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 \( s \) to node \( t \) in \( d \) time-steps on a given tree:

\[
T_d^s = \{(x_r, x_{r+1}, \ldots, x_{r+d}) \in \mathcal{S}^d | x_r \in S, x_d \in T\}.
\]

Learning transition forests
- Objective function for one layer of the tree:

\[
\min \left\{ E_{\text{class}}(h_{v_{0}v_{1}v_{2}}) + E_{\text{trans}}(h_{v_{0}v_{1}v_{2}}), \right\}
\]

where \( E_{\text{class}} \) is the classification loss on single frames, and \( v_{0}, v_{1}, v_{2} \) are the layer nodes randomly assigned to be optimized using either classification or transition objective functions.
- Transition objective function:

\[
\mathbb{H}(v_{0}v_{1}v_{2}) = \sum_{i=1}^{N} - \frac{1}{N} \sum_{j=1}^{M} \sum_{k=1}^{M} \sum_{l=1}^{M} \mathbb{P}(h_{v_{0}}(a_i | v_{1}), h_{v_{2}}(a_i | v_{2})) \cdot \mathbb{P}(a_i | v_{1}, v_{2}).
\]

Inference
- Transition probability:

\[
P_d(a_i | v_{1}, v_{2}) = \frac{1}{|\mathcal{M}|} \sum_{n=1}^{N} \frac{1}{|\mathcal{M}|} \sum_{m=1}^{M} \mathbb{P}(h_{v_{0}}(a_i | v_{1}), h_{v_{2}}(a_i | v_{2})).
\]

- Inference equation (frame-based):

\[
\mathbb{P}(a_i | v_{1}, v_{2}) = \frac{1}{|\mathcal{M}|} \sum_{n=1}^{N} \frac{1}{|\mathcal{M}|} \sum_{m=1}^{M} \mathbb{P}(h_{v_{0}}(a_i | v_{1}), h_{v_{2}}(a_i | v_{2})).
\]

Experimental results
- Comparison with decision forest baselines (action recognition):
- Comparison with state-of-the-art approaches (action recognition):
- Comparison with baselines and state-of-the-art (online action detection):