

動画の行動認識用データセット

UCF101

UCF101: A Dataset of 101 Human Actions Classes
From Videos in The Wild

sasakiy84

1. UCF101 の行動分類と動画の構造
2. 動画の収集方法とカテゴリ別の規模
3. 行動認識の評価指標とモデル性能
4. UCF101 の利用動向と後続データ
5. 配布動画と評価分割の入手方法

1. UCF101 の行動分類と動画の構造

UCF101 の原著の公開ページ

 Search Submit Donate Log in

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 3 Dec 2012]

UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild

Khurram Soomro, Amir Roshan Zamir, Mubarak Shah

We introduce UCF101 which is currently the largest dataset of human actions. It consists of 101 action classes, over 13k clips and 27 hours of video data. The database consists of realistic user uploaded videos containing camera motion and cluttered background. Additionally, we provide baseline action recognition results on this new dataset using standard bag of words approach with overall performance of 44.5%. To the best of our knowledge, UCF101 is currently the most challenging dataset of actions due to its large number of classes, large number of clips and also unconstrained nature of such clips.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**

Report number: CRCV-TR-12-01

Cite as: [arXiv:1212.0402](https://arxiv.org/abs/1212.0402) [cs.CV]
(or [arXiv:1212.0402v1](https://arxiv.org/abs/1212.0402v1) [cs.CV] for this version)
<https://doi.org/10.48550/arXiv.1212.0402> 

Submission history

From: Khurram Soomro [[view email](mailto:khurram@cs.cmu.edu)]

[v1] Mon, 3 Dec 2012 14:45:31 UTC (2,861 KB)

Access Paper:
[View PDF](#)
[HTML \(experimental\)](#)
[TeX Source](#)
[view license](#)

Current browse context:
cs.CV
[< prev](#) | [next >](#)
[new](#) | [recent](#) | [2012-12](#)
Change to browse by:
[cs](#)

References & Citations
[NASAADS](#)
[Google Scholar](#)
[Semantic Scholar](#)

DBLP - CS Bibliography
[listing](#) | [bibtex](#)
[Khurram Soomro](#)
[Amir Roshan Zamir](#)
[Mubarak Shah](#)

Export BibTeX Citation

Bookmark


動画クリップを 101 種類の行動へ分類する

- 動画の行動認識には、クリップ全体の行動を分類する課題と、行動が起こる時間範囲を特定する課題がある。
- UCF101 が扱うのは前者である。入力は短い動画クリップ、正解はその動画が表す 1 つの行動カテゴリとなる。
- スポーツ・楽器演奏・日常動作を含む 13,320 クリップを、動画共有サービス YouTube から集め、2012 年に公開した。

101 種類の行動は 5 つの型に整理されている

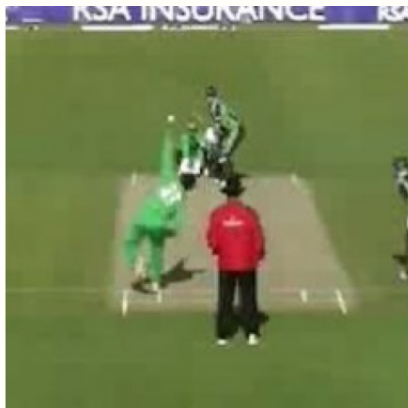
行動の型 [1]	カテゴリ数	クリップ数 [2]
人と物との相互作用	20	2619
身体の動き	16	1910
人と人との相互作用	5	690
楽器の演奏	10	1428
スポーツ	50	6673

[1] Soomro et al. (2012). CRCV-TR-12-01; arXiv:1212.0402. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Figure 2

<https://arxiv.org/abs/1212.0402>

[2] https://github.com/yxiong/temporal-segment-networks/tree/master/data/ucf101_splits

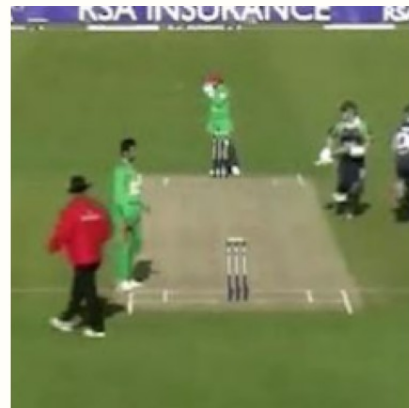
クリケットの打撃（CricketShot）の例



0.00 秒



1.56 秒



3.12 秒

行動の型から動画クリップまでの包含関係

単位 [1]	何を表すか	具体例
型	カテゴリの大区分。全 5 種類	スポーツ
カテゴリ	全 101 種類の行動ラベル	CricketShot
グループ	撮影場所や人物などが似た動画のまとめ。各カテゴリに 25 組	CricketShot の g04
クリップ	1 本の動画。各グループに 4 ~ 7 本	g04 の c01

