Activity Recognition from Low Resolution and Low Bit Rate Videos

Sheelabhadra Dey, Savinay Narendra
Texas A&M University
Department of Computer Science and Engineering
sheelabhadra@tamu.edu, savinay@tamu.edu

Abstract

Privacy protection from surreptitious video recordings is an important societal challenge. We desire a computer vision system that can recognize human activities and assist our daily life, yet ensure that it is not recording video that may invade our privacy. In this paper, we present a 3D CNN model for action recognition. Our approach extracts features from both the spatial and temporal dimensions from low resolution videos with spatial size 32x24 by performing 3D convolutions and captures the motion encoded in multiple adjacent frames. The model generates information from the input frames with the final feature representation containing all the information.

1. Introduction

Human activity recognition is one of the intensively studied areas in computer vision. Most existing works do not assume video resolution to be a problem due to general applications of interests. However, with continuous concerns about global security and emerging needs for intelligent video analysis tools, activity recognition from low-resolution and low-quality videos has become a crucial topic for further research.

A system capable of inferring the behaviour of humans would have many applications, from visual surveillance to automatic sports commentary. In particular, a method for classifying an instantaneous human action, or even better, determining a behaviour that may comprise several actions in sequence, would inevitably be a core building block of the system.

Cameras have become ubiquitous and are being extensively used for surveillance applications. They are helpful in monitoring the actions of people and maintaining security. At the same time it is important to not use very high resolution (HR) videos for this purpose as it becomes difficult to apply activity recognition algorithms to them in real-time. Also, since the cameras are connected via. a grid, using HR videos would lower the transmission speeds. Using low resolution (LR) videos at low bit rates (frame rates) could be a possible solution for the issue. Hence, it becomes important to design recognition algorithms that work robustly on LR and low frame rate videos.

Security cameras installed indoors also pose the threat of being hacked compromising the privacy of the residents. It would be helpful if these cameras only capture low resolution images in which human faces are not identifiable but are capable of analyzing their activity. These days, a lot of wearable devices also sport decent cameras which increases the chances of invading privacy without consent. It is important in these situations to only identify activity using cameras without compromising privacy.

In some cases it is useful to have a computer vision system (e.g., a robot) that can recognize human activities and assist our daily life and at the same time ensure that it is not recording video that may invade our privacy. In sports and defense applications sometimes videos need to be captured from a far distance due to which the person who is being observed covers only a few pixels in the screen. It becomes important in these situations to identify the subject's action from such LR videos. Inspired by the above challenges we wish to address the issue of recognition of activities from extreme LR videos; in the range of 32x24 pixels.

Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers. They have a state of the art performance on videos and images classification. Convolutional Neural Networks are very similar to ordinary Neural Networks. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other. And they have a loss function (e.g. SVM/Softmax) on the last (fully-connected) layer. ConvNet architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture.

These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network. CNNs when trained with proper regularization can achieve superior performance on visual object recognition tasks. CNNs have been shown to be invariant to certain variations such as pose, lighting, and surrounding clutter.

CNNs have primarily been used on performing image recognition for 2D images. We explore the use of CNNs for recognizing human activities from low resolution videos. Video frames can be treated as images and we can apply 2D CNN to recognize actions at the individual frame level. However, this model is not able to capture the motion information in the videos. Hence, we propose to use 3D CNNs to recognize human activities from videos so that discriminative features along both the spatial and the temporal dimensions are captured. In this paper we present progress towards such a system by demonstrating a deep learning model to capture the temporal and spatial features of the video using 3D Convolutional Neural Networks.

We evaluate our architecture on UCF-101 and HMDB-51 datasets. The network was able to achieve good results on both the datasets.

