methods in untrimmed setups where target actions occupy a relatively small part of longer videos. In particular, we explore the following questions:

- What are the important differences between trimmed and untrimmed videos for action recognition?
- How well methods designed for trimmed videos perform on untrimmed videos?
- What are the different approaches to represent content and context for action recognition in untrimmed videos?

Since we aim to study the role of actions (content) and background (context) in untrimmed videos - which requires temporal annotations - we perform experiments on the 20 action classes with manually annotated action intervals (see Section 4). Recall that the THUMOS'15 Validation set was formed by merging THUMOS'14 Validation and Test sets, and we collected a new Test set for THUMOS'15. For all the experiments in this section, we used THUMOS'14 Validation Set and/or THUMOS'14 Training Set (UCF101) for training, and the THUMOS'14 Test set for testing.

8.1. Representations

To systematically investigate the role of context or background, we construct several representations simulating different amounts of trimming around the action instances (content). These representations are illustrated in Figure 15 and are described below:

R1 - Global: In the global representation, we extract action descriptor from the full video without using any knowledge about the ground truth action intervals. This is the most straightforward application of standard techniques to untrimmed settings.

R2 - Content Only: Here we assume all action boundaries to be known and extract one descriptor for each action interval. This setup resembles the majority of common action methods and datasets with trimmed action boundaries.

R3 - Context Only: Video intervals outside action boundaries often correlate with temporally close actions and can provide contextual cues for action recognition. For example, tennis swing action co-occurs with running and typically appears on tennis courts. To investigate the effect of contextual cues, we extract descriptors from an entire video excluding action intervals.

R4 - Sliding Window: Here we do not use any knowledge about action boundaries and assume actions occupy compact temporal windows. We model the uncertainty in temporal position of an action and compute descriptors for overlapping windows of length 4 seconds using temporal stride of 2 seconds.

R5 - Loose crop: This setup is derived from the *Content Only* representation by gradually extending the initial action interval into background. We extend initial action

boundaries by 1, 3, and 7 seconds before and after the action. Note that the extension of temporal boundaries to the full video is equivalent to the *Global* representation above.

R6 - Content & Context Modeling: Given a mechanism that can separate content from context, this representation aims to understand if there is any benefit in modeling them separately. Therefore, we combine *Content Only* and *Context Only* representations by concatenating representations computed from action intervals and the temporal background.

8.2. Features

Local video features are a standard choice for action representation. We adopt common, standard, and well performing features, in particular Improved Dense Trajectory Features (IDTF) [3], to focus on experiments on various representations and methods. Following [3], we use HOF and MBH features based on optical flow to capture the motion information in the video. We also use HOG features based on the orientation of spatial image gradients to captures static information in the scene. All descriptors are computed in space-time volumes along 15-frames long point tracks, hence, they capture information in motion-aligned local neighborhood of a video.

To aggregate local features into video descriptors we use Fisher Vector encoding (FV) [63]. FV has been shown to consistently outperform histogram-based bag-of-feature aggregation techniques [4]. We use Gaussian Mixture Model with K=256 learned separately for each type of local feature, after reducing the dimensionality of HOG, HOF and MPH using PCA.

Since computing features is the most expensive step to represent video intervals with different temporal locations and temporal extents, we compute FVs for consequent chunks of 10 frames of a video without FV normalization independently for HOG, HOF and MBH. To obtain a FV descriptor for a given video interval, we used the additivity property of Fisher Vectors [64] by taking weighted sum of FVs corresponding to 10-frames chunks followed by L2 normalization. Thus, this approach allowed us to avoid re-computation of features for generating different representations as required by our setup.

8.3. Experimental Results

In this subsection, we report results and analysis of our experiments on action classification and temporal detection in untrimmed videos. We also investigate the role context plays in detecting actions in untrimmed videos. Context refers to the background portion of a positive video which does not contain any instance of the labeled action (R3). We evaluate the different representations in Section 8.1 to convert the localized (e.g., frame-level) annotations into video-level action labels: *Global, Content Only, Context Only, Sliding window, Loose Crop, and Content & Context modeling*.