[1] Soomro et al. (2012). CRCV-TR-12-01; arXiv:1212.0402. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. §2. <https://arxiv.org/abs/1212.0402>

目元の化粧（ApplyEyeMakeup）の例



0.04 秒

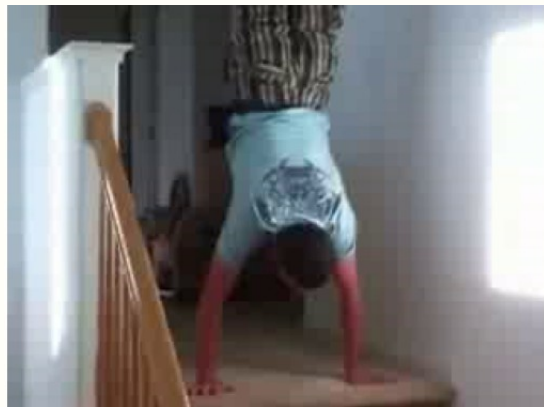


3.28 秒



6.52 秒

逆立ち歩き（HandstandWalking）の例



0.00 秒



2.16 秒



4.32 秒

2. 動画の収集方法とカテゴリ別の規模

先行データより行動と撮影条件を広げた

- KTH（2004）は、統制された撮影条件で 6 行動を扱う先行データである。HMDB51（2011）は、映画などの映像で 51 行動を扱った。
- また、先行する UCF50 は YouTube 映像の 50 行動を収録した。UCF101 は 101 行動へ、カテゴリ数を約 2 倍に広げた。
- UCF101 は、多種類の行動をカメラの動き・遮蔽・混雑した背景などの変動を含む映像で評価する。

自然な映像の多様な行動を評価する

- 撮影条件を統制した小規模なデータだけでは、日常の映像に見られる変動を十分に扱えなかった。
- そこで、YouTube から行動カテゴリに合う動画を集め、見た目や撮影条件の変化を含む認識用データを作った。
- カテゴリの選定、不要動画の除去、形式の統一、撮影条件が似たクリップのグループ化を行った。

収集した動画を選別して形式を統一する

- YouTube から動画をダウンロードし、収集する行動カテゴリに該当しない動画を人手で除去した。
- 映像は 320×240 画素、1 秒当たり 25 フレーム（25 fps）へ統一し、動画ファイル形式 AVI でカテゴリ別に配布した。
- 追加した 51 行動の動画では、音声も保持した。

撮影条件の似たクリップをグループにまとめる

- 各行動カテゴリには 25 グループがある [1]。
- 1 グループには、撮影場所や映る人物、カメラの視点などが似た 4 ～ 7 クリップをまとめる。
- ファイル名の g04 はグループ番号、c01 はグループ内のクリップ番号。
- 近い映像が学習とテストの両方に入ることを避けるため、グループ単位で分ける。

平均 7.21 秒の 13,320 クリップを収録する

- 101 カテゴリ、13,320 クリップを収録する [1]。
- 平均クリップ長は 7.21 秒、総時間は 1,600 分。映像は 25 fps、320×240 ピクセル。
- 動画をまとめた圧縮 ZIP は約 7 GB [3]。
- 25 グループ × 101 カテゴリで計 2,525 グループ [2]。
- 1 本の動画に 1 つの行動ラベルが付く。

[1] Soomro et al. (2012). CRCV-TR-12-01; arXiv:1212.0402. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Table 1
<https://arxiv.org/abs/1212.0402>

[2] グループ数は分割リストの集計 . https://github.com/yjxiong/temporal-segment-networks/tree/master/data/ucf101_splits

[3] ZIP は 6,960,400,115 bytes. https://github.com/tensorflow/datasets/blob/master/tensorflow_datasets/url_checksums/ucf101.txt

バイオリンの演奏（PlayingViolin）は 100 本 [2]



0.00 秒



3.72 秒



7.40 秒

[1] https://github.com/RichardMinsooGo-ML/5_TF2_UCF101_video_classification v_PlayingViolin_g01_c01.avi

[2] カテゴリ別集計 . https://github.com/yjxiong/temporal-segment-networks/tree/master/data/ucf101_splits

クリケットの打撃（CricketShot）は 167 本 [2]