2. Literature Survey

Recognition of activities from LR images is an area of active research. Convolutional Neural Networks (ConvNets) have been employed in a number of recent papers to recognize actions and have been reported to outperform algorithms using handcrafted features. [1], [10], and [14] focused on using HR training videos to learn the LR decision boundaries. [3] explored emotion recognition from low bit rate video using a ConvNet and a max-mix training strategy. In this strategy a super-resolution Fully ConvNet was pre-trained with LR-HR pairs generated with the maximum down-sampling factor, followed by fine-tuning the ConvNet model, on a mixture of LR frames that are generated from HR frames using the range of all down-sampling factors. [10] used one HR training video and converted them to multiple LR videos by using different types of LR transforms. This idea was extended in [9] in which a Siamese ConvNet was used to learn the embedding space of the transformation between HR images and LR images. It basically helped the ConvNet learn to cluster LR images derived from the same HR image together while learning to put LR images derived from different HR images in different clusters. This paper also used a 2 stream network that consisted of both a spatial stream as well as a temporal stream to learn the structure in a particular frame and the relationship between adjacent frames respectively. This idea was also employed in [1] in which a 2 stream ConvNet with spatial and temporal streams was used and jointly trained with HR and LR images and their optical flows respectively. This approach of using 2 separate ConvNets on video recognition is very popular and has been used in [11], [5], for activity recognition from HR images.

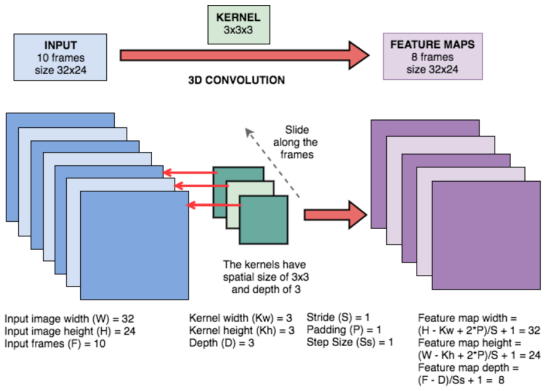

Figure 1. 3DCNN operations

[13] introduced the idea of a 3D ConvNet (C3D) and the authors concluded that they are more suitable for spatio-temporal feature learning as compared to 2D ConvNets. 3D ConvNets learn spatio-temporal features by stacking consecutive RGB frames in the input. They have also been used in [6] in which the architecture generates multiple channels of information from adjacent input frames and performs convolution and sub-sampling separately in each channel.

Combining ConvNets with Recurrent Neural Networks (RNNs) for emotion recognition has been employed in [4] and [15] for activity recognition. However, these implementations did not consider activity recognition from LR videos.

3. 3D Convolutional Neural Networks

In this section we explain in detail the basic operations of 3D ConvNets, talk about different architectures for 3D ConvNets which give us the best results.

3D ConvNet is suitable for spatio-temporal feature learning. When compared to 2D ConvNet, 3D CNNs have a better ability to model temporal information because of 3D convolution and pooling operations. In 3D ConvNets, convolution and pooling operations are performed spatio-temporally while in 2D ConvNets they are done only spatially. 2D CNNs loose temporal information after every convolution operation. 3D Convolution preserves the temporal information of the input signals.

From various studies [13], it is found that 3x3x3 convolution kernels in all layers work best. Small receptive fields of 3 3 convolution kernels with deeper architectures yield best results.

3.1. 3D ConvNet for LR videos

So far 3D ConvNets have been used extensively for activity recognition tasks when the videos have good spatial

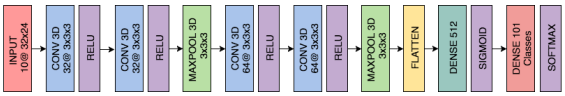

Figure 2. Our 3DCNN Architecture

resolution (112x112 or 224x224 or 320x240). To the best of our knowledge, we couldn't find any work in which 3D ConvNets have been used for activity recognition from LR videos. This motivated us to experiment the performance of 3D ConvNets on LR videos at low spatial resolutions of 16x12 and 32x24.