8.3.1. Action Classification in Untrimmed Videos

We investigate the first five representations R1–R5 at test time and report action classification results. For training, we assume a fully-supervised setup with known action intervals. We use trimmed videos from the THUMOS'14 Training Set (UCF101) and annotated action instances from the THUMOS'14 Validation Set as positive samples for a particular action class, i.e., one descriptor per positive instance. For negative samples, we generate a single descriptor from each background video in THUMOS'14 Validation Set, and one descriptor per sliding window from the background portion of positive videos (*Context Only*). We learn one-vs-rest classifiers for all action classes, where the negative samples include positive instances from the other classes in addition to background samples. Table 8 summarizes the results of the video-level classification task. For each case, we report the mean average precision, reweighted by the number of instances in each test set. This makes the number of test instances identical for all cases and enables direct comparison between them. We make several observations:

- The *Global* case in the second row corresponds to the real-world deployment of a traditional action recognizer, which is trained on trimmed data (*Content Only*) and tested on features aggregated over an entire untrimmed test video. However, comparing this to *Context Only* in the first row is heartening: we confirm that the method is strongly influenced by the frames containing the action of interest (rather than context alone). Removing the action frames drops mAP from 0.68 to 0.46 for IDTF.

Training Setup	Testing Representation	mAP
Context Only (R3)	Global (R1)	0.46
Content Only (R2)	Global (R1)	0.68
Content Only (R2)	Content Only (R2)	0.72
Content Only (R2)	Sliding Window (average pooling) (R4)	0.77
Content Only (R2)	Sliding Window (max pooling) (R4)	**0.78**

Table 8: Comparison of the various training and aggregation representations. The mean average precision (mAP) presented is obtained after re-balancing, where we ensure that number of testing instances is identical for all the five cases. This is achieved through repeating each video proportional to the number of action instances contained within that video.

- The *Content Only* in the third row corresponds to the (artificial) scenario, where the action of interest is manually segmented from the untrimmed video, enabling each representation to be aggregated only over relevant frames. As expected, mAP improves from 0.68 to 0.72.
- The *Sliding Window* scenario is a systematic way (though computationally expensive) way to deploy an action recognizer trained on trimmed data on untrimmed videos. We see that it performs the best and that the choice of pooling strategy (max vs. average) has little impact, with max pooling (0.78 mAP) better by only 0.01.

Figure 16 shows results of these experiments individually for the 20 classes. We also investigate the reason for superior performance of *Sliding Window* approach over other cases. In this regard, Figure 17 shows examples of temporal detection results for several categories of sample videos. We note that the action of interest (black curve) rises above the average of responses from other actions (green curve) when the action is present. This explains why *Sliding Window* approaches work well for video-level classification with either form of pooling compared to the *Global* representation. The actions are usually much shorter than an entire untrimmed video and the detector gives better performance for those short durations. Another interesting result that highlights the difference between action recognition in trimmed and untrimmed videos is *Sliding Window* outperforming *Content Only* testing representation. The is primarily due to

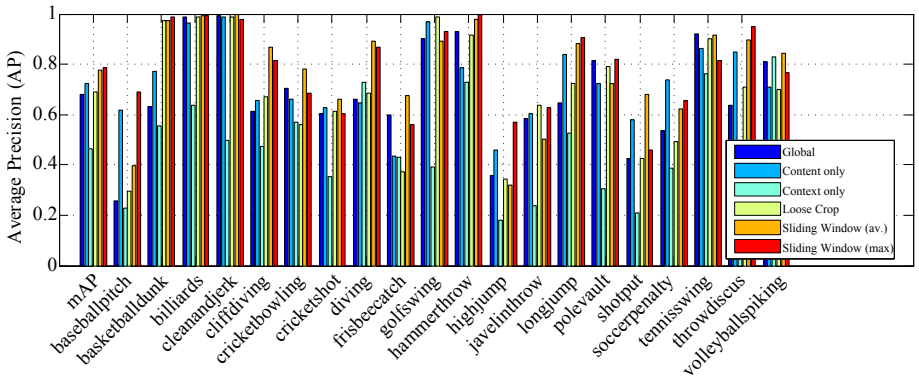
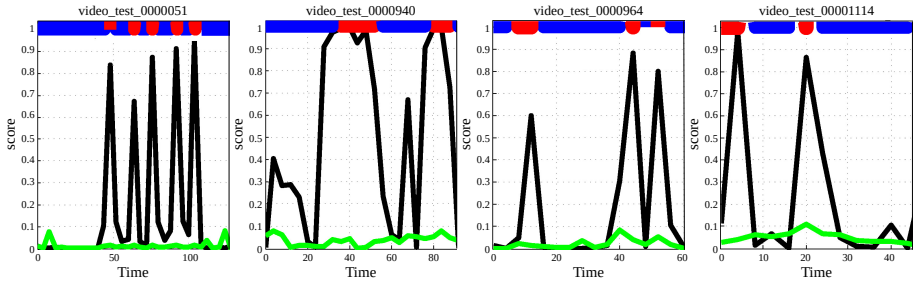