0.00 秒



1.68 秒

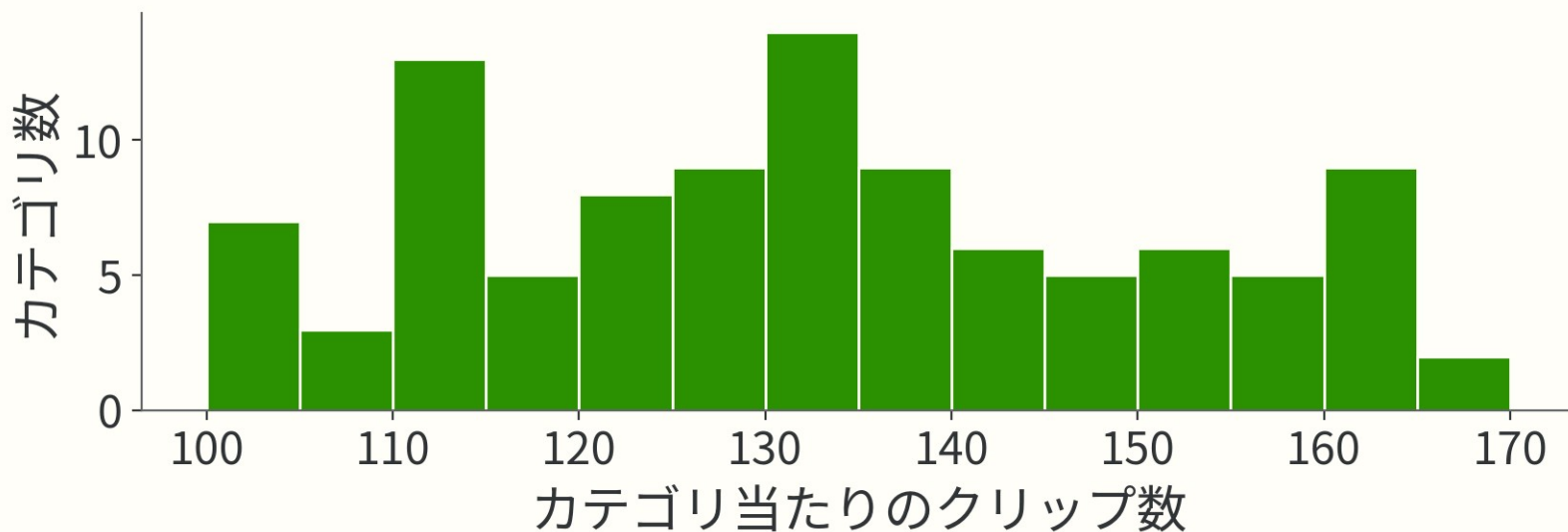


3.32 秒

[1] https://github.com/RichardMinsooGo-ML/5_TF2_UCF101_video_classification v_CricketShot_g01_c01.avi

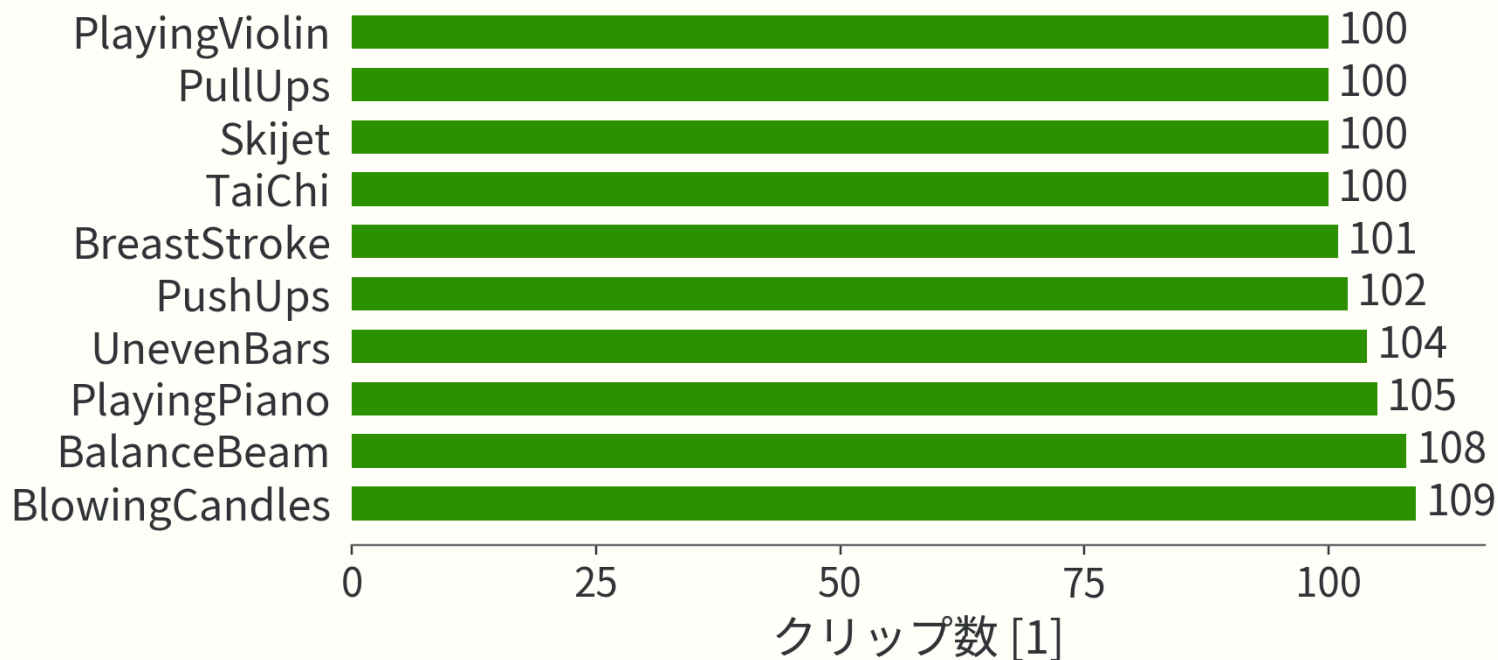
[2] カテゴリ別集計 . https://github.com/yjxiong/temporal-segment-networks/tree/master/data/ucf101_splits

