Imperial College London

icvl

Transition Forests: Learning Discriminative Temporal Transitions for Action Recognition and Detection

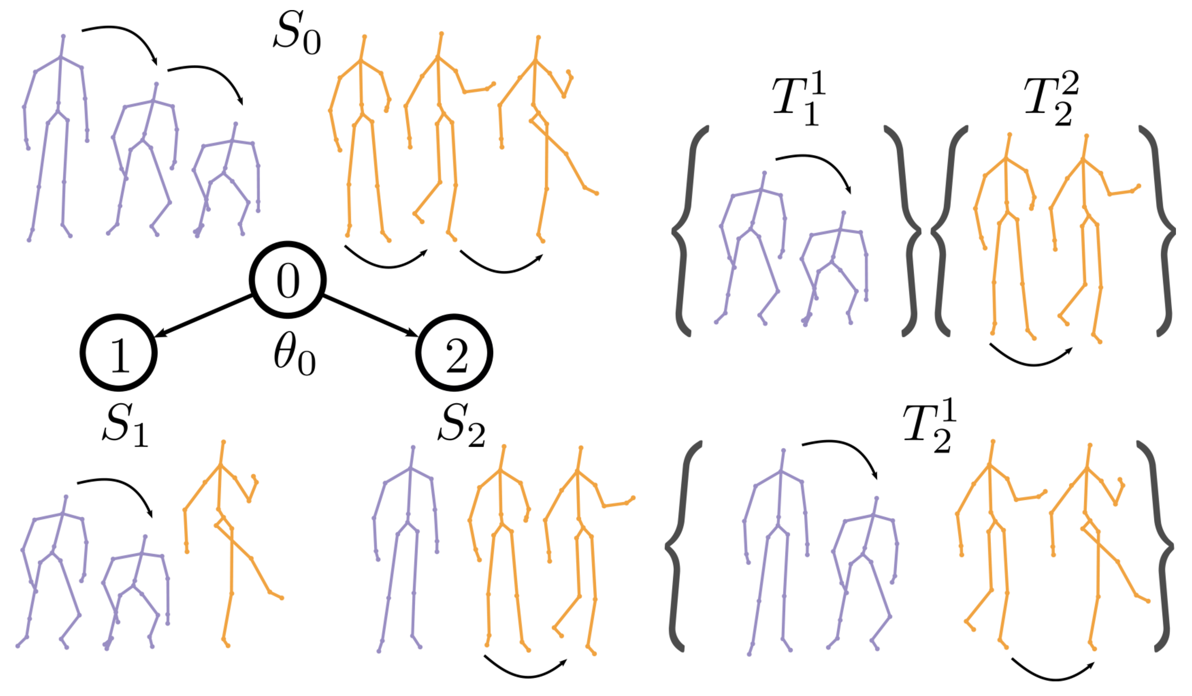
Guillermo Garcia-Hernando and Tae-Kyun Kim

Imperial College London

Overview

- We propose Transition Forests: a temporal decision forest model for human action recognition and detection.
- Growing trees tends to group frames that have similar temporal transitions and share same action label.
- Trees are grown for different temporal orders and combined in prediction.
- Efficient and online per-frame inference.

Transitions



- Transitions as frames traveling from node i to node j in d time-steps on a given tree:

$$T_i^j = \{\{(x_{t-d}, y_{t-d}), (x_t, y_t)\} \mid (x_{t-d}, y_{t-d}) \in S_i \wedge (x_t, y_t) \in S_j\}.$$

Learning transition forests

- Objective function for one layer of the tree:

$$\min_{\{\theta_i\}} E_c(\{\theta_i\}_{i \in N_c}) + E_t(\{\theta_i\}_{i \in N_c \cup N_t}),$$

