THORN: Temporal Human-Object Relation Network for Action Recognition

Mohammed Guermal, Rui Dai, and François Brémond
Inria, Université Côte d'Azur, 2004 Route des Lucioles, 06902 Valbonne
{mohammed.guermal, rui.dai, francois.bremond}@inria.fr

I. Qualitative Study on Learned Adjacency Matrix

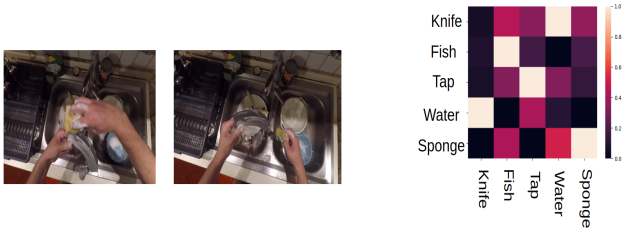

Fig. 1. Example of the learned adjacency matrix of the action from Epic-Kitchen55 dataset. We notice a strong correlation between the classes *knife* and *water* for the action *wash knife*. Thus, we are able to collect high inter-class relation to recognize the right verb and its relevant objects. Moreover, the irrelevant classes such as *fish* are not activated, showing robustness of the learned attention.

In this section, we provide more insight of our THORN model. We show the strength of using the adjacency matrix and the attention mechanism.

In Fig. 1 we show an example of the learned adjacency matrix for the action *wash knife*. In this figure, we find that there is a high correlation between the classes *knife* and *water* in both directions. Whereas the classes *tap, fish* and *sponge* are only correlated to themselves since they are not directly relevant to the objective action class *wash knife*. This example shows the effectiveness of THORN to capture the inter-object relations in the clipped HOI videos.

II. Qualitative Study on the Object Representation Filter

The object representation filter is one of the main parts of our architecture as it allows extraction of a good representation for different objects related to the action. To make sure our filtering work, we extract the activation maps for the different object and see what do they highlight in the scene.

Figures 2, 3, 4, represent different actions with their Class Activation Map (CAM). The example in Fig. 2 represents the action *wash leaf*, when looking at the output of the object representation filter the highest activation where on the classes *leaf* and *tap*. As specified in the main paper, we want to learn features specific to each class. The CAM of tap and leaf in this example clearly shows that only the pixels relevant to the object were highlighted, hence, the feature in the nodes are more representative of the objects of interest.

Moreover, this result shows that our work does similar work to unsupervised object segmentation. Hence, unlike other methods that rely on pre-trained object detectors and tracking methods to extract object and then use ROI-Align to extract objects features, our method is capable of yielding the same result in a unsupervised manner and in a more simplified way. Besides that, our THORN model learns to only focus on objects of interest.

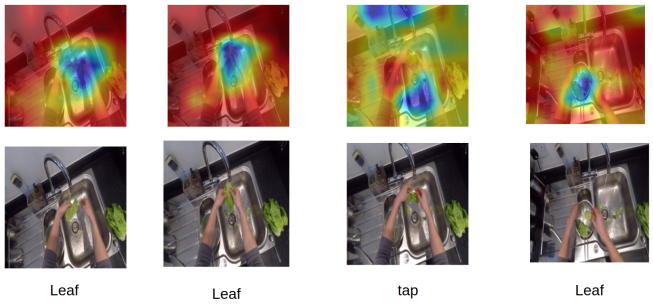

Fig. 2. Example of action *washing leaf*. the highest activated classes were leaf and tap and when inferring the class activation map we can see that most activated pixels are around the objects of interest. Hence, the features extracted are more significant which makes it easier to predict the right action.

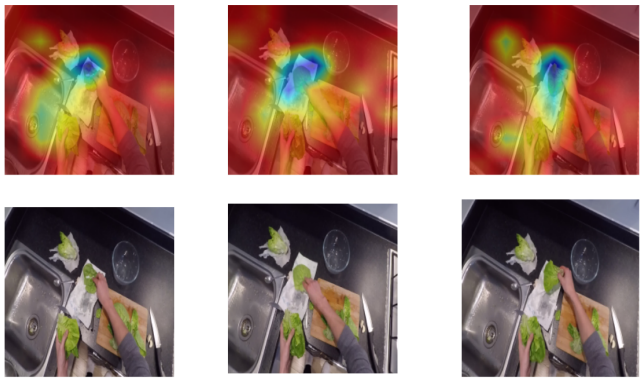

Fig. 3. Example of action *put leaf*. In this example the most activated object was leaf and its activation map shows that the focused-on pixels actually belongs the leaf, proving the strength and robustness of our approach.

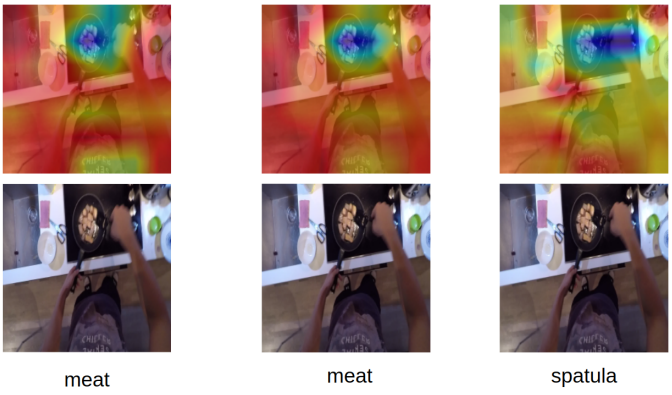

Fig. 4. The action in this figure is *mix meat*, and looking at the figure we notice that the highlighted pixels are the ones corresponding to the spatula and the meat. Therefore, it is easier to predict the right action.