カテゴリ当たりのクリップ数の中央値は 131 本



全 101 カテゴリ。最少は 100 本、最多は 167 本 [1]。

110 本未満のカテゴリは 10 種類



[1] 13,320 クリップをカテゴリ別に集計 . https://github.com/yjxiong/temporal-segment-networks/tree/master/data/ucf101_splits

3 分割ではテストにするグループが異なる

分割 [1]	テスト用グループ
split 1	g01 ~ g07
split 2	g08 ~ g14
split 3	g15 ~ g21

それ以外の 18 グループを学習用に使う。

[1] 全 101 カテゴリの分割リストを照合 . https://github.com/yjxiong/temporal-segment-networks/tree/master/data/ucf101_splits

3つの分割で学習用とテスト用を入れ替える

分割 [1]	学習用	テスト用
split 1	9,537	3,783
split 2	9,586	3,734
split 3	9,624	3,696

各分割の合計は 13,320 クリップ。

[1] 配布分割リストの集計 . https://github.com/yjxiong/temporal-segment-networks/tree/master/data/ucf101_splits

3. 行動認識の評価指標とモデル性能

動画単位の正解率を計算する例

正解カテゴリ	予測カテゴリ	得点
PlayingViolin	PlayingViolin	1
PlayingGuitar	PlayingGuitar	1
CricketShot	CricketBowling	0

3 動画中 2 動画が正解なので、正解率は $2 \div 3 \times 100 = 66.7\%$ となる。

動画の正解率を分割ごとに求めて平均する

- 各動画の最上位予測が正解カテゴリと一致すれば 1、不一致なら 0 とする。全動画の得点を平均し、百分率で正解率を報告する。
- 公式の 3 分割では、それぞれ別の学習用で学習し、対応するテスト用で評価する。
- 3 分割の平均は、分割ごとの正解率を平均した値である。

2012 年の原著の正解率は 44.5% だった

- 原著は局所的な動きと見た目を特徴量にして、映像を特徴の出現頻度で表した [1]。
- 分類にはサポートベクトルマシンを使った。
- 25 グループのうち 1 グループをテストにし、25 回入れ替える交差検証を行った。
- 平均正解率は 44.5% 。

I3D （ 2017 ） の入力別の正解率

I3D の入力 [1]	Top-1 正解率
RGB 画像	95.6%
フレーム間の動きを表す光学フロー	96.7%
RGB 画像と光学フロー	98.0%

400 行動の動画を扱う Kinetics-400 と ImageNet で事前学習し、UCF101 で微調整。公式 3 分割の平均。

[1] Carreira et al. (2017). CVPR. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Table 5 <https://arxiv.org/abs/1705.07750>

VideoMAE （ 2022 ） の事前学習データ別の結果

ViT-B の事前学習データ [1]	Top-1 正解率
UCF101 のみ	91.3%
Kinetics-400	96.1%

基本サイズの動画 Transformer （ ViT-B ） を UCF101 で微調整。隠した映像を復元して学ぶ VideoMAE の報告値。

2024 年の固定特徴による行動分類の正解率

固定する特徴モデル [1]	Top-1 正解率
DINOv2-g	91.2%
TVTSv2-H	91.8%
InternVideo2 第 2 段階・6B	97.3%

特徴を固定し、UCF101 の学習用で線形分類器を学習する。

[1] Wang et al. (2024). arXiv:2403.15377. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. Table 4 <https://arxiv.org/abs/2403.15377>

事前学習したモデルを転用して評価する

- 事前学習は、別の画像や映像から特徴の取り出し方を先に学ぶ工程である。微調整では、UCF101 の学習用でそのモデルを更新する。
- 線形分類では特徴モデルを固定し、行動を選ぶ分類器だけを UCF101 で学習する。
- ゼロショットでは UCF101 による学習を行わず、映像と行動名の対応などから予測する。

[1] Carreira et al. (2017). CVPR. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. <https://arxiv.org/abs/1705.07750>

[2] Wang et al. (2024). arXiv:2403.15377. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. <https://arxiv.org/abs/2403.15377>