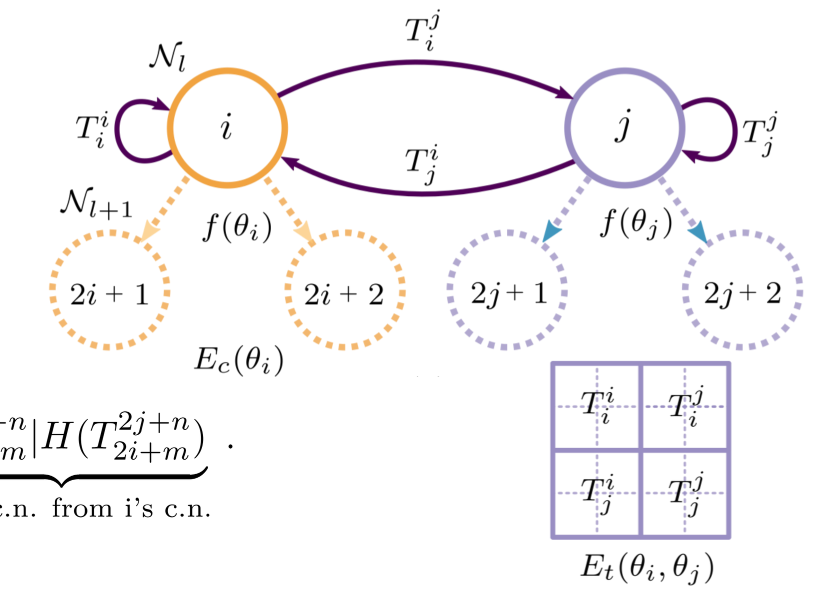
where E_c is the classification loss on single frames, N_c and N_t are the layer nodes randomly assigned to be optimized using either classification or transition objective functions.

- Transition objective function:

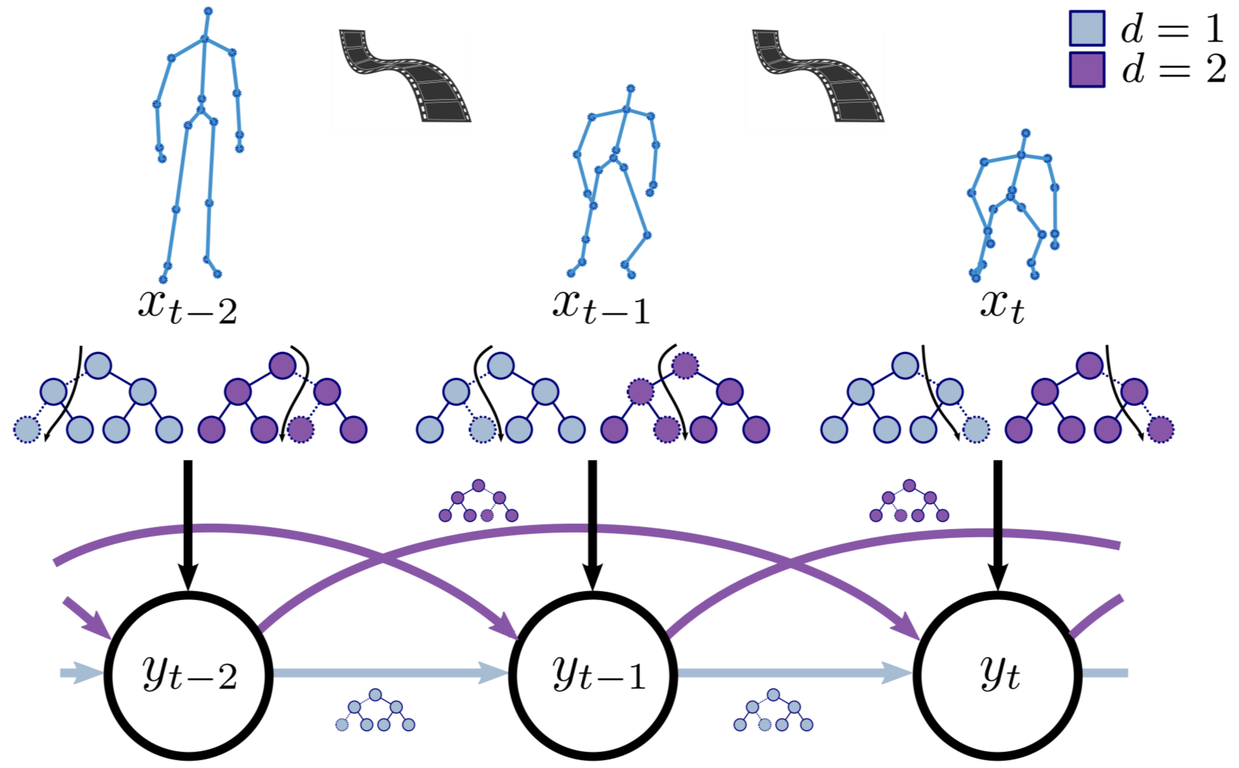
$$E_t(\theta_j) = \sum_{m,n \in \{1,2\}} |T_{2j+m}^{2j+n}| H(T_{2j+m}^{2j+n}), \quad \text{with } H(T_{(\cdot)}^{(\cdot)}) \text{ the Shannon entropy.}$$