Figure 16: Classification Performance: This graph shows the results when training is performed on UCF-101 and trimmed action instances of the THUMOS'14 Validation subset, and testing is performed on THUMOS'14 Test set using representations R1 to R5. *Sliding Window* with both average and max pooling is reported in this graph.

the reason that untrimmed videos, and especially sports videos, usually contain multiple instances of a particular action. Then, pooling simply makes results robust by aggregating scores over multiple instances. A video can get a high score if most of the instances in it obtain high scores, thus, average pooling serves as a regularizer. Similarly, in max pooling, if one instance obtains a high score, then the entire video gets that score and weaker detections within the video benefit as a consequence. Remember that we evaluate the performance of all methods by first obtaining score at the video level, and then reweigh each video with the number of instances within it.

We also performed experiments for different parameters of *Sliding Window* (R4) and *Loose Crop* (R5) with results shown in Table 9. For *Loose Crop* experiments in the first five rows, the performance of action classification drops as window length is increased around the action instance. The 120 second loose crop corresponds to the *Global* (R1) case as can be seen with mAP of 0.68 from Table 8. The results for *Sliding Window* (R4) are shown in the bottom part of Table 9. The optimal performance is achieved when the window length is 4 seconds and drops when it is either smaller or larger. This is because the average duration of actions for the 20 classes is around 3.75 seconds, and thus the detector output is optimized around this window length.



(a) *Billiards* (b) *Cliff Diving* (c) *Baseball Pitch* (d) *Shotput*

Figure 17: Examples of temporal detection scores over time: This figure shows four graphs for different actions. The x-axis is time along the video and y-axis shows the detector scores. The ground truth is shown at the top in red (action) and blue (background). The black curve is detector score for the ground truth action, whereas green curve shows mean of scores from all other detectors. Note that the action of interest rises above the mean response when the action is present, showing why sliding window works well for video-level classification (when pooled) as well as temporal detection.

Nonetheless, the drop in performance is nominal for longer windows and shows *Sliding Window* is not sensitive to window length.

8.3.2. Role of Context for Classification in Untrimmed Videos

Context plays an important role in the ability of the classifiers to make good predictions. However, context alone is not sufficient for obtaining good performance. Removing the action of interest from training decreases performance from 0.68 mAP to 0.46 mAP (Table 8). The mAP for different runs evaluating the role of context are summarized in Table 10, while Fig. 18 shows the same for the 20 concepts individually. This particular experiment evaluates on *Content & Context* (R6) representation and thus the training data requires untrimmed videos containing action instances. Thus, we cannot use UCF101 since its videos are trimmed (no additional context), and background videos from THUMOS'14 Validation set that do not contain content. The training is performed on positive videos from the THUMOS'14 Validation Set, while testing performed on THUMOS'14 Test Set.

In Fig. 18, the blue bars denote the *Global* descriptor for untrimmed videos (R1), light-blue shows *Context Only* (R3), yellow depicts *Content Only* (R2) i.e., trimmed actions, while red marks the results obtained by concatenating descriptors for *Content*

Testing Representation	Window Length	Pooling	mAP
Loose Crop (R5) (1FV per loose GT window)	0 sec loose	-	**0.72**
	1 sec loose	-	0.71
	3 sec loose	-	0.69
	7 sec loose	-	0.69
	120 sec loose	-	0.68
Sliding Window (R4) (1FV per sliding window)	2 sec long	Max	0.76
	2 sec long	Average	0.77
	4 sec long	Max	**0.78**
	4 sec long	Average	0.77
	7 sec long	Max	0.77
	7 sec long	Average	0.76
	10 sec long	Max	0.76
	10 sec long	Average	0.76

Table 9: Video classification with *Loose Crop* (R5) and *Sliding Window* (R4): For all experiments we train models on UCF101 (1FV per video) + background set (1FV per video) + Validation (1FV per GT window, 1FV for each sliding window on the background part).

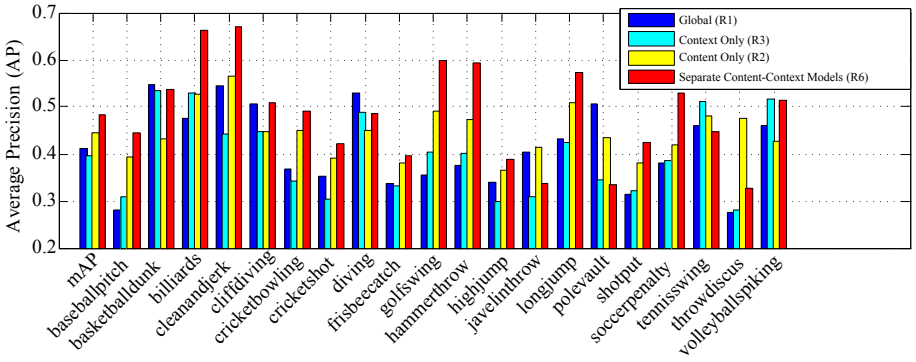


Figure 18: This graph shows average precision (AP) for 20 actions using the THUMOS'14 Validation and Test sets for different combinations of content and context for representing the videos.

Training Setup	Testing Representations	mAP
1FV per GT win, 1FV for each sliding win on BG	Global (R1)	0.42
1FV per GT win, 1FV for each sliding win on BG	Content only (R2)	0.45
1FV per GT win, 1FV for each sliding win on BG	Context only (R3)	0.39
1FV per GT win + 1FV for BG	Content & Context (R6)	0.49

Table 10: This table shows the experimental results on different approaches to handling context. The training is performed using positive videos of THUMOS'14 Validation Set and testing is performed on THUMOS'14 Test Set.

& Context (R6). The graph reveals an important insight that context described separately but used in conjunction with content gives the best performance compared to training using *Content Only* (R2). Therefore, gains in performance can be achieved through separate modeling content and context for action classification. For this run, we used information about action boundaries during testing. In realistic scenario, this is expected to be obtained with methods that can generate generic action proposals.

8.3.3. Temporal Detection in Untrimmed Video

We also report some results for the task of temporal detection on 20 action classes. In this case, we use the same training setup as for action classification using Training and Validation subsets. At test time we use the classifier in a sliding window manner in combination with temporal non-maximum suppression to select a single action interval

for each action hypothesis on the THUMOS'14 Test set. Fig. 19 reports AP per class using sliding windows. IDTF achieves a mAP of 0.67 on this task. Furthermore, a sliding window for 4 seconds outperforms that of 2 seconds by a margin of 0.03.

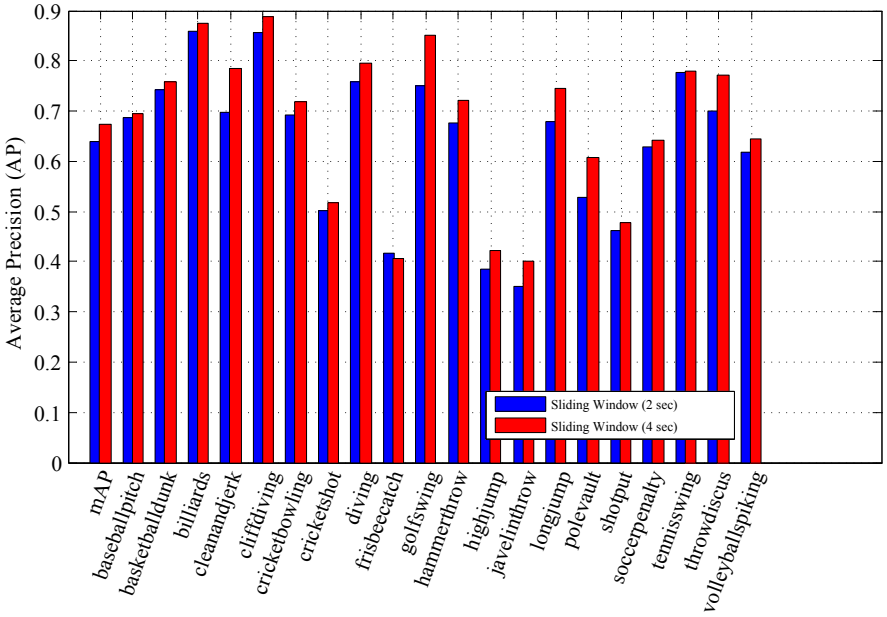


Figure 19: This figure shows the temporal detection performance on 20 action classes. The blue and red bars represent sliding windows of 2 and 4 seconds, respectively.

