Weakly supervised actor-action segmentation in videos

Jie Chen

Research Prospective

Believe it or not, around 400 hours of new videos are uploaded to the YouTube every minute [1]. While this number may be astonishing, it also calls into the question as how many hours of these newly uploaded videos are truly watched by people within that hour. The uploaded videos, besides frames that are purely scenic spots, mostly contain one or more actors conducting certain activities. It could either be a lion and a deer chasing in a life and death contest, or an old man walking his beloved golden retriever along a seaside beach, or a fierce football match involving dozens of players and hundreds of cheering audience. If we can automatically recognize the actor-action in every frame of those videos, it would save us a fairly amount of time extracting useful information from the videos. Especially, for the newly uploaded video, which lacks in-time manual annotations, they may keep silent in the YouTube repository for quite a while until someone find them. Therefore, this comes out the first direct benefit of our project, i.e., to help people retrieve the most relevant videos, and even the most relevant frames that they are interested in.

Once we have recognized the actor-action in the video, we can do one thing further, i.e., mood analysis, which facilitates the recommendation of the background music that best fits the current scene. For example, the play of Hungarian Rhapsody will definitely increase the intensity of mental tension as we watch a fierce bullfighting. While a piece of soft melody would be a better choice for a scene depicting a couple that ride a bicycle along the riverside. All of these are based on the correct actor-action detection, so we can map the who does what to how they feel, and finally find a suitable piece of music to describe that mood.

Actually, the actor-action recognition not only benefits the video uploaded for fun, but also the surveillance video for security. Nowadays, we have electronic monitoring equipment installed everywhere, from public zones to right inside our homes. When the real-time actor-action recognition is implemented on these videos, it seems as if we own a smart camera that not only
records silently, but also alerts in case of an emergency. For example, once the system spots a thief puts his/her hand into another person’s bag, it calls the police. Or, once it spots an old lady falling down from the staircase in her home, it immediately calls the ambulance. Situation like this requires people to respond very efficiently, but now, it mostly relies on manual monitoring of the surveillance video. Since no one but only machine can work 24 hours a day without rest, and an emergency can happen at any time, our project could contribute greatly to our society.

Note that, the goal of our project is one step further than recognizing the actor-action in the video. As the title indicates, we also want to segment the objects from the background with clear boundaries. This will be very useful for science fiction movie shooting. Constructing a studio in the outer space is extremely expensive, but it’s cheap to have it on a regular grassland. With our project accurately segmenting out the actor, it is at the director’s favor to switch to whatever theme he/she likes.

In all, our project will positively influence our society in a number of ways, no matter security or entertainment is concerned.

**Outline of the Research Plan**

Our project aims at weakly supervised learning with deep neural network, which means no pixel-level annotation is involved during training. This makes our project challenge. Recently, the Fully Convolutional Networks (FCN) proposed by Long *et al.* [2] is very successful in semantic segmentation. However, it requires full supervision for pixel-wise segmentation. Later on, Tokmakov *et al.* [3] built upon FCN to solve segmentation in videos with weakly supervised learning, in which motion segmentation provides additional information to complement the simple category label. The dataset they tested on is relatively simple with almost one object.

In our project, we want to tackle videos with multiple actors on a more complex Actor-Action Dataset (A2D) [4], thus distinguish different objects by utilizing both appearance and action. The research plan is roughly outlined below, and it is subject to change as the project goes on.

- **Sep. 13 – Sep. 30** Implement the fully-supervised baseline model, FCN on the PASCAL VOC.
- Oct.16 – Oct.31  Design our model for weakly supervised segmentation on A2D dataset.
- Nov. 1 – Nov. 31  Optimize and refine the model.

Resources

Nvidia GPUs for deep neural network.

Reference:


