技術マップ：動画認識・行動認識

原 健翔（産総研）

タスク・データセット

モデル・学習手法

動画・テキスト

検索	質問応答	キャプション生成	動画テキストペア	一人称視点	一人称x三人称	手順理解
MSRVTT Xu+, 2016	MSRVTT-QA Xu+, 2017	ActivityNet Captions Krishna+, 2017	HowTo100M Miech+, 2019	Ego4D Grauman+, 2021	Ego-Exo4D Grauman+, 2023	Ego4D Goal-Step Song+, 2023

文献リスト

検出タスク

日常動作	時空間検出	時空間検出	動物動画	関係性検出
THUMOS Zamir+, 2014	ActivityNet Heilbron+, 2015	Charades Sigurdsson+, 2016	AVA Gu+, 2017	HACS Zhao+, 2017	AVA-Kinetics Li+, 2020	MammalNet Chen+, 2023	SportsHHI Wu+, 2024

長時間動画理解

EgoSchema Mangalam+, 2023

識別タスク

KTH Schuldt+, 2004

HMDB51 Kuehne+, 2011

Sports-1M Karpathy+, 2014

Kinetics-{400, 600, 700} [Kay+, 2017], [Carreira+, 2018, 2019]

MiT Monfort+, 2018

Action Genome Ji+, 2019 シーングラフ

多様なドメインのYouTube動画のデータセットが主流

Weizmann Blank+, 2005

人・物体インタラクション認識

詳細行動認識

Hollywood Laptev+, 2008

UCF101 Soomro+, 2012

YouTube-8M Haija+, 2016

Something-Something Goyal+, 2017

Diving48 Li+, 2018

FineGym Shao+, 2020

動物動画

Animal Kingdom Ng+, 2022

マルチモーダル動画理解

MVBench Li+, 2023

LLMを活用した動画テキストデータ構築

InternVid Wang+, 2023

Step Differences Nagarajan+, 2024 手順の差異理解

Hand-Crafted

STIP Laptev+, 2003

Dense Traj. Wang+, 2011

3DHOG Klaeser+, 2008

iDT Wang+, 2013

Sparse Sampling

Dense Sampling

局所特徴を抽出し動画全体で符号化

CNN

2D CNN

Two-Stream Simonyan+, 2014

TDD Wang+, 2015

TS Fusion Feichtenhofer+, 2016

CNN+LSTM Ng+, 2015

TSN Wang+, 2016

RGB+Optical FlowのTwo-Streamから始まりTwo-Stream構造の検討やLSTMを用いた時間情報のモデル化など動画の表現獲得に向けて研究.

3D CNN

Non-local Wang+, 2017

IG-Kinetics Ghadiyaram+, 2019

C3D Tran+, 2014

LTC Varol+, 2016

I3D Carreira+, 2017

3D ResNet Hara+, 2017

SlowFast Feichtenhofer+, 2018

CorrNet Wangi+, 2020

P3D Carreira+, 2017

R(2+1)D Tran+, 2017

S3D Xie+, 2017

CSN Tran+, 2019

X3D Feichtenhofer+, 2020

基礎モデルの検討から時間情報の表現や効率的な学習・認識に向けて研究が進展.

Transformer

ViTの応用

TimeSformer Bertasius+, 2021

ViViT Arnab+, 2021

Video Swin Transformer Liu+, 2021

Masked ModelingによるSSL

VideoMAE Tong+, 2022

VideoMAE V2 Wang+, 2023

マルチモーダル学習

Omnivore Girdhar+, 2022

TubeViT Piergiovanni,+, 2022

画像での事前学習モデルの活用

UniformerV2 Li+, 2022

FROSTER Huang+, 2024

動画・言語モデル，動画基盤モデル，LLMの活用

VideoCLIP Xu+, 2021

LaViLa Zhao+, 2022

InternVideo2 Wang+, 2024

VideoPrism Zhao+, 2024

2000 2012 2016 2020 2024