Motivation

Most Human Action Understanding models **do not understand the action -> fragile** and **unable to adapt** to new settings.

To **gain more in-depth knowledge about human actions** -> new action understanding task: which actions are likely to occur in the same time interval.

Most **human actions are interconnected**, as an action that ends is usually followed by the start of a related action, not a random one.

Interconnection of human actions is **very well depicted in lifestyle vlogs**, vloggers record their **everyday routine**.

Dataset

Action Co-occurrence Task

Are the actions in the videos co-occuring within 10 seconds?

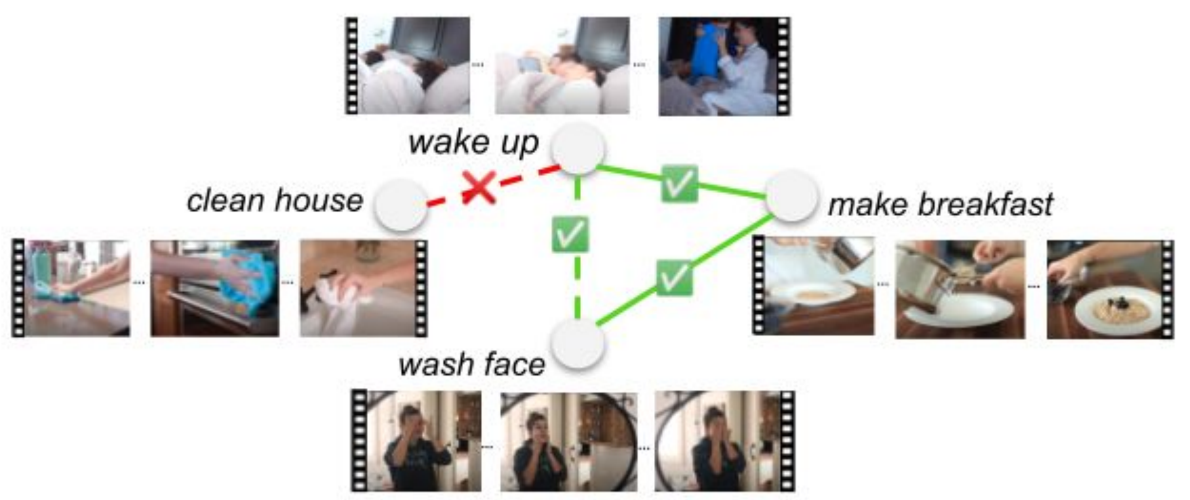


Fig. 1: A natural way to model the connections between human actions is through a **graph representation**, where actions are as **nodes**, and their co-occurrences as **edges**.

Data Pre-processing steps

Step1: Action Co-occurrence

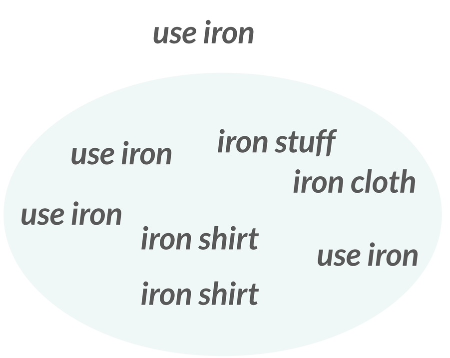
Transcript ⇒ Co-occurring Action pairs

00:50 morning I **wake up** — (wake up, wash face)
00:53 after that I **wash** my **face** — (wake up, make breakfast)
00:59 I then **make breakfast** — (wash face, make breakfast)