Capturing distant node transitions (within a layer): We propose an iterative approach to minimize the objective function (algorithm box in the paper).

$$E_t'(\theta_j | \{\theta_i\}_{i \neq j \in N_c \cup N_t}) = \sum_{m,n \in \{1,2\}} \overbrace{|T_{2j+m}^{2j+n}| H(T_{2j+m}^{2j+n})}^{\text{between j's child nodes (c.n.)}}$$
$$+ \sum_{\substack{i \\ m,n \in \{1,2\}}} \underbrace{|T_{2j+m}^{2i+n}| H(T_{2j+m}^{2i+n})}_{\text{from j's c.n. to i's c.n.}} + \underbrace{|T_{2i+m}^{2j+n}| H(T_{2i+m}^{2j+n})}_{\text{to j's c.n. from i's c.n.}}.$$



Inference



- Transition probability:

$$p_d(y_t | x_t, x_{t-d}, y_{t-d}) = \frac{1}{|\mathcal{M}_d|} \sum_{m \in \mathcal{M}_d} (\pi_{\ell(x_t)}^{\ell(x_{t-d})}(y_t | y_{t-d}))^{(m)},$$

where $\pi_{\ell(x_t)}^{\ell(x_{t-d})}(y_t | y_{t-d})$ is the probability of observing label y_t given that x_t and x_{t-d} reached leaf nodes $\ell(x_t)$ and $\ell(x_{t-d})$ respectively, and previous label hypothesis y_{t-d}.

- Inference equation (frame-based):

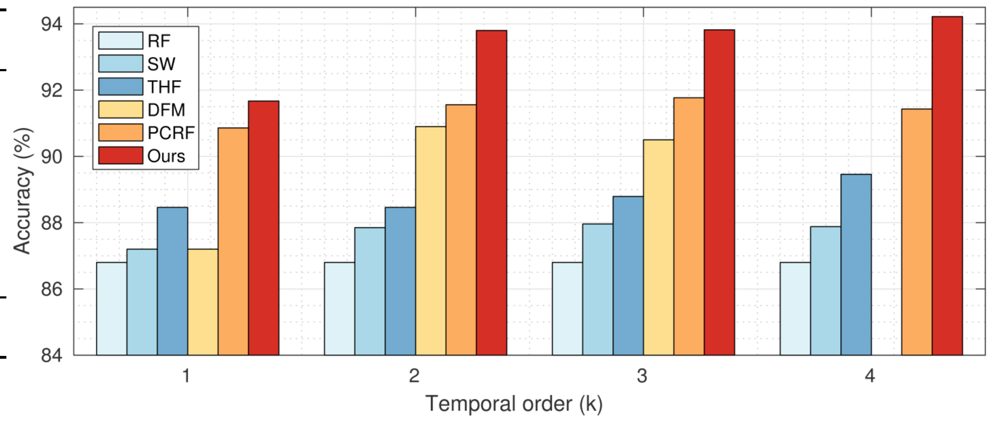
$$p(y_t | x_t, x_{t-1}, ..., x_{t-k}, y_{t-1}, ..., y_{t-k}) = \frac{1}{|\mathcal{M}|} \sum_m (\pi_{\ell(x_t)}(y_t))^{(m)} \frac{1}{k} \sum_{1 \leq d \leq k} p_d(y_t | x_t, x_{t-d}, y_{t-d}),$$

where $\pi_{\ell(x_t)}(y_t)$ is the classification probability (static frame), k is the temporal order of the transition forest and $|\mathcal{M}|$ the ensemble size.