Our model has 4 convolutional layers with the 2^{nd} and the 4^{th} layer maxpooled. Each layer is a 3D convolutional layer with 32 and 64 filters having a kernel size of 3x3x3. The convolutional layers have a stride of 1x1x1 and has a ReLU activation layer. The final layer is a fully connected layer with a softmax activation function to predict the classes. We used Adam optimizer for our back propagation process and the model was trained for 100 epochs. To prevent overfitting, a dropout rate of 0.25 was used after the 2^{nd} 3D convolutional layer and the dense (512) layer.

[13] explains in detail the feature extraction and the architecture of a 3D ConvNet that has been successful in activity recognition. The learned features, called C3D model appearance and motion simultaneously. It has been observed that C3D starts by focusing on appearance in the first few frames and tracks the salient motion in the subsequent frames. To incorporate a similar architecture it is essential to use a less deeper network than the one proposed in [13]. This is because in [13] the input to the network are videos with a high spatial resolution of 128x171 while our work is concerned with videos having a spatial resolution of 32x24. So, we designed a scaled down version of the original network and ran our experiments on it. The original C3D net has 8 convolution, 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. In our network we used only 1 max-pooling layer as the spatial resolution of the video frames at the input is low. We also reduced the number of convolution layers from 8 to 4 due to the same reason mentioned earlier and used 1 hidden layer for classification. We also used lesser number of filters; 32 in the 1^{st} and 2^{nd} convolution layer, and 64 in the 3^{rd} and 4^{th} convolution layer since the number of features to learn in low resolution images would be low. 512 units were used in fully- connected hidden layer. We added dropouts between layers to reduce overfitting and used batch-normalization after each convolution layer to reduce the network's variability to different initializations. 'relu' activations were used in all the convolution layers and 'sigmoid' activation in the fully-connected layer.

We trained our network using 'adam' optimizer with a batch size of 128 for 50 epochs. Earlystopping on the validation accuracy was used to prevent overfitting.

3.2. 3D ConvNet Ensemble

Ensemble learning is a technique multiple versions of a predictor network are generated to obtain an aggregated prediction. In the last few years, several papers have shown that ensemble method can deliver outstanding performance in reducing the testing error. Most notably, [7] showed that on the ImageNet 2012 classification benchmark, their ensemble model with 5 ConvNets achieved a top-1 error rate of 38.1%, compared to the top -1 error rate of 40.7% given by the single model. In addition, [16] showed that by the ensemble of 6 ConvNets, they reduced the top -1 error from 40.5% to 36.0%.

In [2], experiments were performed on the MNIST dataset using ensemble of ConvNets. Instead of simple averaging the prediction numerically, the outputs of all the ensemble models were stacked to form a new feature map. This feature map was then fed to a new Softmax layer which outputs a vector of the length 10, which is the number of classes. The elements of the new output still represent the probability of the original input belonging to a certain class. It was found that stacking the outputs of ensemble models gave better results than simple averaging. Hence, we followed a similar strategy of ensembling. We used 5 3D ConvNets, trained each of them for 25 epochs, and then stacked their output to form a new feature map. The new feature map was then input to the classification layer with the output layer containing the number of action classes.

4. Dataset

We tested both our architectures on two datasets namely, UCF-101[12] and HMDB-51[8].

4.1. UCF-101

The UCF dataset has 101 action classes and 13320 videos. UCF101 gives the largest diversity in terms of actions and has a presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. The videos in 101 action categories are grouped into 25 groups, where each group can consist of 4-7 videos of an action. The videos from the same group may share some common features, such as similar background, similar viewpoint, etc.