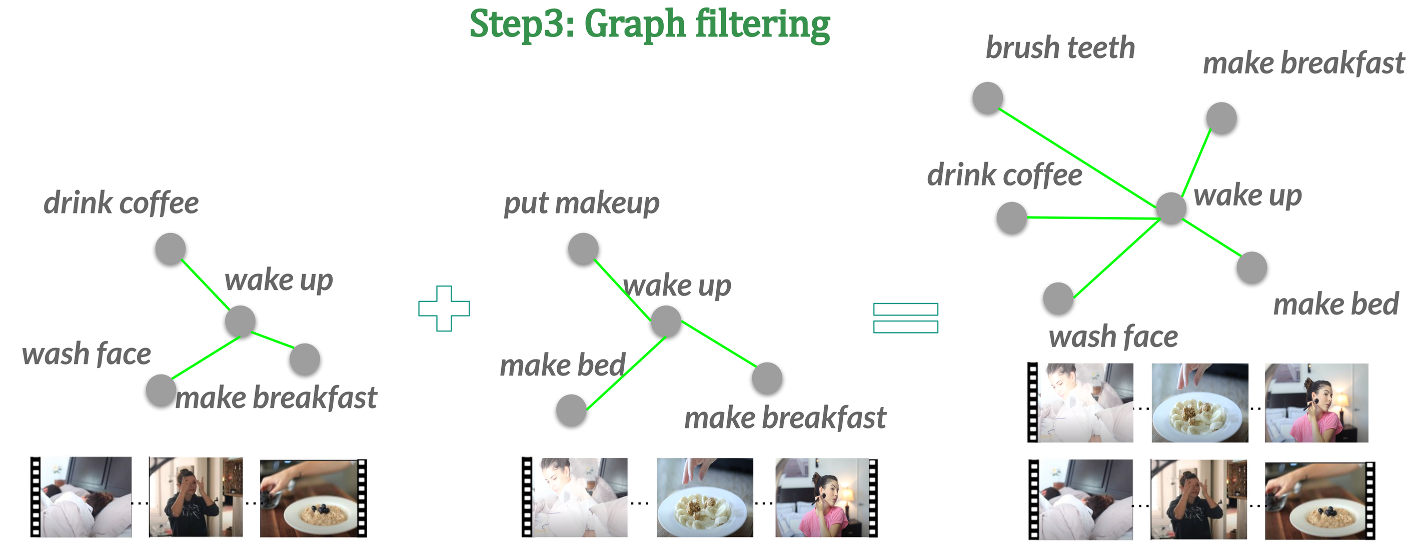
Two actions co-occur if they occur in the same interval of time (**10 seconds**) in a video.

- **10 sec** is **intermediate value threshold**, allows for **short** (e.g. "open fridge") & **long** (e.g. "cook meal") **actions** to co-occur.

Step2: Action Clustering



Step3: Graph filtering



Data Statistics

	#Verbs	#Actions	#Action pairs
Initial	608	20,718	-
Co-occurrence	439	18,939	80,776
Clustering	172	2,513	48,934
Graph	164	2,262	11,711

Video-clips	19,685
Verbs	164
Actions	2,262
Action pairs	11,711
(Action, Video-clip) pairs	12,994

Tab. 1& 2: Data statistics, at each stage of data pre-processing.

TextGraphs-17
August 15, 2024
at ACL 2024

UNIVERSITY OF MICHIGAN

Human Action Co-occurrence in Lifestyle Vlogs using Graph Link Prediction

Oana Ignat, Santiago Castro, Weiji Li, Rada Mihalcea

oignat@umich.edu

Our Contributions:

1. Novel **task**: **Human Action Co-occurrence Identification**.
2. ACE (Action Co-occurrencE) **Dataset**: a graph of ~12k co-occurring pairs of actions & video clips.
3. Graph link prediction **Models**: use visual & textual information to infer if two actions are co-occurring.
   a. Graphs are particularly well suited to **capture relations between human actions.**
   b. Graph representations capture novel and relevant information **across different data domains.**

Data and Code:
github.com/

Models & Evaluation

Heuristics-based Graph Topology (e.g. Common Neighbours):

score_A1A2 = node_A1 & node_A2 *#common-neighbours*

score_A1A2 > thresh. -> A1 & A2 co-occur

Embedding-based:

cosine-similarity (emb_A1, emb_A2) > thresh. -> A1 & A2 co-occur

Learning-based (SVM):

