Inferring Tasks and Fluents in Videos by Learning Causal Relations

Haowen Tang, Ping Wei, Huan Li, and Nanning Zheng

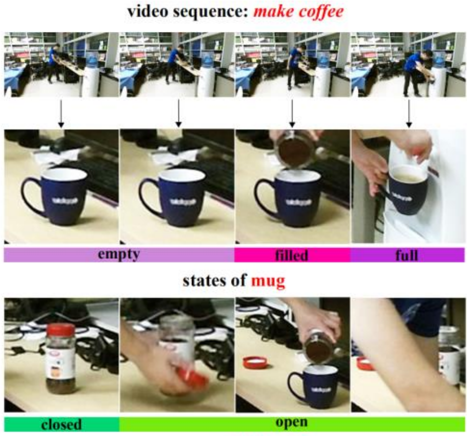
Xi'an Jiaotong University, Xi'an, China

Overview

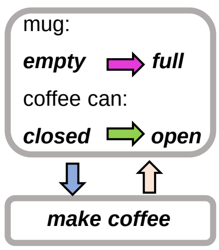
Recognizing time-varying object states in complex tasks is a challenging issue. In our work, we propose a novel model to jointly infer object fluents and complex tasks in videos, in our model:

- A task is a complex human activity with specific goals.
- A fluent is defined as a time-varying object state.
- A hierarchical graph represents a task as a human action stream and multiple concurrent object fluents which vary as the human performs the actions.

In this process, the human actions serve as the causes of object state changes which conversely reflect the effects of human actions.

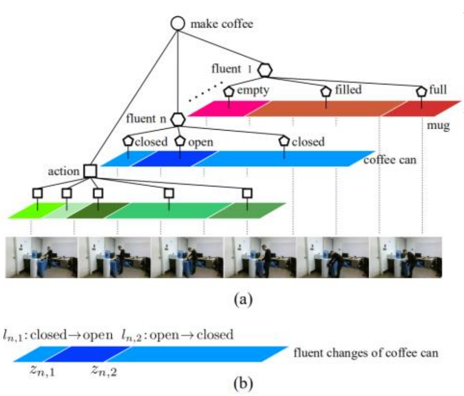


We can infer the fluents of mug according to task and infer task according to fluents of mug and coffee can:



For a given input video, a causal sampling search algorithm is proposed to jointly infer the task category and the states of objects. For model learning, a structural SVM framework is adopted to jointly train the task, fluent, cause, and effect parameters. We test the proposed method on a task and fluent dataset. Experimental results demonstrate the effectiveness of the proposed method.

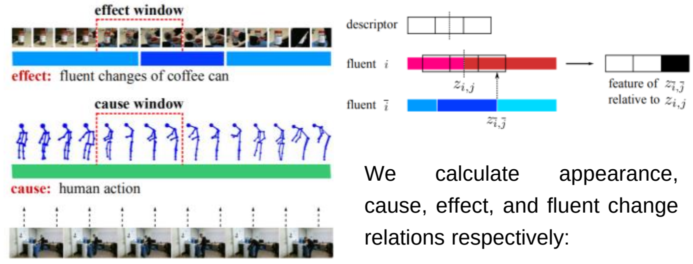
Model



We propose a hierarchical graph model to describe fluents and tasks. On this graph, the task category is the root node, which is decomposed into a human action process and several object fluent processes. Each fluent process is composed of several continuous sequential object states in temporal domain and the action includes several sub-action processes. The score of labelling video I with fluent states f and task y is defined as:

$$S(y,\mathbf{f},\mathbf{I}) = \underbrace{\sum_{i=1}^{n_y}\sum_{t=1}^{\tau}\boldsymbol{\omega}^{\mathrm{T}}_{y,f_{i,t}}\psi(i,I_t)}_{\text{fluent appearance}} + \underbrace{\sum_{i=1}^{n_y}\sum_{j=1}^{m_i}\boldsymbol{\alpha}^{\mathrm{T}}_{y,l_{i,j}}\phi(\mathbf{I},z_{i,j})}_{\text{cause}} + \underbrace{\sum_{i=1}^{n_y}\sum_{j=1}^{m_i}\boldsymbol{\beta}^{\mathrm{T}}_{y,l_{i,j}}\varphi(\mathbf{I},z_{i,j})}_{\text{effect}} + \underbrace{\sum_{i,j}^{n_y,m_i}\sum_{\bar{i},\bar{j}}^{n_y,m_{\bar{i}}}\boldsymbol{\gamma}^{\mathrm{T}}_{y,l_{i,j},l_{\bar{i},\bar{j}}}\boldsymbol{\lambda}(z_{i,j},z_{\bar{i},\bar{j}})}_{\text{fluent relation}}$$

This equation combines the appearance, cause, effect, and fluent change relations to measure the compatibility of the task and fluent state. It provides a unified framework to jointly represent, learn, and infer the task and fluents in videos.



We calculate appearance, cause, effect, and fluent change relations respectively:

- **Fluent appearance:** VGG-16 network ➡ fluent state classifier;
- **Cause:** SVM ➡ fluent change classifier;
- **Effect:** an effect classifier with histogram;
- **Fluent change relation:** a temporal descriptor ➡ represent fluent

Total loss:

$$\arg\min_{\mathbf{w},\xi_n\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{n=1}^{N}\xi_n \qquad \text{s.t. } \forall n, \forall y, \forall \mathbf{f},$$

$$S_{\mathbf{w}}(y^n,\mathbf{f}^n,\mathbf{I}^n) - S_{\mathbf{w}}(\mathbf{I},y,\mathbf{f}) \geq \Delta(y,y^n,\mathbf{f},\mathbf{f}^n) - \xi_n$$

$\Delta(y,y^n,\mathbf{f},\mathbf{f}^n)$ measures the joint loss between the hypothesized task-fluent labels and the ground-truth ones:

$$\Delta(y,y^n,\mathbf{f},\mathbf{f}^n) = \Delta_s(y,y^n) + \Delta_f(\mathbf{f},\mathbf{f}^n)$$

Experiments

Overall task recognition accuracy

Methods	Accuracy
Frame CNN	0.39
LSTM	0.31
Two-Stream CNN	0.54
4DHOI	0.62
ALE	0.67
Our Method	**0.72**

Overall accuracy of 50-class fluent states

Methods	Accuracy
SFCNN	0.25
Our Method	**0.37**

Ablation analysis of different model terms

Methods	Task Acc	Fluent Acc
App	0.609	0.290
App + Csl	0.614	0.294
App + Csl + Rel	**0.72**	**0.37**

Visualization