Experimental results

- Comparison with decision forest baselines (action recognition):

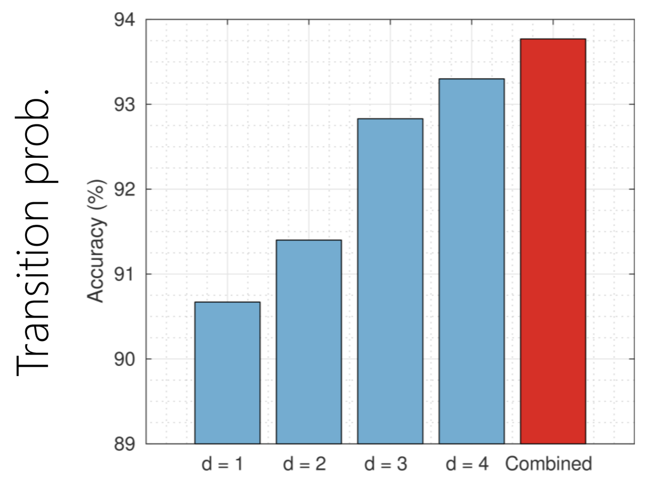
Method	MSRC-12	MSR-Action3D	Florence-3D
Random forest	86.83	87.77	85.46
Sliding window	87.81	90.48	88.44
Hough forest	89.46	91.31	89.06
DFM (CVPR'14)	90.90	-	-
PCRF (ICCV'15)	91.77	92.09	91.23
Ours	**94.22**	**94.57**	**94.16**



- Comparison with state-of-the-art approaches (action recognition):

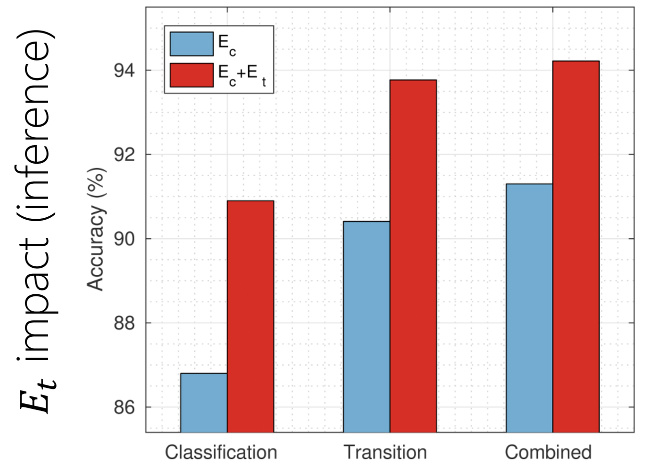
MSR-Action 3D

Method	Year	Real-time	Online	Acc (%)
BoF forest	CVPR'13	✗	✗	90.90
Lie group	CVPR'14	✗	✗	92.46
HBRNN-L	ICCV'15	✓	✗	94.49
Graph-based	ECCV'16	✗	✗	94.77
Gram matrix	CVPR'16	✓	✗	96.97
Key-poses	CVPR'16	✓	✗	**97.44**
PCRF (our result)	ICCV'15	✓	✓	92.09
HURNN-L	ICCV'15	✓	✓	**93.57**
Ours	CVPR'17	✓	✓	**94.57**



Florence 3D

Method	Year	Real-time	Online	Acc (%)
Bag of poses	CVPR'13	✗	✗	82.15
Lie group	CVPR'14	✗	✗	90.88
PCRF (our result)	ICCV'15	✓	✓	91.23
Rolling rot.	CVPR'16	✗	✗	91.40
Graph-based	ECCV'16	✗	✗	91.63
Key-poses	CVPR'16	✓	✗	92.25
Ours	CVPR'17	✓	✓	**94.16**



- Comparison with baselines and state-of-the-art (online action detection):

OAD dataset

	Baselines			State-of-the-art		
	RF	SW	PCRF	RNN	JCR-RNN (ECCV'16)	Ours
F1-score (Action)	0.578	0.556	0.607	0.600	0.653	**0.712**
F1-score (Start frame)	0.361	0.366	0.378	0.366	0.418	**0.514**
F1-score (End frame)	0.391	0.326	0.412	0.376	0.443	**0.527**
Inference time (s)	0.59	0.61	3.58	3.14	2.60	**1.84**

CVPR July 21-26 HONOLULU 2017

Imperial Computer Vision and Learning Lab

Visit us!