Saliency Tubes: Visual Explanations for Spatio-Temporal Convolutions

Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp and Ronald Poppe

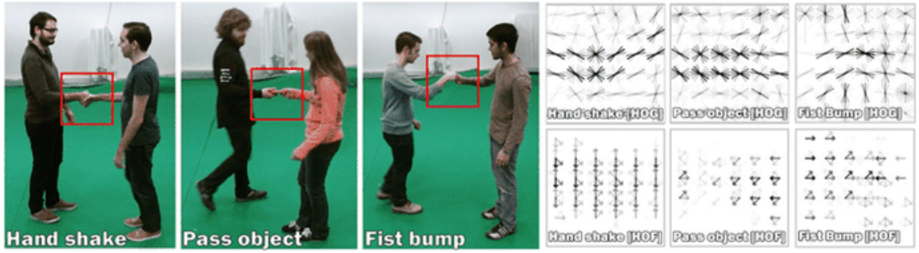
Universiteit Utrecht

University of Essex

London South Bank University
EST 1892

Action Recognition - brief overview

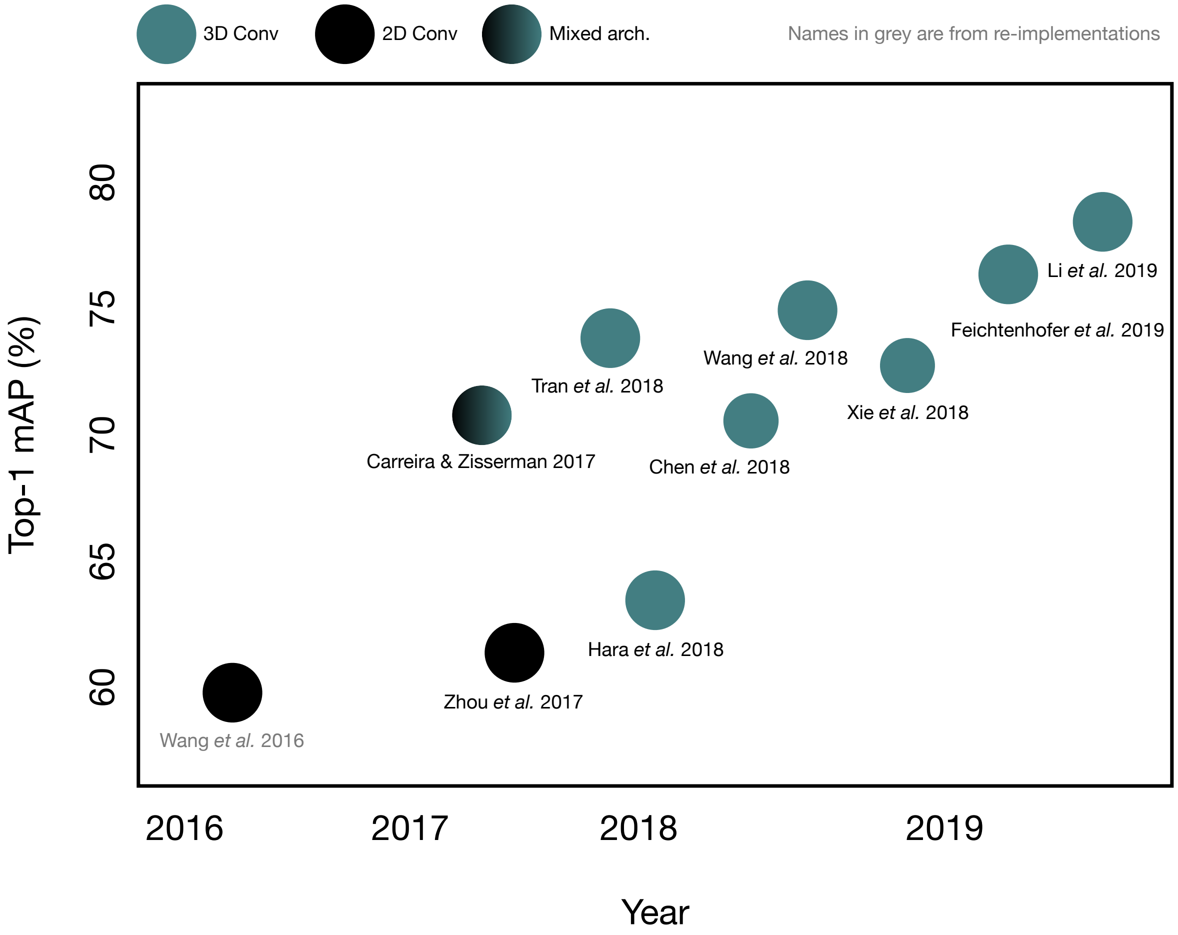
- Automated recognition of actions remains a challenging topic as the coordination of movements across time defines the nature of their collective behavior.
- Early works focused on the classification of videos through the use of programmed/hand-crafted features (SIFT, HOG, MBH, FV, DPMs, DT, etc.)