9. Future Directions

There are several thrusts for improving action recognition, we focus on two of them in line with the THUMOS challenge: the dataset and evaluation tasks that quantify performance on different aspects of action recognition. We believe having a denser, more comprehensive, and more generalizable understanding of a video is the way forward. One possibility is to introduce the **spatio-temporal localization** task in weakly supervised setting, where training is performed on untrimmed videos without the availability of frame-level annotations or bounding boxes. The test set, however, can contain frame-level and bounding box annotations for the detection and localization tasks, respectively.

THUMOS'15 contains about 13,000 trimmed videos for training the 101 action classes, as well as approximately 10,500 untrimmed videos in the validation and test sets. The dataset amounts to $\sim$370 gigabytes of data, making it the largest dataset for actions and activities. However, there is still room for extension in the THUMOS'15 dataset, which despite being the largest video dataset for action recognition, is still deficient both in the number of classes as well as number of instances per class. One possible approach is to define action and activity classes associated with a variety of *verbs*. This will result in the most comprehensive set of classes specifically aimed at capturing human motion. The number of classes will be several times larger than current dataset, with at least 200 instances per action. The space requirement are expected to be on the order of terabytes.

Moreover, the aim should be to move from visual (appearance and motion) perception in videos to a deeper semantic understanding by describing different objects, actions and their interaction among themselves and the environment in terms of attributes, semantic relationships and textual descriptions. Hence, the goal is not only to detect objects and actions, but also explain their complex spatial and temporal interactions. For this, one idea is to add a wide variety of videos with primary focus on actions and activities performed by humans, both as individuals or in groups, and then perform **dense annotations** for objects, actions, scenes, attributes, and the inter-relationships between objects, actions and environment.

For assigning labels to objects, actions and scenes, WordNet can be used as it allows modeling of structured knowledge. The WordNet Synsets can be used to relate the different nouns, verbs and adjectives. Here, it will be important to consider the trade-off between **consistency** and **diversity**. The consistency requires that labels are reused, so that a particular object or action has the same label across videos. However, this objective conflicts with diversity, as it limits the number of new labels that can be assigned to objects and actions. For instance, the terms 'person' and 'man' might refer to the same subject. Similarly, the actions 'jump' and 'plunge' are interchangeable in some contexts. WordNet Synsets are able to relate these words as 'person' is a hyperonym of 'man', and 'jump' and 'plunge' are synonyms. Thus, the trade-off could be controlled by preferring specific labels over more general labels and using them consis-

tently, however, other specific labels can be used whenever relevant and available. This will also allow the transfer of many appearance attributes directly from WordNet. The label 'grass' will be immediately labeled with green due to the structured knowledge available in WordNet. Indeed, this will require verification from the annotators, but the transfer of attributes and properties will save time and effort while generating richer and dense annotations for a large video dataset.

For cognitive understanding of videos, it is important that training data contains detailed annotations about how the objects, actions and scenes interact with each other. Moreover, qualitative properties of objects and actions, termed **attributes** also add to the semantic understanding of video data. Both *appearance attributes* that capture the visual qualities of objects including color, size, shape, as well as *motion attributes* which are related to the actor, such as the body parts used, their articulation, and type and speed of movement etc. should preferably be included. Next, these relationships can be expressed using a structured representation with WordNet. For instance, a man playing violin could be *playing (man, violin)*, and a woman holding eye brush as *holding (woman, eye brush)*. Once these relationships have been constructed for objects, actions, scenes and attributes, they can be merged together to form a graphical representation. The annotators will verify the validity of tree-graphs relating nouns, verbs and adjectives.