4.2. HMDB-51

HMDB-51 dataset has 51 action classes and contains around 7,000 manually annotated clips extracted from a variety of sources ranging from digitized movies to YouTube

The action categories can be divided into five types: 1) Human-Object Interaction, 2) Body-Motion Only, 3)

Figure 3. 320x240 images converted to 32x24 images

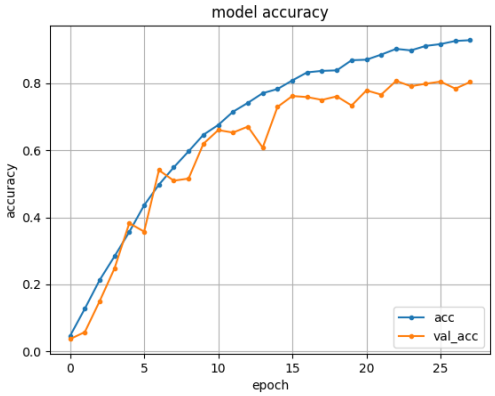

Figure 4. Accuracy curves for UCF-101

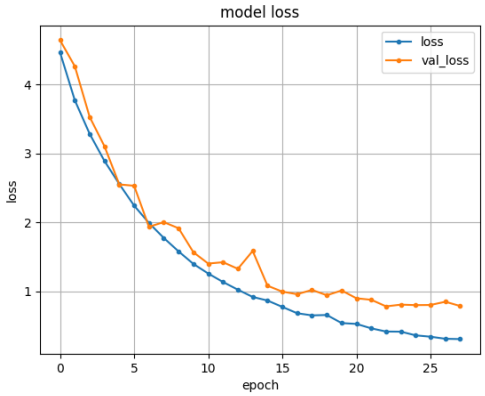

Figure 5. Loss curves for UCF-101

Human-Human Interaction, 4) Playing Musical Instruments, 5) Sports.

4.3. Data Preparation

We converted all the videos to low resolution by downsizing from 320 x 240 to 32 x 24. We divided videos into smaller clips that contained 15 continuous frames for input to the network. We also converted raw RGB images to Numpy.zip format for faster data access. Figure 3 shows samples of images that have been converted from the original 320x240 spatial resolution to 32x24 spatial resolution. It can be observed the operation to a large extent preserves privacy when the subject is close to the camera, but it makes images with a lot of background clutter more messy and hence it is difficult to identify actions from them.

5. Results and Observations

Our model gives us a very promising result for further work and improving upon it. On the UCF-101 dataset, after 28 epochs, we got a training accuracy of 92.87% and a validation accuracy of 80.40%. The training loss on the dataset was 0.3110 while the validation loss was equal to 0.7904. The accuracy obtained on the on the test set was 33.09%. Figure 4 shows the accuracy curves for UCF-101 and Figure 5 shows the loss curves for UCF-101. Figure 6 shows the Confusion matrix on the test data for UCF-101. It can be observed that classes such as 'BasketballDunk' (label 8), 'BlowDryHair' (label 12), 'CricketShot' (label 24), 'HorseRace' (label 40), 'HorseRiding' (label 41), 'IceDancing' (label 43), 'PlayGuitar' (label 62), and 'Surfing' (label 87) have high accuracy rates while 'BreastStroke' (label 18), 'TrampolineJumping' (label 93), and 'PizzaTossing' (label 57) have very low accuracy rates.

The results on the HMDB-51 dataset were higly overfitting on the model developed for UCF-101. After 30 epochs, the training and validation accuracy was equal to respectively 80.16% and 44.31% respectively. Figure 7 showa the accuracy curves for HMDB-51 and Figure 8 shows the loss curves for HMDB-51. Figure 9 shows the Confusion matrix on the test data for HMDB-51. It can be observed that classes such as 'walk' (label 14), 'push' (label 29), 'laugh' (label 37), 'throw' (label 17), and 'kiss' (label 45) have high accuracy rates while 'punch' (label 19), and 'flic' (label 16) have very low accuracy rates.

As expected, running an ensemble of 5 3D ConvNets helped us gain about 2% more accuracy on the test set. The test accuracy and test loss obtained after using the ensemble model was 35.29% and 2.776 respectively. Figure 10 shows the accuracy and loss curves of all the models separately.