- However, the extension of CNNs to spatio-temporal series such as videos, have shown noticeable improvements with a higher degree of feature complexity being iteratively learned.
- This is also due to the wide availability of raw video material which lead to the creation of datasets such as Kinetics, Moments in Time, HACS etc.

Milestones in action recognition

Progress in action recognition through the years (based on the Kinetics-400 dataset).



Spatio-temporal convolutions at a glance

- **Differences to 2D Convs**: Additional temporal signal is integrated in the input and at the output volume produced.
    - No separate stream of temporal-only information is created. Input is convolved over space and time.
    - No recursion cells required as time is embedded.

2D Convolutions are based on the notion that each of the image dimensions is processed symmetrically (width and height). However, this strong connection does not extend to time-inclusive volumes where the relationship between the temporal and both spatial dimensions are weaker.

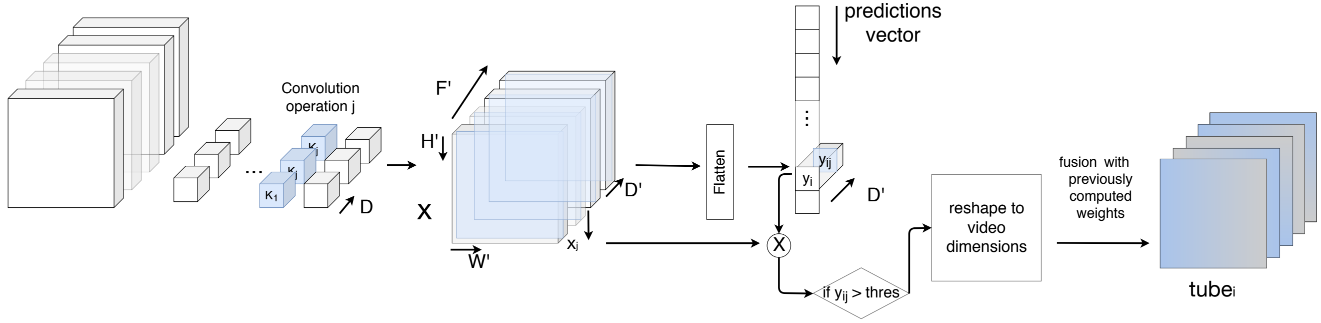
Human interpretative networks

Altough 3D-CNNs have shown great leaps in terms of recognition performance compared to other methods, there are still difficulties on creating **visual huaman-friendly feedback**.

What is required is an way of mapping the learned class-associated featured features to the same dimensionality as the input.

Methods as such have been widely explored for static images where the dimensionality is significantly less complex given the direct relationship between width and height which is not preserved in the time dimension as well.