I3D は見た目とフレーム間の動きを利用する

- I3D（2017）は、画像の畳み込みを時間方向へ広げた動画認識モデルである。
- RGB 画像は各フレームの色と見た目を表す。光学フローは、連続するフレームの間で画素がどの方向へ動いたかを表す。
- 大規模な 400 行動の動画データ Kinetics-400 と、画像分類用 ImageNet で事前学習し、UCF101 で微調整した結果が報告されている。

VideoMAE は隠した映像領域の復元で学習する

- VideoMAE（2022）は、動画の一部を隠し、見えている部分から復元する自己教師あり事前学習の手法である。
- ViT は、画像や映像を小さな領域に分け、その間の関係を Transformer で扱うモデルである。表の ViT-B は基本サイズを表す。
- 同じモデルでも、UCF101 だけで事前学習する場合と、大規模な Kinetics-400 で学ぶ場合とで、微調整後の正解率が変わる。

画像特徴の DINOv2 でも 91.2% に達する

- 背景や道具だけでも、行動カテゴリの手がかりになりうる。
- 2024 年の比較では画像特徴モデル DINOv2 を使う方式も 91.2% だった [1]。

4. UCF101 の利用動向と後続データ

Kinetics は行動数と動画数を拡大した

- Kinetics (2017) は、YouTube から多様な人物・撮影条件の行動クリップを集めた大規模データである。
- 最初の 400 行動版は 306,245 クリップを扱い、UCF101 の約 4 倍の行動数、約 23 倍のクリップ数となる。
- 大規模な映像から事前学習し、UCF101 へ微調整する利用を支えた。

手と物の動きを扱う Something-Something

- Something-Something （2017）は、物を押す・持ち上げるなど、人と物の相互作用を 174 種類に分けた動画データである。
- UCF101 のスポーツや日常動作に対し、手と物の関係や移動方向を細かく区別する。
- 物体や背景が似ていても動きの違いで答えが変わるため、時間方向の変化を捉える評価へ対象を広げる。

UCF101 の転移学習・特徴評価への利用

- 2012 年の公開後、行動分類から事前学習モデルの転移評価へ用途が広がった [1]。
- InternVideo2 （2024）は、固定特徴の線形分類とゼロショットの評価に用いた [1]。
- 2026-10-15 の確認時点でも、MMAction2 は取得・前処理・評価手順を公開している [2]。

[1] Carreira et al. (2017). CVPR. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. <https://arxiv.org/abs/1705.07750>

Wang et al. (2024). arXiv:2403.15377. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. <https://arxiv.org/abs/2403.15377>

[2] MMAction2 contributors (n.d.). MMAction2 documentation. 参照 2026-10-15. Preparing UCF-101.
<https://github.com/open-mmlab/mmaction2/tree/main/tools/data/ucf101>

5. 配布動画と評価分割の入手方法

動画本体と 3 分割のリストを取得する

- 公式配布元から、動画のアーカイブと学習・テスト分割の ZIP を取得する [1]。
- 展開したカテゴリ別フォルダと分割リストを対応させる。
- データを扱う MMAction2 などの手順で、フレーム抽出と入力用リストを準備できる [2]。

[1] MMAction2 contributors (n.d.). MMAction2 documentation. 参照 2026-10-15. Preparing UCF-101.
<https://github.com/open-mmlab/mmdetection/tree/main/tools/data/ucf101>

カテゴリ別の動画を分割リストと対応付ける

- 配布動画は行動カテゴリごとのフォルダに置かれ、ファイル名にはカテゴリ・グループ番号・クリップ番号が入る。
- 学習用とテスト用のリストは、そのフォルダ名とファイル名で動画を指定する。同じカテゴリ内でも、グループ番号によって役割が変わる。
- 動画認識用ライブラリ MMAction2 の手順では、動画からフレームを抽出し、分割リストに沿って学習・評価用ファイルを準備する。

UCF101 は動画モデルの行動認識と転移を評価する

- 101 種類の行動を扱う 13,320 クリップのデータセット。
- 自然な YouTube 映像で、多様な行動を評価するため 2012 年に作られた。
- 撮影場所・人物・カメラの視点が似た動画のグループを単位に分割する。
- 登場時の分類評価から、事前学習モデルの転移・特徴評価にも用途が広がった。

Appendix

1. 開発者の研究と発表先
2. 動画と行動ラベルの追加例
3. 5つの型に属する全 101 行動カテゴリ
4. 行動の型・グループごとの動画数
5. 行動の型別・ゼロショットの評価結果
6. 動画の取得と本数の集計コード
7. 動画の行動認識に関連する文献

UCF101 の開発者と 2012 年の技術報告

- CRCV-TR-12-01 としてまとめ、2012-12-03 に arXiv へ投稿した [1]。
- 第一著者の Khurram Soomro は、UCF101 を構築した行動認識研究者。
- 最終著者の Mubarak Shah は、動画の行動認識を研究し、UCF50 や UCF Sports にも取り組んだ。
- 当時の著者の所属は、セントラルフロリダ大学のコンピュータビジョン研究センター。