The annotations can also be supplemented with text, as the ability to produce valid text descriptions of videos is one of the measures of cognitive and high-level understanding. Also, **textual descriptions** may be added for all interesting occurrences and events in a video by first annotating with bounding boxes and tubes. Different video regions can have both spatial and temporal overlap with each other, and a description of their own. For instance, to be able to detect the action 'BasketballDunk', one only needs to detect the person performing the action. However, for high-level reasoning such as whether the actor is performing the action independently during practice, or while playing a game with others, it is important that we are able to locate all other objects and detect behaviors of other actors in the video. These dense text captions for each video region will give local summaries and help train better models for cognitive video understanding. The descriptions can be written in third-person present tense, and

be verified for vocabulary and grammatical consistency.

Region-level descriptions in addition to shots selected for summarization through manual annotation can allow evaluation of video-to-text approaches as well. While annotating the videos for descriptions, it is important that the textual summary for regions are not repeated and are diverse enough to delineate the events captured in the video. This can be achieved in an online manner, where new descriptions from an annotator will be n-gram matched to existing descriptions, and highly matching descriptions will be flagged for an immediate update.

Finally, with the graphical structure representing the objects, actions and attributes in addition to the textual descriptions for regions, it is straightforward to create **Question and Answer pairs** that go beyond the detection and localization and allow computers to exhibit cognitive understanding. These questions should emphasize the motion of actions, such as:

- Which hand did the person use to apply makeup? Which eye?
- How long did the person hold the arrow in the bow?
- Was the baby crawling on his/her belly?
- What instrument was the person playing?
- Where were the people ice dancing?
- Who was performing gymnastics?

Conclusion

This paper describes the THUMOS dataset and the challenge is detail. The two tasks include action classification and temporal detection. We presented an overview of the relationship of THUMOS to existing datasets, the procedure used to collect and annotate thousand of videos. Furthermore, we described evaluation metrics used in the challenge and methods and analysis of results for the THUMOS'15 competition. Next, we presented a study on untrimmed videos which were introduced in the 2014 challenge. The results show that sliding window outperforms global representation, and separate modeling of content and context is certainly helpful for improving the perfor-

mance. We also presented several directions to improve the challenge and proposed spatio-temporal localization and weakly supervised action recognition tasks in the future challenges. Finally, by providing a large-scale benchmark dataset of untrimmed videos to the vision community constituting dense annotations of objects, actions and textual descriptions, we hope to foster research in holistic understanding of video data.

Acknowledgement: We thank Saad Ali, Mikel Rodriguez, Aakif Tanveer, Shahzad Aziz, Jingen Liu, Khurram Soomro, Jiebo Luo, George Toderici, Massimo Piccardi and the numerous diligent annotators for their contributions to the UCF action datasets and the THUMOS challenge.

References

[1] K. Soomro, A. R. Zamir, M. Shah, UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild, Tech. Rep. CRCV-TR-12-01, UCF, 2012. 2, 3, 4, 6, 7, 10, 18

[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: IEEE ICCV, 2011. 2, 4, 6, 10

[3] H. Wang, C. Schmid, Action Recognition with Improved Trajectories, in: IEEE ICCV, 2013. 2, 4, 22, 24, 36

[4] D. Oneata, J. Verbeek, C. Schmid, Action and Event Recognition with Fisher Vectors on a Compact Feature Set, in: IEEE ICCV, 2013. 2, 36

[5] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: NIPS, 2014. 2, 22, 26

[6] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: IEEE CVPR, 2009. 2, 6

[7] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos “in the wild”, in: IEEE CVPR, 2009. 2, 4, 6, 10

[8] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: IEEE CVPR, 2012. 2

[9] M. S. Ryoo, L. Matthies, First-Person Activity Recognition: What Are They Doing to Me?, in: IEEE CVPR, 2013. 2

[10] S. Satkin, M. Hebert, Modeling the temporal extent of actions, in: ECCV, 2010. 3, 18

[11] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, J. Sivic, Weakly supervised action labeling in videos under ordering constraints, in: ECCV, 2014. 3, 33

[12] O. Duchenne, I. Laptev, J. Sivic, F. Bach, J. Ponce, Automatic annotation of human actions in video, in: IEEE ICCV, 2009. 3, 33

[13] M. Hoai, Z.-Z. Lan, F. De la Torre, Joint segmentation and classification of human actions in video, in: IEEE CVPR, 2011. 3, 33

[14] H. Pirsiavash, D. Ramanan, Parsing videos of actions with segmental grammars, in: IEEE CVPR, 2014. 3, 33

[15] Z. Shou, D. Wang, S.-F. Chang, Action Temporal Localization in Untrimmed Videos via Multi-stage CNNs, in: IEEE CVPR, 2016. 3