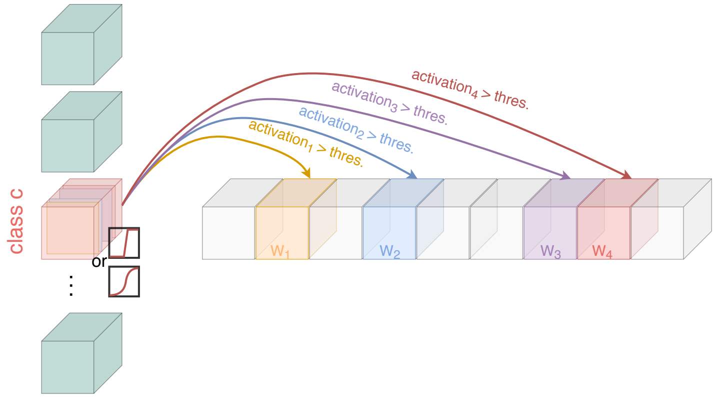
Saliency tubes pipeline



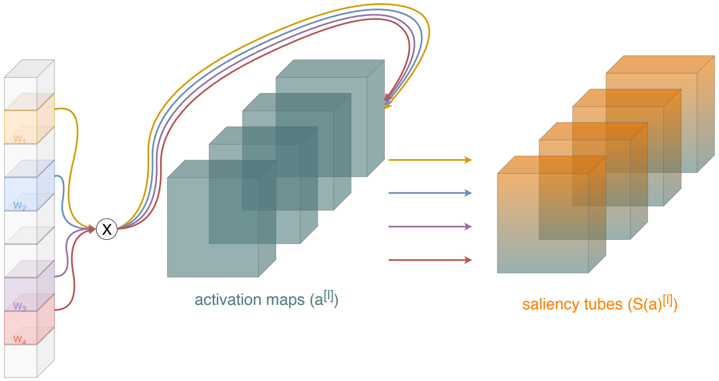
- Specific class weights are selected given a corresponding class to visualize the features of.
- Thresholding activations/weights can be done in expense of possible **global information bias** (i.e. dropping smaller activations that may be class specific).
- Weights or the weighted activations are concatenated to a single volume that represents the information in a layer-specific feature-space.
- The layer-specific feature space is extended through spline interpolation for adjacent spatio-temporal regions.
- The produced tube-like activations are applied through either a heatmap or focus effect on top of the original input creating **heat tubes** and **focus tubes** respectively - as visualization variants of **Saliency tubes**.

Back-stepping through associations between class activations and weights

- The activation maps for a specific class are found based on the produced class index from a softmax function.
- Each activation directly relates to a weight as it is the dot product of that weight and the input activation.
- Selected weight can be found by either threshold like activation (or a sigmodial equivalent).


class c
or
$\text{activation}_4 > \text{thres.}$
$\text{activation}_3 > \text{thres.}$
$\text{activation}_2 > \text{thres.}$
$\text{activation}_1 > \text{thres.}$
W_1
W_2
W_3
W_4

Tube class activations



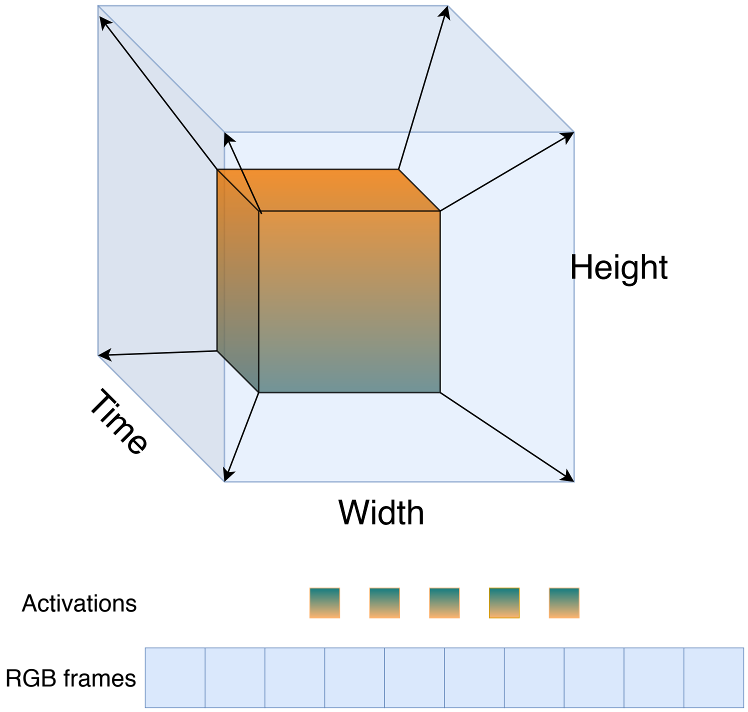
The multiplication process between the class weight vector and the activations can be expressed as a sum over space and time:

$$z_{i,j} = \sum_{f}^{F} \sum_{w}^{W} \sum_{h}^{H} weigth_{i,j} \times \alpha_{f,w,h,j}$$

Where i corresponds to a specific class ($i \in 0, ..., N$, for N classes) and F, W, H are the space-time dimensions or the activations and j is the feature index.

Cosequently the **Saliency tubes** are defined as the sum of $tube_i = \sum_{j}^{D} z_{i,j}$, where D is the set of features equal to the weights to be visualized.

Spatio-temporal dimensionality mismatch



Spatio-temporal activation maps of the last convolution layer, do not exhibit a one-to-one correspondence with the video/clip input volume.

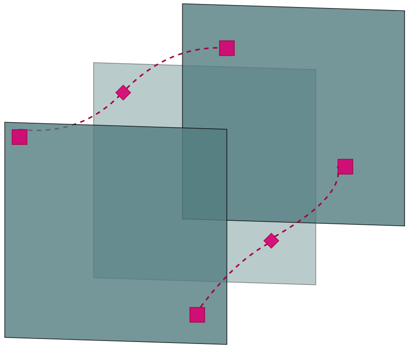
Priors:

- Both are different representations of the same information.
- The space-time activation maps encapsulate the input signal in a (significantly higher) feature space of smaller space-time dimensionality.
- Singular feature vectors in the activations are based on the network's receptive field over a region (in space and time).

