Video Analysis

Abstract

When moving from image to video analysis with CNNs, the complexity of the task is increased by the extension into the temporal dimension. This dimension can be processed by introducing 3D convolutions, additional multi-frame optical flow images, or RNNs. The architectures can be split into models sensitive to local or global motion. Local methods capture a shorter time period and are suited to detect e.g., gestures, while global methods span a larger time interval and can capture a sequence of actions.

Error rendering macro 'toc'

null

Introduction

A video is a sequence of consecutive images or frames, which form a continuous and smooth impression to the human eye. Video analysis tries to make sense of image features and their behavior in time. From a classical computer vision point of view, the Harris-3D detector and the Cuboid detector are likely the most used space-time salient points detectors. However, they rely on hand-crafted features and are thereby highly problem dependent. Therefore, latest research has focused on learning low-level and mid-level features by supervised or unsupervised learning.

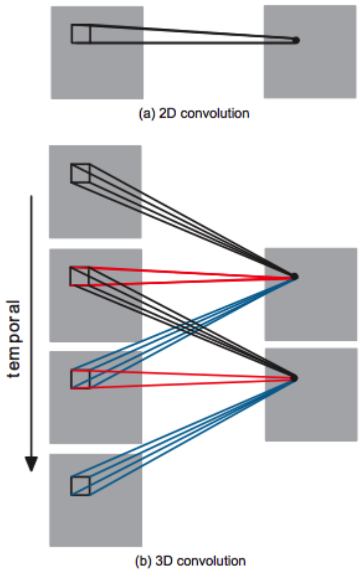


Figure 1: **Comparison of 2D** (a) **and 3D** (b) **convolutions**. In (b) the size of the convolution kernel in the temporal dimension is 3, and the sets of connections are color-coded so that the shared weights are in the same color. In 3D convolution, the same 3D kernel is applied to overlapping 3D cubes in the input video to extract motion features. *Source: (2)*

After the tremendous success of deep learning and especially CNNs with images, the extension to the video domain is obvious. However, compared to the image counterparts, there has been less work and groundbreaking success on video analysis. One obvious reason is the increased complexity, due to the video’s additional temporal dimension.

This extension can be tackled in two ways: On the one hand, space and time can be treated as equivalent dimensions and processed via e.g., 3D convolutions, which is illustrated in *figure 1*. This was explored in the works of Baccouche et al. (1) and Ji et al. (2). On the other hand, one can train different networks, responsible for time and space, and finally fuse the features, which can be found in publications of Karpathy et al. (3) and Simonyan & Zisserman (4).

Besides these supervised learning methods, unsupervised learning schemes for training spatio-temporal (ST) features have also been introduced. Two known approaches involve Convolutional Gated Restricted Boltzmann Machines (5) and Independent Subspace Analysis (6).

The computer vision community has been working on video analysis for decades and tackled different problems such as *event and action recognition*, *anomaly detection*, *video retrieval*, and *activity understanding*. Especially human action recognition finds applications in a variety of domains including intelligent video surveillance, customer attributes, and shopping behavior analysis. However, accurate recognition of actions is a highly challenging task due to cluttered backgrounds, occlusions, and viewpoint variations.

A very demonstrative example for action recognition can be found within the C3D project of Tran et al. (11), which perform sports classification on the Sports-1M (12) dataset. The video footage is displayed with the two top class predictions and the respective confidence scores.

In terms of activity understanding, there has been work on semantic video analysis, which has already seen great success in the image domain. This can be understood as features being merged to objects, whose appearances are kept track of over time. With the use of RNNs, this data is then processed e.g., into video descriptions.

Concepts

There are a variety of promising concepts shown in different publications. We can generally distinguish between architectures that model local or global motion. In other words, a local motion only covers short periods of time and tries to draw an inference from them. If the task is e.g., to distinguish between different arm gestures, the important information is probably encoded in the local details. However, if we want to capture information about the plot of a movie, a longer time window and thereby a more global approach is needed. In other architectures, a fusion between both approaches can be found.

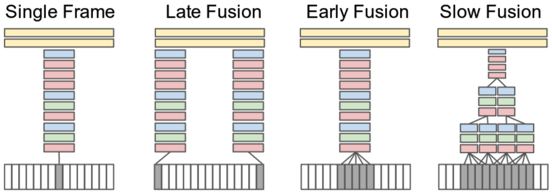


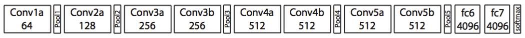
Figure 2: **Approaches for fusing information over the temporal dimension through the network** [3]. They involve convolutional, normalization, and pooling layers, as well as different types of output or fully connected layers. *Source: (3)*

Another major distinction is the method how the temporal dimension is included into the network. Besides 3D convolution, the usage of optical flow, RNNs, and connection via fully connected layers have been proposed.

Figure 2 demonstrates a few different concepts of how to fuse information over the temporal dimension through the network. The *single-frame* model gives the baseline, performing a standard 2D convolution on an individual frame. The *late fusion* places two separate single-frame networks with shared parameters a distance of 15 frames apart and then merges the two streams in a fully connected layer. The *early fusion* model combines information across an entire time window by extending the filter by one dimension and condensing the temporal information in one step. Lastly, the *slow fusion* model is a balanced mix of the two former networks that slowly fuses temporal information throughout the network. This is implemented via 3D convolutions with some temporal extent and stride.

