Raesetje Bonjo Sefala 844165

3D Convolutions for deep learning and its applications to surveillance

Annotated Bibliography

References

Jayabalan, Adhavan, et al. "Dynamic Action Recognition: A convolutional neural network model for temporally organized joint location data." *arXiv preprint arXiv:1612.06703* (2016).

Aim: Recognise human actions in a video using joint movements in 3D space and Convolutional Neural networks rather than Recurrent Neural Networks.

Style/Type: Journal article, Practical

Cross references: This article extends the work of Ji, Shuiwang [2013] which uses 3D CNNs to recognise actions. In addition to Ji, Shuiwang [2013], the model used here can read data from a bigger variety of sources such as Kinect sensors, LIDAR or even smart clothing. They use CNNs to analyse skeletal joint movements in humans to classify the action.

Summary: In this paper the 3D Convolutional neural networks (CNNs) are used to recognise human actions in videos by analysing skeletal joint movements of humans in a video in order to try and classify the action. The algorithm can read data from sources such as Kinect sensors, LIDAR or even smart clothing as they only need data on skeletal joint locations to determine the action. For the CNN model, they used a 3D feature vector consisting of the maximum number of frames in a video, the number of joints associated with any frame and the number of attributes per joint. They used Tensor flow to train the 5-layered CNN model with functions such as pooling, softmax regression and dropout applied to try mitigate overfitting and still allow good accuracy from less training data. Despite the fact that not all actions can be classified using skeletal representations, the model still performed with accuracies of around 87% on standard deviations of 0.3 on the dataset's Gaussian noise.

References

Liu, Zhi, Chenyang Zhang, and Yingli Tian. "3d-based deep convolutional neural network for action recognition with depth sequences." *Image and Vision Computing* 55 (2016): 93-100.

Aim: To use a deep learning-based approach to recognise actions using depth sequences and their respective skeletal joint information

Style/Type: Journal article, Practical

Cross references: This article extends the work of Jayabalan [2016] by including depth sequences to improve the results of Jayabalan [2016]. Unlike Jayabalan [2016] the use of depth sequences can help with representing actions which are difficult to represent using skeletal representations. Additionally this article's model is more regularised for use on other different datasets.

Raesetje Bonjo Sefala 844165

Summary: This paper proposes a 3D Convolutional Neural Network (3D CNN) based approach to better recognise actions in videos n using depth sequences and the corresponding skeleton joint information rather than RGB videos which are sensitive to light variances. The model automatically learns spatial and temporal features from raw depth video sequences using Deep Neural Networks. They propose a framework which depends on 3D CNNs, Support Vector Machine (SVM) classifiers and Joint Vector calculations in order to classify the actions. Torch7 is used for the model's computations and LIBLINEAR is used to classify actions. The trained model also achieves high results when used with other different datasets

References

Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2013): 221-231.

Aim: Use a 3D Convolutional Neural Network model to extract features from both temporal and spatial dimensions automatically without relying on hand crafted features which are usually used in 2D Convolutions.

Style/Type: Journal article

Cross references: In this paper they develop a 3D CNN model which performs subsampling and convolutions separately in multiple channels of information from adjacent video frames rather than in Feichtenhofer, Christoph [2016] where they use use spatial and temporal information advantageously by fusing Convolutional Neural Network towers both spatially and temporally.

Summary: This paper proposes to effectively include motion information in video analysis using 3D Convolutional Neural Networks (CNNs) for human action recognition in videos. Another approach of recognising human actions is to use a 2D CNN and treat videos as a stack of image frames and then analyse each frame individually using CNNs to recognise the action. The problem with this approach is that it does not include motion information within the frames. In this paper they develop a 3D CNN model which performs subsampling and convolutions separately in multiple channels of information from adjacent video frames. A combination of all information from the channels yields the final feature representation. The model was evaluated on airport video surveillance data and the results of the supervised 3D CNN model on the dataset demonstrated great performances in real world environments

References

Chen, Chenyi, et al. "R-cnn for small object detection." *Asian Conference on Computer Vision*. Springer, Cham, 2016.

Aim: To identify small objects in an image using a Convolutional Neural Network-based model by detecting big objects and leveraging their relationship to the small not easily detected object such as the computer screen and a mouse on its left side.

Style/Type: Conference paper

Cross references: In this paper they use 3D Convolutional Neural Networks like in Ji, Shuiwang [2013] to identify objects but they use the model in conjunction with the idea of trying to recognise smaller difficult to detect objects using correlated model identified objects

Summary: This paper investigates small object recognition using their strong co-occurrence spatial relation with bigger object by using a CNN- based model to classify the object. Since it is easier to recognise big objects such as computer monitors, the idea is to use the spatial relation between a mouse (small object) and a monitor to compute the probability mapping of finding a mouse next to the monitor. The spatial relation maps can be learnt as convolutional filters in a CNN and the filter sizes are made large enough to accommodate the space between the objects. Scales were explicitly selected in order to allow convolution with only one filter. After training the model and testing the model, it was observed that the co-occurrence approach almost always produced incorrect results in low recall – high precision range.

References

Bilen, Hakan, et al. "Dynamic image networks for action recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Aim: To propose a long –term pooling operator which is used in a Convolutional neural network to represent long term dynamics for human action recognition.

Style/Type: Conference Paper

Cross references: In this paper they use 3D CNNs like in Ji, Shuiwang [2013] but they approximate rank pooling Convolutional network layer which allows generalising of dynamic images to dynamic feature maps together with automatic feature extraction

Summary: This paper proposes an approximate rank pooling Convolutional network layer which allows generalising of dynamic images to dynamic feature maps. They also demonstrate the successes of their new representations on standard benchmarks in action recognition technology outperforming state of the art performances. When evaluating the effect of end to end training, it was found that multiple dynamic images are better to be pooled on top of the static RGB frames and that multiple dynamic images yield better performances when used in end to end training. A combination of dynamic and static images increases the recognition accuracy by 6% which showing evidence of a complementary relationship. They conclude that dynamic images also allow for state of the art accuracy of action recognition.

References

Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Raesetje Bonjo Sefala 844165

Aim: To use spatial and temporal information advantageously by fusing Convolutional Neural Network towers both spatially and temporally.

Style/Type: Conference paper

Cross references: In this paper they they use use spatial and temporal information advantageously by fusing Convolutional Neural Network towers both spatially and temporally unlike in Ji, Shuiwang [2013] where they develop a 3D CNN model which performs subsampling and convolutions separately in multiple channels of information from adjacent video frames

Summary: This paper proposes an architecture that is able to fuse spatial and temporal cues in feature abstraction at some granular levels with spatial and temporal integration to try and better recognise actions in RGB videos. They found out that fusing a spatial and temporal network at a convolutional layer, this can be done without overall loss of performance but with a significant saving in parameters: that it is better to fuse these networks spatially at the last layer of convolution and that accuracy can be improved by fusing at the class prediction layer: and that pooling of abstract convolutional features over spatiotemporal further increases performance. Although some concussions made in this paper should be treated with caution as datasets can either be too small or noisy.