Cubic spline interpolation

Given the sets of triplets corresponding to width,height and time $(x_0, y_0, z_0), (x_1, y_1, z_1)\ldots$ we can create quadratic splines for each dimension such that:

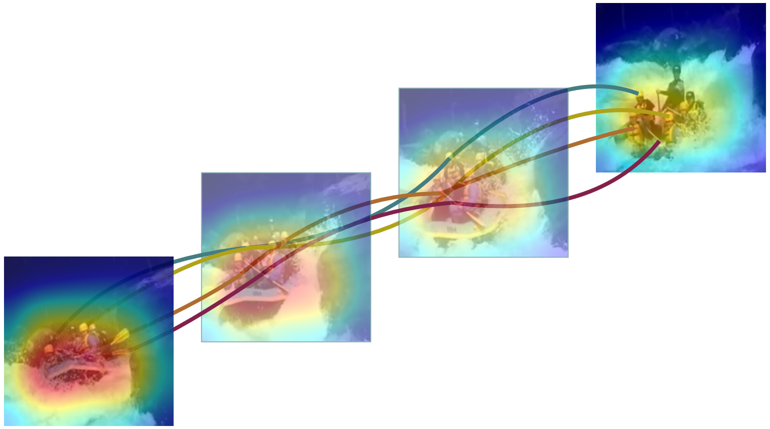
$$\begin{cases} f_x = ax^2 + bx + c, \ \textit{with derivative } 2ax + b \\ g_y = ay^2 + by + c, \ \textit{with derivative } 2ay + b \\ v_z = az^2 + bz + c, \ \textit{with derivative } 2az + b \end{cases}$$



Spline interpolation ins mainly used in order to **differentiate** and additionally **integrate** information that have been merged together.

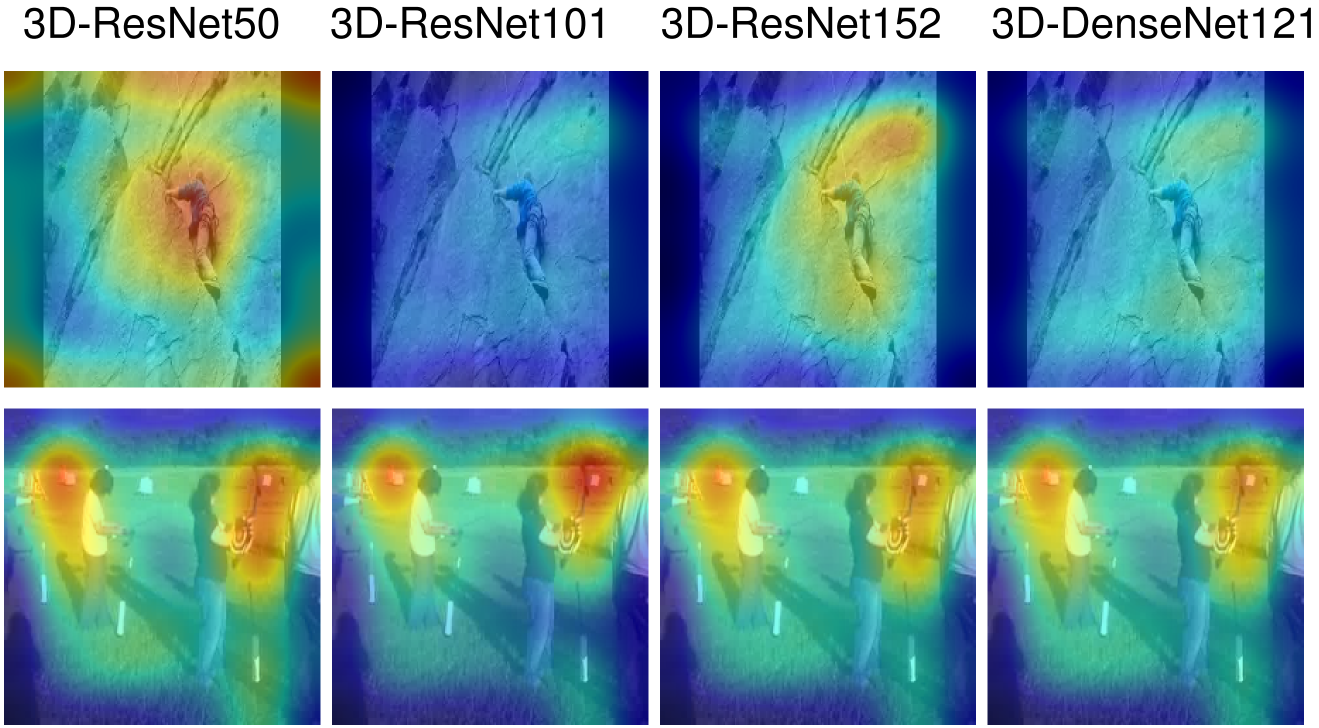
Note: This is only used as an **approximation** for bridging the gap between the dimensionalities. It should **not** be considered as a fixed solution.

