Space-Time Tree Ensemble for Action Recognition

Shugao Ma\textsuperscript{1}, Leonid Sigal\textsuperscript{2}, Stan Starlack\textsuperscript{1}
\textsuperscript{1}Computer Science Department, Boston University. \textsuperscript{2}Disney Research Pittsburgh.

Introduction and Motivation: Human actions [2] and interactions are inherently structured patterns of body movements. A single structured model, such as those explored in [1, 3, 4], is insufficient to represent an action category in all but the simplest scenarios. Foremost, the execution of the action may differ from subject to subject; furthermore, the video capture process introduces intra-class variations due to occlusions and/or variations in camera viewpoint. As a consequence, the resulting space-time and appearance variations necessitate using a collection of spatio-temporal structures that can best represent the action at large.

We propose a method that discovers a collection of hierarchical space-time trees from video training data and subsequently learns a discriminative action model that builds on these discovered trees to recognize and spatially localize actions in videos. Both the model parameters and the topology of the tree structures are learned automatically from training data. The only supervision that is needed for learning is the action labels of the training videos, i.e., bounding box annotations on video frames are unnecessary. Fig. 1 illustrates one simple discovered tree and its best match in a test video.

Formulation: A video is represented as a graph \( G = \{ V, A^t, A^s, A^h, E \} \). \( V \) is the set of vertices that are the space-time sub-volumes of the video. \( A^t, A^s \) and \( A^h \) are the time, space and hierarchical adjacency matrices containing edge labels. The rows of matrix are visual features extracted from the vertices. For each action class \( a \), a collection of trees is then used in constructing an ensemble classifier:

\[
S_a(G, T) = w^T \cdot \Phi(G, T) = \sum_{m \in \{1, \ldots |T| \}} w_m \phi_m(G, T_m),
\]

where \( G \) denotes a test input video, \( T \) is the set of learned tree structures for class \( a \) and \( T_m \) is one of such trees in this set, and \( w = \{ w_m | m \in \{1, \ldots |T| \} \} \) is the learned weight vector. Each \( \phi_m \) is a scoring function that measures compatibility (or degree of presence) of \( T_m \) in video \( G \). In the multi-class classification setting, the predicted action class \( a^* \) of \( G \) is computed by \( a^* = \arg \max_a S_a(G, T) \).

We formalize a tree as \( T_m = \{ N, E^t, E^s, E^h, \beta \} \) where \( N, \{ E^t, E^s, E^h \} \) are the nodes and adjacency matrices respectively. \( \beta \) are discriminative weights associated with the nodes and edges. Each node \( n_i \in N \) is an index into a learned discriminative action word vocabulary \( \mathcal{W}_a \) for class \( a \); each edge \( E^k_i \) (\( k \in \{t, s, h\} \)) is associated with a corresponding temporal, spatial or hierarchical relationship between nodes \( i \) and \( j \), similar to the relations defined for \( A^k \) in graph \( G \). The matching score of a tree to a graph is computed as follows:

\[
\phi_m(G, T_m) = \psi \{ (\beta \cdot \varphi(G, T_m, z) | z \in Z(G, T_m)) \},
\]

where \( z \) is a latent variable that represents a match of a tree \( T_m \) to the video \( G \); \( z \) is realized as \( z = (z_1, \ldots z_{|N|}) \) where \( z_i \) is the index of the vertex in \( G \) that is matched to the \( i \)th node in \( T_m \). \( \psi \) is a pooling function over the matching scores of the set of all possible (partial) matches \( Z(G, T_m) \).

The matching score of a specific match \( z \) to \( T_m \) is:

\[
\beta \cdot \varphi(G, T_m, z) = \sum_{n_i \in N} \beta_1 \cdot p_n(z_i, n_i) + \sum_{k \in \{t, s, h\}} \sum_{E^k_i \in E^k} \beta^k_{ij} \cdot p_k(\langle a^k_{z_2}, E^k_{ij} \rangle).
\]

where \( \beta_1 \) and \( \beta_{ij}^k \) (\( k \in \{t, s, h\} \)) are the tree node weights and edge weights respectively. The function \( p_n \) scores compatibility of the tree nodes with graph vertices; \( p_t, p_s \) and \( p_h \) score compatibility of the temporal, spatial

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.