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

Problem

- **Definition:**

  task: a complex human activity with specific goals;

  fluent: a time-varying object state;

- **Objective:**

  Jointly infer **object fluents** and **complex tasks** in videos;

- **Method:**

  A causal sampling search algorithm.

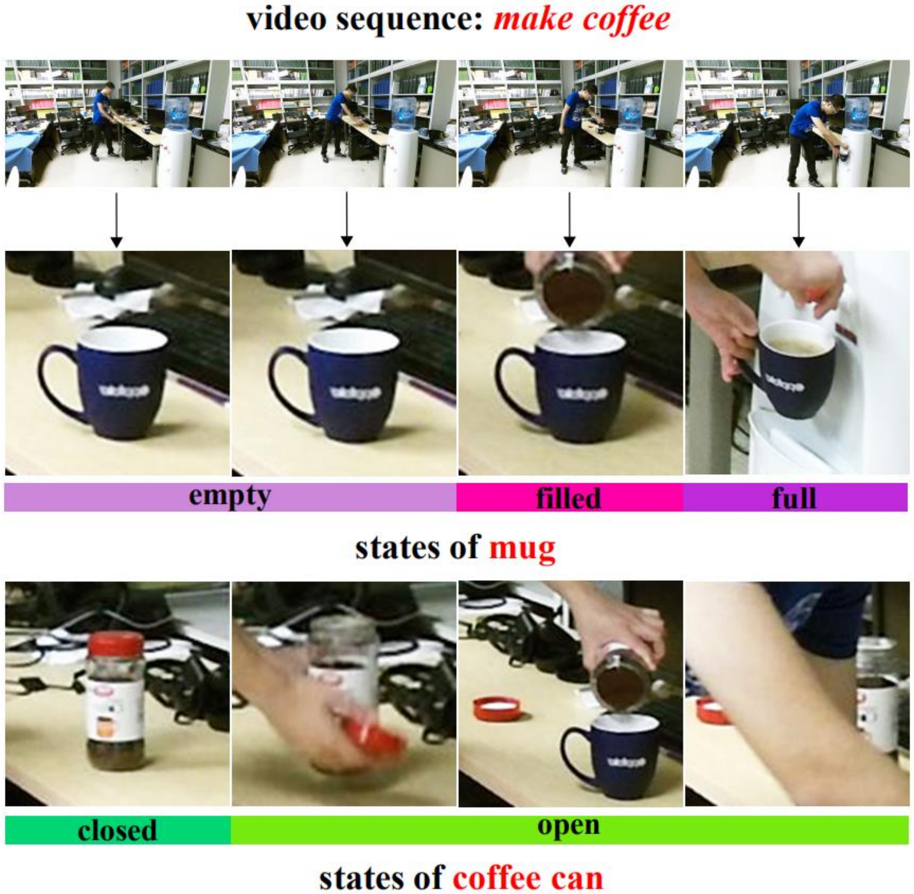

Fig. 1. Tasks and fluents in videos.
The bars represent the fluent states.

Model

- The **score** of labelling video ***I*** with fluent states ***f*** and task *y*:

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

- Inferring the fluent states ***f*** and task *y* by:

$$(y^*, \mathbf{f}^*) = \arg\max_{y,\mathbf{f}} \ S(y,\mathbf{f},\mathbf{I})$$

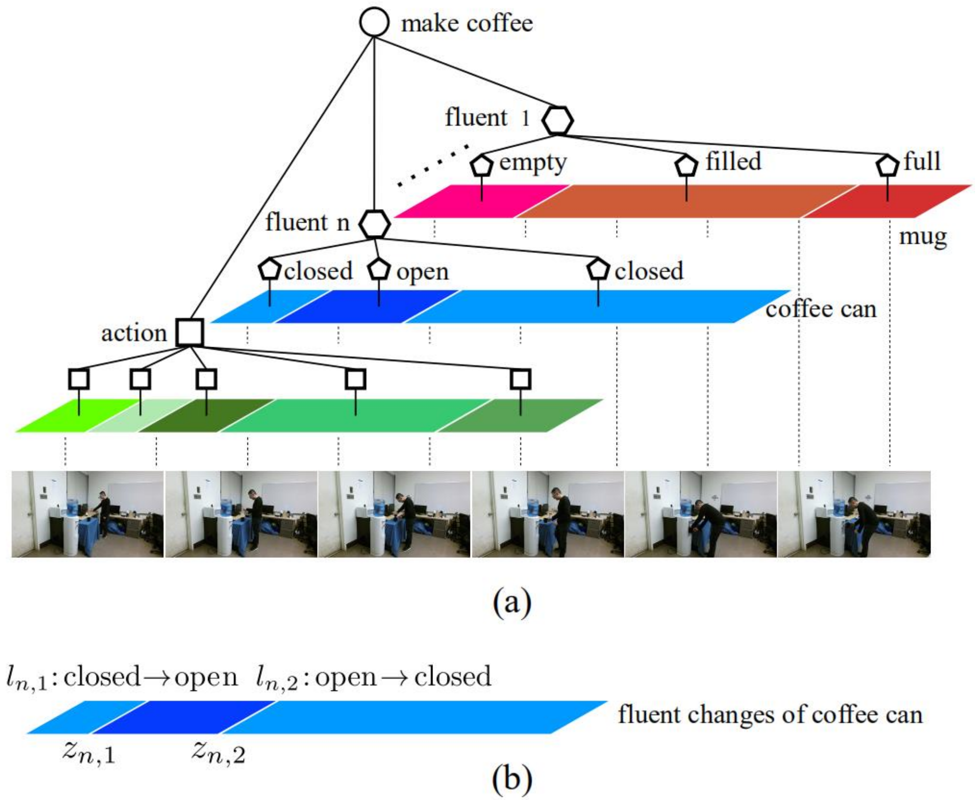

Fig. 2. Hierarchical models of tasks and fluents.

Model

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 change relations.

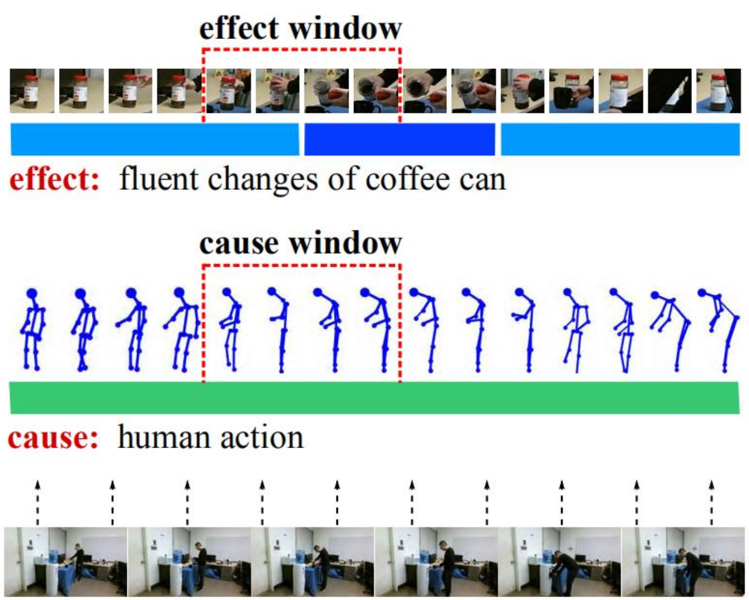

Fig. 3. Cause and effect windows in a task.

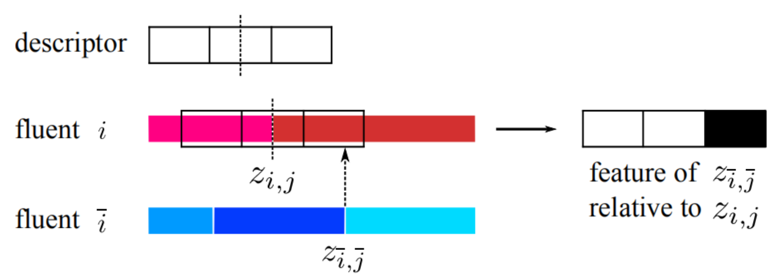

Fig. 4. Fluent change relation descriptor.

Loss Function

- We learn the model parameters with **structural SVM** method:

$$\arg\min_{\mathbf{w},\xi_n\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \ \forall n, \ \forall y, \ \forall \mathbf{f},$$

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

where ξ_n is a slack variable and *C* is a positive constant which balances the training error and margin maximization.

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

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

Results & Ablation

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

Table. I. Overall task recognition accuracy.