Using spline interpolation for intermediate activation-maps



Through interpolating each of the dimensions in width, height and time the activations can be extended to the same dimenionality as the input creating the **Saliency Tube** effect.

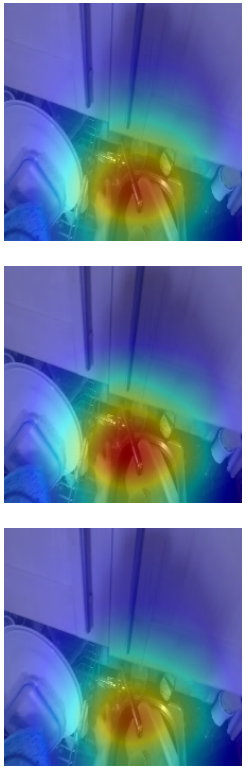
Results on different architectures



3D-ResNet50 3D-ResNet101 3D-ResNet152 3D-DenseNet121

As the produced visualizations are directly related to the overall network structure and inheritedly its complexity, smaller shallow networks produce larger spatio-temporal saliency regions.
In turn, as the complexity of the architecture increases, the saliency regions become more specific.

Focus variations across different classes



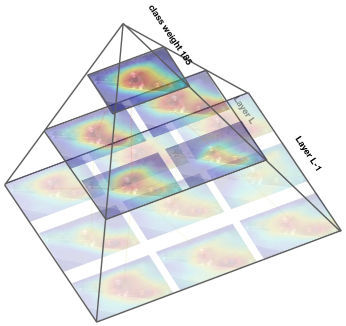
In many instances, and especially ego-centric videos, classes can include similar features. Example of **wash** class while the saliency tubes created are for classes:

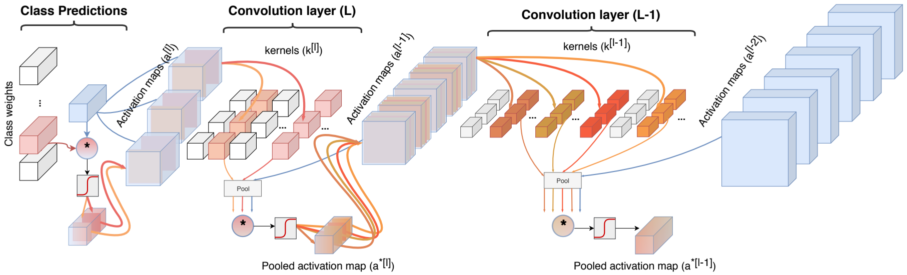
- **Take**: which has similarity by picking and holding item(s).
- **Wash**: ground truth label.
- **Turn-on**: which is similar in the sense of the overall movement performed by the actor.

Shortfalls

- **Visualized features are only of certain degree of complexity.** This is directly related to the curse of dimensionality. As the number of features related to a specific class can only correspond directly to the final convolution layer only a certain degree of feature complexity can be visualized.
- **The entire network is not explored.** Since connections across layers are not based on linear operations there is not a direct procedure for visualizing features across different network layers.
- **Spatio-temporal salient regions are general representations.** As the number of features increase the activations of different kernels include different parts of the video an thus their combined activations are less specific to individual regions. This becomes more prominent in shallow-er architectures

Moving Forward – Hierarchical class visualizations across layers and kernels





- Features cane be expressed as a non-linear conjunction of lower-level features.This creates a link between features across different layers
- If layer kernels and input activation maps are both pooled a vectorized representation of both volumes can be made.
- the pooled kernel includes weight information for individual features in the previous layer, while the pooled activation map consists of feature values corresponding to a video/clip used.
- With multiplying the layer kernels and the input, the most dominant features in the activation map, based on that kernel, can be found and further **backstepped**.

Closing Remarks

- We proposed a method for creating visual explanations of classes in videos called **saliency tubes**.
- Saliency tubes can effectively represent informative regions across space and time providing a human-friendly view of the learned class features.
- Saliency tubes can be user regardless of the spatio-temporal dimensions of the final activation maps as they provide approximations based on the given video dimensions.
- Video explanation is available at (`https://youtu.be/JANUqoMc3es`) and our github repo can be found at (`https://goo.gl/xX4nnv`).