A video sequence can be used to generate an image of multi-frame optical flow, highlighting how much single pixels change over time. Using 2D convolutions on these images helps to extract the temporal information. In RNNs, the outputs of the hidden layers are functions of the input and their previous values, which thereby introduce time dependency into the system.

Locally modeled temporal motion



Early experiments with 3D convolution have been performed by Baccouche et al. [1] and Ji et al. [2]. Karpathy et al. [3] used a multi-resolution architecture to speed up training and were thereby able to train their network on a larger data set. Comparing the different models of *figure 2*, they showed that spatio-temporal convolutions work best if they are done step-wise, s.t. the temporal and spatial information is constantly merged to a larger degree, as illustrated in the slow fusion model. Tran et al. [7] constructed a very clean network with convolutions as 3x3x3 and (almost) all polling layers as 2x2x2. The architecture shown in *figure 3* is also used for the C3D project [11] mentioned in the introduction.

A very interesting 2D convolution approach was introduced by Simonyan & Zisserman [4], which separates the spatial and temporal component by training one ConvNet on a single frame and another one on multi-frame optical flow (*figure 4).* The results of both networks are then fused in the end. The architecture was motivated by the two-stream hypothesis, according to which the human visual cortex contains two pathways: the ventral stream (which performs object recognition) and the dorsal stream (which recognizes motion).

Globally modeled temporal motion

In order to make the system able to process the temporal dimension on a more global scale, RNNs have to be introduced. Common choices are the Long Short-Term Memory (LSTM) network or the Gated Recurrent Unit (GRU), which are very similar in their architecture.

An architecture including temporal processing and a consecutive recurrence was already introduced by Baccouche et al. [1] in 2011. However, it received little attention at that time. In 2015, Donahue et al. [10] demonstrated an architecture (*figure 5*) including 3D convolutions and LSTMs, which could thereby go deeper in temporal space. As the illustration shows, such an architecture is suitable to work on problems with sequential inputs and fixed outputs as activity recognition (deep in time), with fixed inputs and sequential outputs as image description (deep in space), and with sequential inputs and outputs as video description (deep in space and time).

In another publication, Ng et al. [9] combined Simonyan’s and Zisserman’s [4] approach of a spatial and temporal stream with LSTMs to achieve a global model.

In a very recent approach, Balles et al. [10] fused a feed-forward, convolutional and a recurrent network into a single recurrent convolutional network (RCN), which provides a very elegant structure. The GRU-RCN method thereby only requires existing 2D convolution routines.

Figure 3: **C3D architecture.** C3D net has 8 convolutions, 5 max-pooling , and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are 3x3x3 with stride 1 in both spatial and temporal dimensions. The number of filters are denoted in each box. The 3D pooling layers are denoted from *pool1* to *pool5.* All pooling kernels are 2x2x2, except *pool1* is 1x2x2. Each fully connected layer has 4096 output units. *Source: (7)*

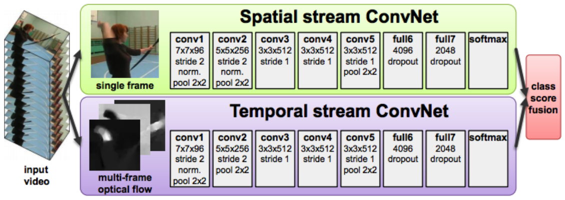


Figure 4: **Two-stream architecture for video classification.** *Source: (4)*

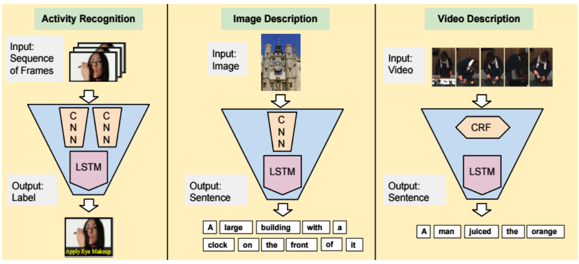


Figure 5: **Task-specific instantiations of our LRCN mode for activity recognition, image, and video description.** CRF stands for Conditional Random Field, a statistical modelling method not further explained here. *Source: (10)*

Literature

1) Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011, November). Sequential Deep Learning for Human Action Recognition. In *International Workshop on Human Behavior Understanding* (pp. 29-39). Springer Berlin Heidelberg.

2) Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEE E transactions on pattern analysis and machine intelligence*, *35*(1), 221-231.

3) Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale Video Classification with Convolutional Neural Networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

4) Simonyan, K., & Zisserman, A. (2014). Two-stream Convolutional Networks for Action Recognition in Videos. In *Advances in neural information processing systems* (pp. 568-576).

5) Taylor, G. W., Fergus, R., LeCun, Y., & Bregler, C. (2010, September). Convolutional Learning of Spatio-temporal Features. In *European conference on computer vision* (pp. 140-153). Springer Berlin Heidelberg.

6) Le, Q. V., Zou, W. Y., Yeung, S. Y., & Ng, A. Y. (2011, June). Learning Hierarchical Invariant Spatio-temporal Features for Action Recognition with Independent Subspace Analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 3361-3368). IEEE.

7) Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4489-4497).

8) Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond Short Snippets: Deep Networks for Video Classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4694-4702).

9) Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

10) Ballas, N., Yao, L., Pal, C., & Courville, A. (2015). Delving Deeper into Convolutional Networks for Learning Video Representations. *arXiv preprint*

Weblinks

11) C3D: Generic Features for Video Analysis (2014). Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M.

12) The Sports-1M Dataset (2014). Accompanies paper **1)**