[16] A. Richard, J. Gall, Temporal Action Detection using a Statistical Language Model . 3

[17] Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: IEEE ICCV, 2007. 3, 33

[18] A. Klaser, M. Marszałek, C. Schmid, A. Zisserman, Human Focused Action Localization in Video, in: International Workshop on Sign, Gesture, and Activity (SGA), ECCV Workshops, 2010. 3

[19] I. Laptev, P. Pérez, Retrieving actions in movies, in: IEEE ICCV, 2007. 3

[20] Y. Tian, R. Sukthankar, M. Shah, Spatiotemporal deformable part models for action detection, in: IEEE CVPR, 2013. 3, 33

[21] K. Soomro, H. Idrees, M. Shah, Action localization in videos through context walk, in: IEEE ICCV, 2015. 3

[22] K. Soomro, H. Idrees, M. Shah, Predicting the Where and What of actors and actions through Online Action Localization, in: CVPR, 2016. 3

[23] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, R. Sukthankar, THUMOS'14: ECCV Workshop on Action Recognition with a Large Number of Classes, http://crcv.ucf.edu/THUMOS14/, 2014. 3, 10

[24] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, R. Sukthankar, THUMOS Challenge: Action Recognition with a Large Number of Classes, http://www.thumos.info/, 2015. 3, 10

[25] Y.-G. Jiang, J. Liu, A. R. Zamir, I. Laptev, M. Piccardi, M. Shah, R. Sukthankar, THUMOS'13: ICCV Workshop on Action Recognition with a Large Number of Classes, http://crcv.ucf.edu/ICCV13-Action-Workshop/, 2013. 4

[26] C. Schuldt, I. Laptev, B. Caputo, Recognizing Human Actions: A local SVM Approach, in: ICPR, 2004. 4, 5, 10

[27] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as Space-Time Shapes, in: IEEE ICCV, 2005. 4, 5, 10

[28] M. Rodriguez, J. Ahmed, M. Shah, Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition, in: IEEE CVPR, 2008. 4, 6, 10

[29] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: IEEE ICCV, 2005. 5, 33

[30] J. Yuan, Z. Liu, Y. Wu, Discriminative Subvolume Search for Efficient Action Detection, in: IEEE CVPR, 2009. 5, 33

[31] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning Realistic Human Actions from Movies, in: IEEE CVPR, 2008. 6, 10

[32] K. K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Machine Vision and Applications 24 (5) (2013) 971–981. 6, 10

[33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE CVPR, 2014. 6, 33

[34] S. Ayache, G. Quénot, Video corpus annotation using active learning, in: European Conference on Information Retrieval, Springer, 187–198, 2008. 8

[35] E. Yilmaz, J. A. Aslam, Estimating average precision with incomplete and imperfect judgments, in: Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, 2006. 8

[36] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, M. Michel, Creating HAVIC: Heterogeneous Audio Visual Internet Collection, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, ISBN 978-2-9517408-7-7, 2012. 9

[37] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: IEEE CVPR, 2015. 9, 10

[38] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 1–7, 2007. 10

[39] J. C. Niebles, C.-W. Chen, F.-F. Li, Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification, in: ECCV, 2010. 10, 33

[40] Z. Xu, L. Zhu, Y. Yang, A. Hauptmann, UTS-CMU at THUMOS 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[41] Z. Qiu, Q. Li, T. Yao, T. Mei, Y. Rui, MSR Asia MSM at THUMOS Challenge 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[42] K. Ning, F. Wu, ZJUDCD Submission at THUMOS Challenge 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[43] X. Peng, C. Schmid, Encoding Feature Maps of CNNs for Action Recognition, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[44] L. Wang, Z. Wang, Y. Xiong, Y. Qiao, CUHK&SIAT submission for THUMOS'15 action recognition challenge, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[45] M. Jain, J. C. van Gemert, P. Mettes, C. G. Snoek, I. ISLA, University of Amsterdam at THUMOS 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[46] Y. Liu, B. Fan, S. Zhao, Y. Xu, Y. Han, Tianjin University Submission at THUMOS Challenge 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[47] C. Gan, C. Sun, R. Kovvuri, R. Nevatia, USC & THU at THUMOS 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[48] K. Ohnishi, T. Harada, MIL-UTokyo at THUMOS Challenge 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[49] J. Yuan, Y. Pei, B. Ni, P. Moulin, A. Kassim, ADSC Submission at THUMOS Challenge 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25, 33