[1] Soomro et al. (2012). CRCV-TR-12-01; arXiv:1212.0402. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. <https://arxiv.org/abs/1212.0402>

編み物（Knitting）の例



0.00 秒



5.32 秒



10.60 秒

ヨ－ヨ－（ YoYo ） の例



0.00 秒

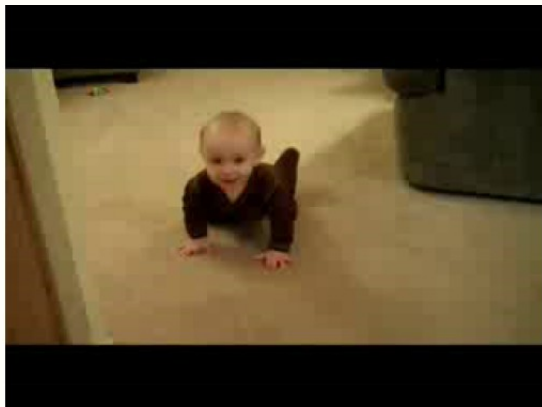


3.72 秒



7.44 秒

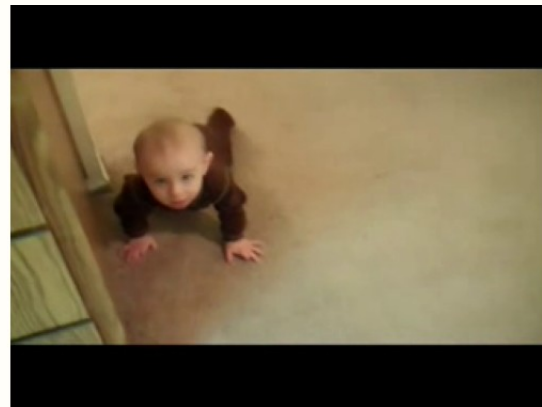
赤ちゃんのはいはい（ BabyCrawling ） の例



0.00 秒

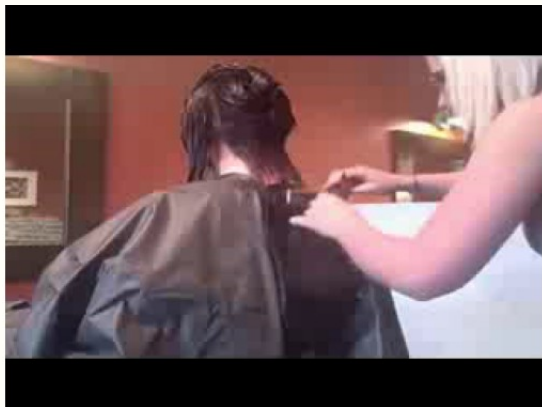


3.28 秒

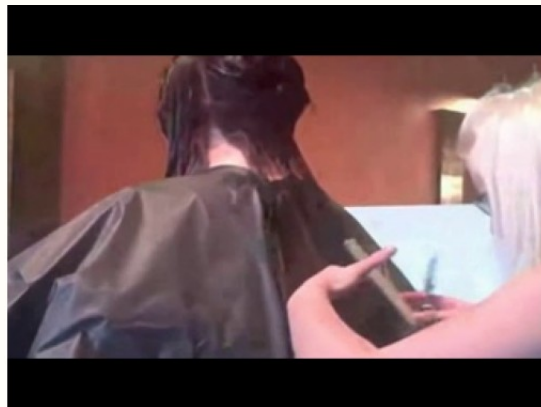


6.56 秒

散髪（Haircut）の例



0.00 秒



2.68 秒



5.36 秒

ギターの演奏（PlayingGuitar）の例



0.00 秒



5.00 秒



9.96 秒

ビリヤード（ Billiards ）の例



0.00 秒



5.48 秒



10.92 秒

バスケットボールのダンク (BasketballDunk)



0.00 秒



1.20 秒



2.40 秒

サーフィン（Surfing）の例



0.00 秒



5.08 秒



10.12 秒

タイピング（ Typing ） の例



0.00 秒



2.52 秒



5.04 秒

スケートボード（SkateBoarding）の例



0.00 秒



4.04 秒



8.04 秒

人と物との相互作用という型の全 20 カテゴリ

ApplyEyeMakeup
ApplyLipstick
BlowDryHair
BrushingTeeth
CuttingInKitchen
Hammering
HulaHoop
JugglingBalls
JumpRope
Knitting

Mixing
MoppingFloor
Nunchucks
PizzaTossing
ShavingBeard
SkateBoarding
SoccerJuggling
Typing
WritingOnBoard
YoYo