Methods	Accuracy
SFCNN	0.25
Our Method	**0.37**

Table. II. Overall accuracy of 50-class fluent states.

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

Table. III. Ablation analysis of different model terms.

Visualization

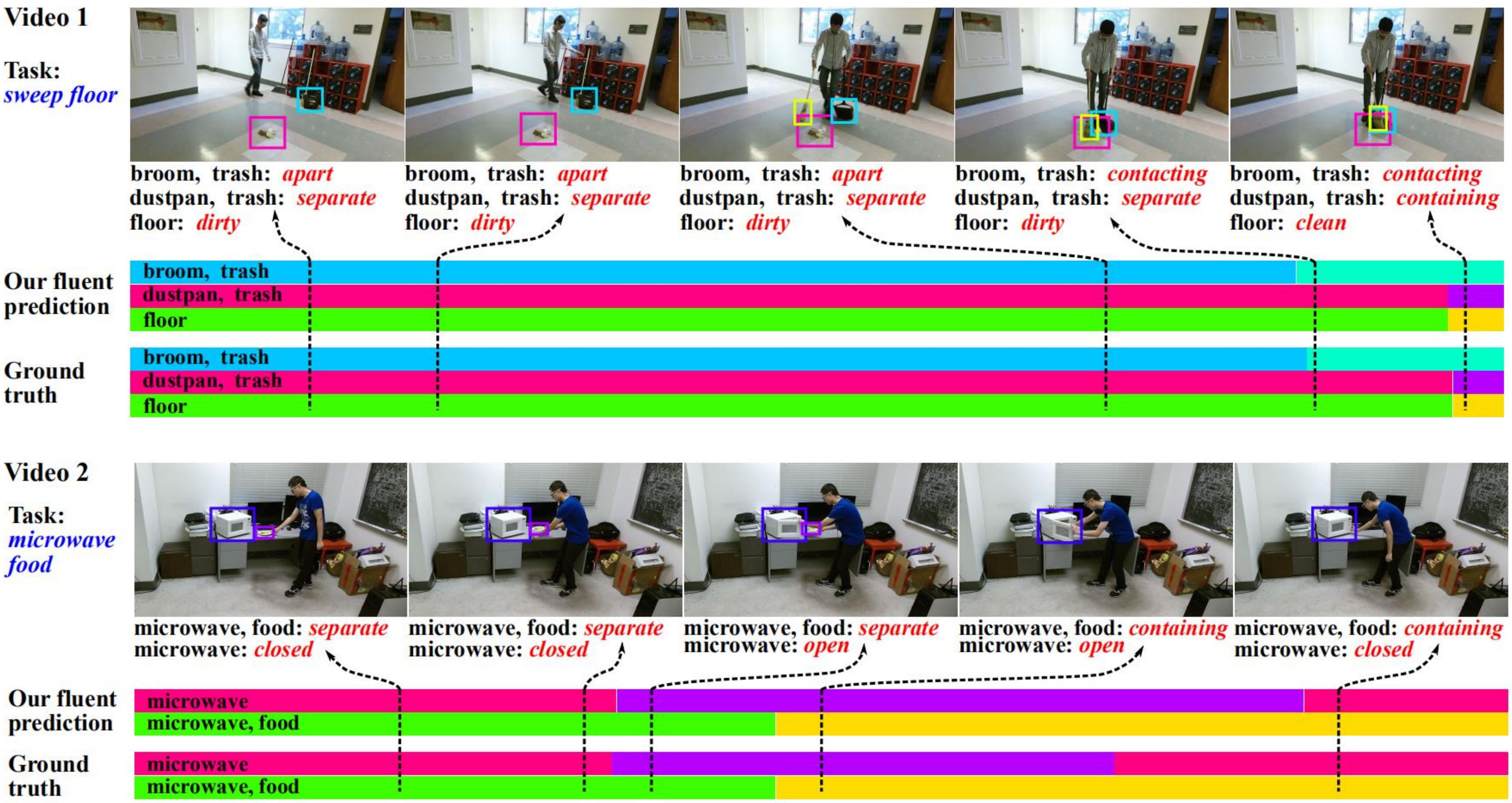

Fig. 5. Visualization of fluent and task recognition in videos.

Thanks for watching

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

Xi’an Jiaotong University, Xi'an, China