Input: emb_A1 + emb_A2 + score_A1A2

Data Representation

Action Embeddings: Sentence-BERT
Transcript Embeddings: Sentence-BERT
} Text

Action Embeddings: CLIP (Text Transformer)
Sequence-level: CLIP (Vision Transformer ViT-B/16)
} Text + Visual

Action Embeddings: Average of neighbour node/ action embed. (Sentence-BERT or CLIP)
} Graph

Model	Accuracy
BASELINE	
Random	50.0
HEURISTIC-BASED	
Common Neighbours	82.9
Salton Index	71.2
Hub Promoted Index	78.3
Hub Depressed Index	61.5
Adamic-Adar Index	82.9
Resource Allocation	67.3
Shortest Path	82.9
EMBEDDING-BASED	
Cosine similarity	82.8
attri2vec	65.7
GCN	77.2
GraphSAGE	78.1
LEARNING-BASED	
SVM	**91.1**

Tab. 3: Results on test data.

Downstream Task: Similar Action Retrieval

- **Novelty vs. Relevance in Action Retrieval.**
- **Diversity in Action Representations.**
- **Location in Action Representations.**

k	INPUT REPRESENTATIONS	
	Textual	Graph
	DIVERSITY/ OVERLAP SCORE ↓	
3	0.35	**0.12**
5	0.31	**0.11**
10	0.26	**0.10**
Dataset	LOCATION / RECALL SCORE ↑	
Breakfast	0.16	**0.22**
COIN	0.23	**0.60**
EPIC-KITCHENS	0.14	**0.26**

Tab. 4: : Scores measuring the difference of information, diversity, and location, between the action kNNs using different types of embeddings: textual and graph-based

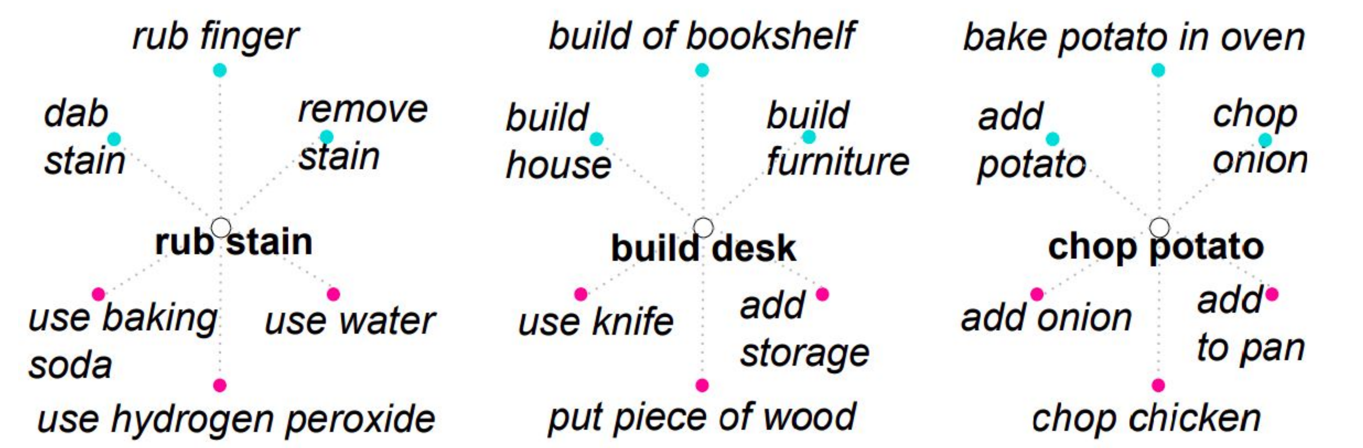


Fig. 2: Top 3 action neighbors, obtained from textual and graph-based representations, for 3 random action queries from our dataset: "rub stain", "build desk", "chop potato".