We observe that the 3DCNN network is much simpler when compared to 2 stream networks. When evaluating the model's performance on the test dataset, its performance suffered. This is because our model overfitted to the training dataset and due to the presence of similar videos in training and validation set. We also observed that reduced resolu-

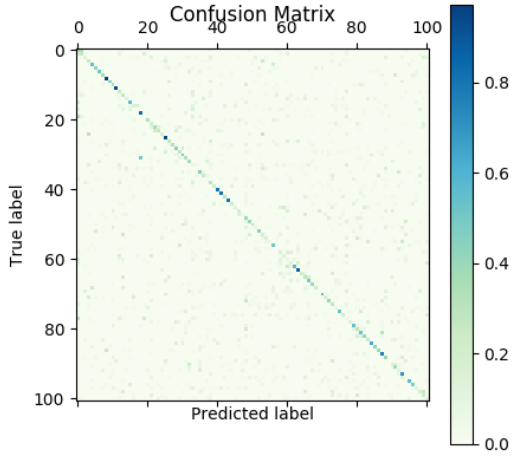

Figure 6. Confusion matrix for UCF-101

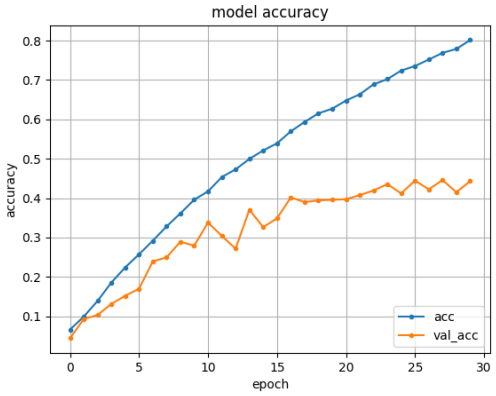

Figure 7. Accuracy curves for HMDB-51

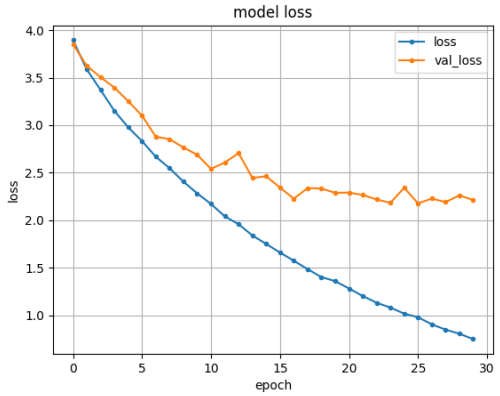

Figure 8. Loss curves for HMDB-51

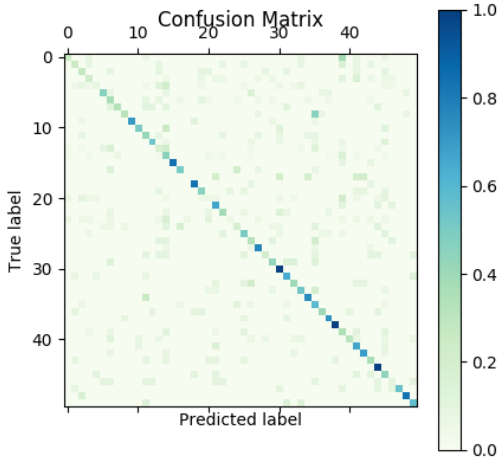

Figure 9. Confusion matrix for HMDB-51

tion makes sports action and clip level action recognition difficult.

6. Conclusion

We developed 3DCNN models for human activity recognition from low resolution videos. These models combine spatial and temporal information by performing 3D convolutions. This architecture combines information from multiple channels of input and performs convolution. The final feature representation is obtained by combining information from all channels. Our model was evaluated on 2 datasets namely UCF-101 and HMDB-51. We were able to achieve good validation accuracy on both the datasets. The test accuracy suffered a bit because there were similar videos in the validation set and the training set and also due to a bit of overfitting. The developed 3D CNN model was trained using a supervised algorithm and requires a large number of labelled samples.