身体の動きという型の全 16 カテゴリ

BabyCrawling

BlowingCandles

BodyWeightSquats

HandstandPushups

HandstandWalking

JumpingJack

Lunges

PullUps

PushUps

RockClimbingIndoor

RopeClimbing

Swing

TaiChi

TrampolineJumping

WalkingWithDog

WallPushups

人と人との相互作用という型の全 5 カテゴリ

BandMarching

Haircut

HeadMassage

MilitaryParade

SalsaSpin

楽器の演奏という型の全 10 カテゴリ

Drumming

PlayingCello

PlayingDaf

PlayingDhol

PlayingFlute

PlayingGuitar

PlayingPiano

PlayingSitar

PlayingTabla

PlayingViolin

スポーツという型の全 50 カテゴリ 1/3

Archery

BalanceBeam

BaseballPitch

Basketball

BasketballDunk

BenchPress

Biking

Billiards

Bowling

BoxingPunchingBag

BoxingSpeedBag

BreastStroke

CleanAndJerk

CliffDiving

CricketBowling

CricketShot

Diving

Fencing

スポーツという型の全 50 カテゴリ 2/3

FieldHockeyPenalty

FloorGymnastics

FrisbeeCatch

FrontCrawl

GolfSwing

HammerThrow

HighJump

HorseRace

HorseRiding

IceDancing

JavelinThrow

Kayaking

LongJump

ParallelBars

PoleVault

PommelHorse

スポーツという型の全 50 カテゴリ 3/3

Punch

Rafting

Rowing

Shotput

Skiing

Skijet

SkyDiving

SoccerPenalty

StillRings

SumoWrestling

Surfing

TableTennisShot

TennisSwing

ThrowDiscus

UnevenBars

VolleyballSpiking

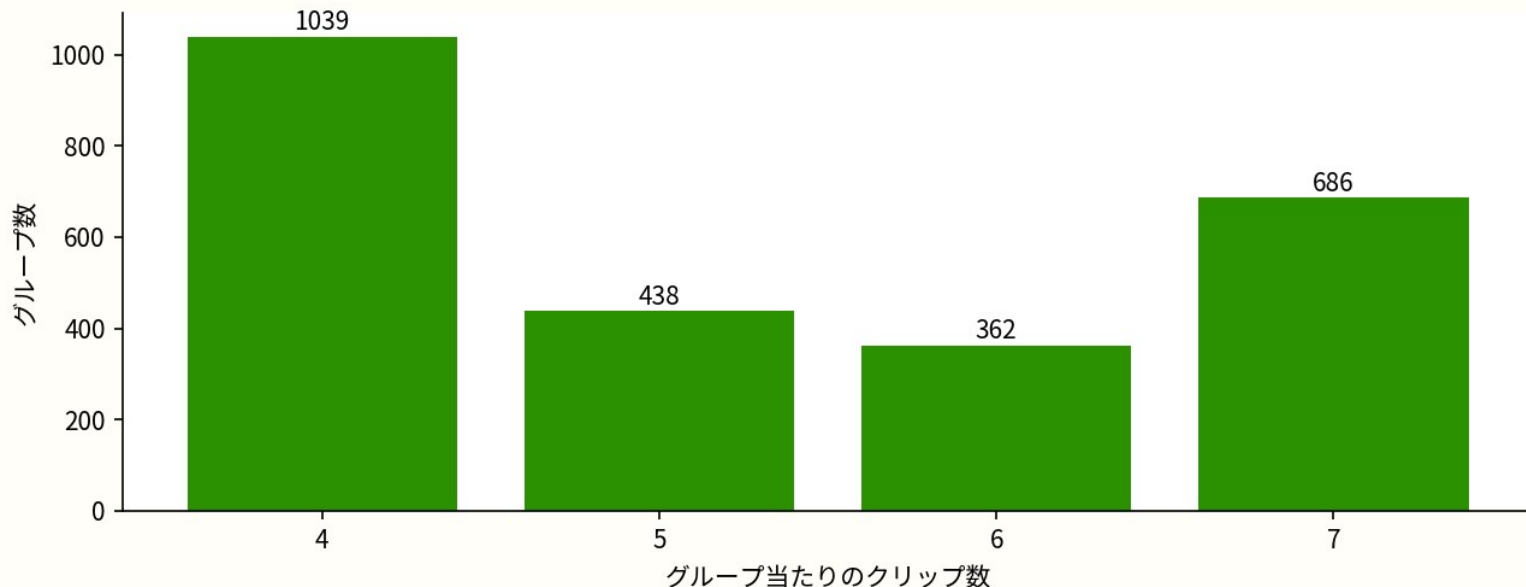
型ごとにカテゴリ数とクリップ数を比較する

型 [1]	カテゴリ数	合計本数	最少	平均	最多
人と物との相互作用	20	2,619	110	131.0	161
身体の動き	16	1,910	100	119.4	144
人と人との相互作用	5	690	125	138.0	155
楽器の演奏	10	1,428	100	142.8	164
スポーツ	50	6,673	100	133.5	167
全 101 カテゴリ	101	13,320	100	131.9	167