[50] J. Cai, Q. Tian, UTSA submission to THUMOS 2015, in: THUMOS'15 Action Recognition Challenge, 2015. 23, 25

[51] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: arXiv preprint arXiv:1409.1556, 2014. 22

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE CVPR, 2015. 22

[53] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: ECCV, 2014. 22

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: ICCV, 2015. 22

[55] Z. Xu, Y. Yang, A. G. Hauptmann, A discriminative CNN video representation for event detection, in: IEEE CVPR, 2015. 23, 25

[56] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE CVPR, 2010. 23

[57] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, IJCV 105 (3) (2013) 222–245. 23

[58] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, B. Raj, Beyond gaussian pyramid: Multi-skip feature stacking for action recognition, in: IEEE CVPR, 2015. 23

[59] S.-I. Yu, L. Jiang, Z. Mao, X. Chang, X. Du, C. Gan, Z. Lan, Z. Xu, X. Li, Y. Cai, et al., Informedia@ TRECVID 2014 MED and MER, in: NIST TRECVID Video Retrieval Evaluation Workshop, 2014. 24

[60] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: BMVC, 2014. 24

[61] M. Raptis, L. Sigal, Poselet Key-framing: A Model for Human Activity Recognition, in: IEEE CVPR, 2013. 33

[62] K. Tang, L. Fei-Fei, D. Koller, Learning latent temporal structure for complex event detection, in: IEEE CVPR, 2012. 33

[63] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: ECCV, 2010. 36

[64] D. Oneata, J. Verbeek, C. Schmid, Efficient action localization with approximately normalized Fisher vectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 36

Appendix A. List of 101 actions

The complete list of actions for UCF 101 and THUMOS is provided below. The actions in **bold face** are used in the evaluation of the temporal detection task.

1. ApplyEyeMakeup
2. ApplyLipstick
3. Archery
4. BabyCrawling
5. BalanceBeam
6. BandMarching
7. **BaseballPitch**
8. Basketball
9. **BasketballDunk**
10. BenchPress
11. Biking
12. **Billiards**
13. BlowDryHair
14. BlowingCandles
15. BodyWeightSquats
16. Bowling
17. BoxingPunchingBag
18. BoxingSpeedBag
19. BreastStroke
20. BrushingTeeth
21. **CleanAndJerk**
22. **CliffDiving**
23. **CricketBowling**
24. **CricketShot**
25. CuttingInKitchen
26. **Diving**
27. Drumming
28. Fencing
29. FieldHockeyPenalty
30. FloorGymnastics
31. **FrisbeeCatch**
32. FrontCrawl
33. **GolfSwing**
34. Haircut
35. Hammering
36. **HammerThrow**
37. HandstandPushups
38. HandstandWalking
39. HeadMassage
40. **HighJump**
41. HorseRace
42. HorseRiding
43. HulaHoop
44. IceDancing
45. **JavelinThrow**
46. JugglingBalls
47. JumpingJack
48. JumpRope
49. Kayaking
50. Knitting
51. **LongJump**
52. Lunges
53. MilitaryParade
54. Mixing
55. MoppingFloor
56. Nunchucks
57. ParallelBars

58. PizzaTossing
59. PlayingCello
60. PlayingDaf
61. PlayingDhol
62. PlayingFlute
63. PlayingGuitar
64. PlayingPiano
65. PlayingSitar
66. PlayingTabla
67. PlayingViolin
68. **PoleVault**
69. PommelHorse
70. PullUps
71. Punch
72. PushUps
73. Rafting
74. RockClimbingIndoor
75. RopeClimbing
76. Rowing
77. SalsaSpin
78. ShavingBeard
79. **Shotput**
80. SkateBoarding
81. Skiing
82. Skijet
83. SkyDiving
84. SoccerJuggling
85. **SoccerPenalty**
86. StillRings
87. SumoWrestling
88. Surfing
89. Swing
90. TableTennisShot
91. TaiChi
92. **TennisSwing**
93. **ThrowDiscus**
94. TrampolineJumping
95. Typing
96. UnevenBars
97. **VolleyballSpiking**
98. WalkingWithDog
99. WallPushups
100. WritingOnBoard
101. YoYo