7. Future Work

Improved dense trajectories (iDT) have shown great performance in action recognition, and their combination with the two-stream approach has achieved state-of-the-art performance. We can have a more effective approach of video representation using improved salient dense trajectories combined with 3D Convolutional Neural Networks. First, detecting the motion salient region and extracting the dense trajectories by tracking interest points in each spatial scale separately and then refining the dense trajectories via the analysis of the motion saliency.

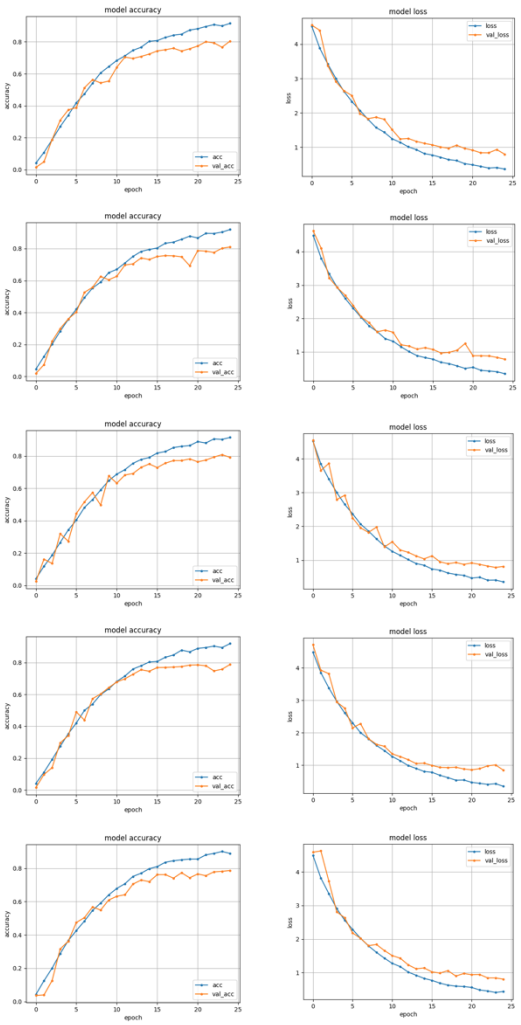

Figure 10. Loss and accuracy curves for ensemble model

We also propose to learn video representations using neural networks with long-term temporal convolutions (LTC). It is also possible to use a Super-Resolution ConvNet to increase the quality of LR images and then use the enhanced images as input to the 3D ConvNet. These are a few techniques and improvements that we would like to explore in the future.

8. Contribution

Tasks covered by Sheelabhadra Dey:

- Preparation of low resolution videos from the original UCF-101 datasets
- Literature review to identify existing work on the topic
- Identification of 3D CNN as the architecture to experiment with
- Development of code for training 3D CNN models on UCF-101 dataset
- Development of code for using 3D CNN ensemble

Tasks covered by Savinay Narendra:

- Preparation of low resolution videos from the original HMDB-51 datasets
- Brainstorming techniques that could be used for LR video activity recognition
- Setting up the system and architecture for developing the code
- Development of code for training 3D CNN models on HMDB-51 dataset
- Preparation of majority of the sections in the final report and presentation

References

[1] J. Chen, J. Wu, J. Konrad, and P. Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 139–147. IEEE, 2017.

[2] L. Chen and G. Shakhnarovich. Learning ensembles of convolutional neural networks, 2014.

[3] B. Cheng, Z. Wang, Z. Zhang, Z. Li, D. Liu, J. Yang, S. Huang, and T. S. Huang. Robust emotion recognition from low quality and low bit rate video: A deep learning approach. *arXiv preprint arXiv:1709.03126*, 2017.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[5] G. Gkioxari and J. Malik. Finding action tubes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[9] M. S. Ryoo, K. Kim, and H. J. Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. *arXiv preprint arXiv:1708.00999*, 2017.

[10] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang. Privacy-preserving human activity recognition from extreme low resolution. In *AAAI*, pages 4255–4262, 2017.

[11] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[12] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[14] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[15] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[16] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.