カテゴリ当たりの本数は中央値 131 本、母標準偏差 18.2 本。25% 点は 117 本、75% 点は 145 本、95% 点は 163 本。

[1] Soomro et al. (2012). CRCV-TR-12-01; arXiv:1212.0402. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Figure 2
<https://arxiv.org/abs/1212.0402> https://github.com/yxiong/temporal-segment-networks/tree/master/data/ucf101_splits

4 クリップのグループが最も多い



原著における行動の型別の正解率

行動の型 [1]	正解率
人と物との相互作用	38.52%
身体の動き	36.26%
人と人との相互作用	44.14%
楽器の演奏	37.42%
スポーツ	50.54%

[1] Soomro et al. (2012). CRCV-TR-12-01; arXiv:1212.0402. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. §3. <https://arxiv.org/abs/1212.0402>

2024 年の報告におけるゼロショット評価

手法 [1]	Top-1 正解率
CLIP	68.9%
TVTSv2	78.0%
InternVideo2clip-1B	88.8%
InternVideo2clip-6B	89.5%

UCF101 で微調整せず、行動名と映像を照合する。

[1] Wang et al. (2024). arXiv:2403.15377. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. Table 5 <https://arxiv.org/abs/2403.15377>

配布リストからカテゴリ別の本数を集計する

```
from pathlib import Path
from collections import Counter

root = Path("ucfTrainTestlist")
train = (root / "trainlist01.txt").read_text().splitlines()
test = (root / "testlist01.txt").read_text().splitlines()
clips = {line.split()[0] for line in train + test}
counts = Counter(path.split("/")[0] for path in clips)
print(len(clips), len(counts))
for label, n in sorted(counts.items()):
    print(label, n)
```

[1] https://github.com/yxiong/temporal-segment-networks/tree/master/data/ucf101_splits

動画と分割リストのアーカイブを取得する

```
base=https://www.crcv.ucf.edu/data/UCF101
curl -L --fail -O "$base/UCF101.rar"
curl -L --fail -O \
    "$base/UCF101TrainTestSplits-RecognitionTask.zip"

unrar x UCF101.rar
unzip UCF101TrainTestSplits-RecognitionTask.zip
```

UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild

Soomro, K., Zamir, A. R., & Shah, M. (2012). CRCV-TR-12-01; arXiv:1212.0402.

UCF101 の原著論文。101 行動の収集、グループの設計、初期の行動認識評価を報告する。

<https://arxiv.org/abs/1212.0402>

Two-Stream Convolutional Networks for Action Recognition in Videos

Simonyan, K., & Zisserman, A. (2014). Advances in Neural Information Processing Systems, 27.

Two-Stream モデルの原著論文。RGB 画像とオプティカルフローを別々のネットワークで扱う。

<https://arxiv.org/abs/1406.2199>

Temporal Segment Networks: Towards Good Practices for Deep Action Recognition

Wang, L., et al. (2016). ECCV.

TSN の原著論文。動画の複数区間からフレームを抽出し、長い時間範囲の情報を統合する。

<https://arxiv.org/abs/1608.00859>

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

Carreira, J., & Zisserman, A. (2017). CVPR.

I3D の原著論文。画像の畳み込みを 3 次元へ拡張し、Kinetics からの転移を評価する。

<https://arxiv.org/abs/1705.07750>

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Advances in Neural Information Processing Systems, 35.

VideoMAE の原著論文。高いマスク率の映像復元により、自己教師あり事前学習を行う。

<https://arxiv.org/abs/2203.12602>

Preparing UCF-101

MMAction2 contributors (n.d.). MMAction2 documentation. 参照 2026-10-15.

MMAction2 の公式データ準備手順。UCF101 の取得、フレーム抽出、分割リスト作成を記載する。

<https://github.com/open-mmlab/mmaaction2/tree/main/tools/data/ucf101>

The Kinetics Human Action Video Dataset

Kay, W., et al. (2017). arXiv:1705.06950.

Kinetics の原著論文。YouTube の大規模な行動クリップの収集と品質管理を報告する。

<https://arxiv.org/abs/1705.06950>

The “something something” video database for learning and evaluating visual common sense

Goyal, R., et al. (2017). ICCV.

Something-Something の原著論文。人と物の相互作用を、時間的な関係を含む動作ラベルで扱う。

<https://arxiv.org/abs/1706.04261>

Training and Test

MMAction2 contributors (n.d.). MMAction2 documentation. 参照 2026-10-15.

MMAction2 の公式実行手順。設定ファイル、チェックポイントを用いた学習・テストを記載する。

https://github.com/open-mmlab/mmdetection/blob/main/docs/en/user_guides/train_test.md