Fully-Coupled Two-Stream Spatiotemporal Networks for Extremely Low Resolution Action Recognition

Mingze Xu[1]* Aidean Sharghi[2]* Xin Chen[3]† David J. Crandall[1]
[1]Indiana University, Bloomington, IN
[2]University of Central Florida, Orlando, FL
[3]Midea Corporate Research Center, San Jose, CA
{mx6, djcran}@indiana.edu, chen1.xin@midea.com

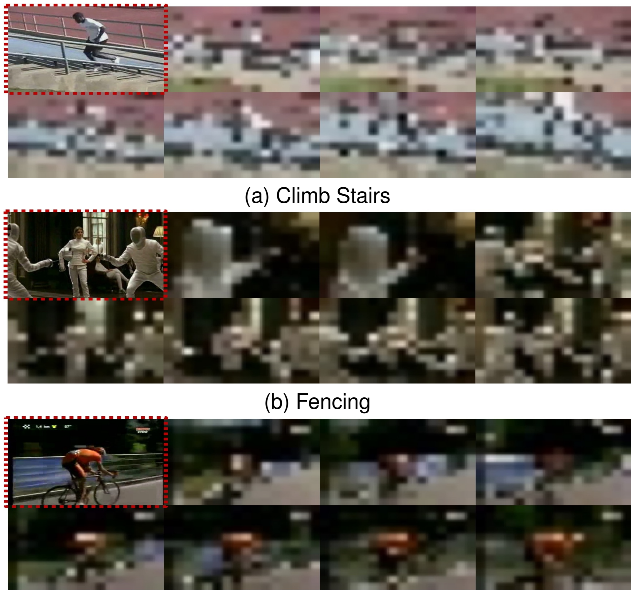
Abstract

A major emerging challenge is how to protect people's privacy as cameras and computer vision are increasingly integrated into our daily lives. A potential solution is to capture and record just the minimum amount of information needed to perform a task of interest. In this extended abstract, we propose a fully-coupled two-stream spatiotemporal architecture for reliable human action recognition on extremely low resolution (e.g., 12×16 pixel) videos. We provide an efficient method to extract spatial and temporal features and to aggregate them into a robust feature representation for an entire action video. We also consider how to incorporate high resolution videos during training in order to build better low resolution action recognition models.

1. Introduction

Cameras are seemingly everywhere, from the traffic cameras in cities and highways to the surveillance systems in businesses and public places. Increasingly we allow cameras even into the most private spaces in our lives. While these cameras have the promise of making our lives safer and simpler, they also record highly sensitive information about people and their private environments.

Perhaps the most effective approach of addressing this privacy challenge is to simply avoid collecting high-fidelity imagery to begin with. For example, low resolution imagery may prevent specific details of a scene from being identified – e.g. the appearance of particular people, or the identity of particular objects – while still preserving enough information for a task like scene type recognition [7]. Particularly important in many home applications of cameras is action and activity recognition, to help give smart devices high-level contextual information about what is going on in the environment and how to react and interact accordingly. Several recent papers have shown that very low resolution videos (around 16×12 pixels) preserve enough information for fine-grained action recognition [1, 2, 4, 5]. This is perhaps surprising, since even a human observer may have difficulty identifying actions from such little information (Figure 1). This raises the question of how much better action recognition on low resolution frames can progress.


(a) Climb Stairs

(b) Fencing

(c) Ride Bike

Figure 1. Sample frames of extremely low resolution (12 × 16 pixel) videos from the HMDB51 dataset. Original high resolution frames are shown in red.

In this extended abstract, we propose a fully-coupled two-stream spatiotemporal network architecture to better take advantage of both local and global temporal features for action recognition in low resolution video. Our model incorporates motion information at three levels: (1) a two-stream network incorporates stacked optical flow images to

[*] Part of this work was done when Mingze Xu and Aidean Sharghi were interns at Midea Corporate Research Center, San Jose, CA.

[†] Corresponding author.

Approach	Accuracy
3-layer CNN [5]	20.81%
ResNet-31 [3]	22.37%
PoT (HOG + HOF + CNN) [6]	26.57%
ISR [5]	28.68%
Semi-coupled Two-stream ConvNets [1]	29.20%
Multi-Siamese Embedding CNN [4]	37.70%
Ours (w/o pre-trained C3D)	**41.04**%
Ours (w/ pre-trained C3D)	**44.96**%

Table 1. Results on the HMDB51 dataset.

Approach	Accuracy
PoT (HOG + HOF + CNN) [6]	64.60%
ISR [5]	67.36%
Multi-Siamese Embedding CNN [4]	69.43%
Ours (w/ pre-trained C3D)	**73.19**%

Table 2. Results on the DogCentric dataset.

capture subtle spatial changes in low resolution videos; (2) a 3D Convolution (C3D) network computes temporal features within local video intervals; (3) a Recurrent Neural Network (RNN) uses the extracted C3D features from videos and optical flow fields to model more robust longer-range features. Our experiments on two challenging datasets (HMDB51 and DogCentric) show that our model significantly outperforms the previous state-of-the-art. Full details of our approach appear in a recent conference paper [10].

2. Our Approach

Spatiotemporal Feature Extractor We propose a feature extractor to capture visual and motion information in low resolution video units. In particular, we use the C3D network, which has proven to be well-suited for modeling sequential inputs such as videos [8]. While the C3D network is able to encode local temporal features within each video unit, it cannot model across the multiple units of a video sequence. We thus introduce a Recurrent Neural Network (RNN) to capture global sequence dependencies of the input video and cue on motion information (e.g., trajectories).

Fully-coupled Networks Inspired by [1, 9], we propose a fully-coupled network architecture where all parameters of both the C3D and RNN networks are shared between high and low resolutions in the (single) training stage. The key idea is that by viewing high and low resolution video frames as two different domains, the fully-coupled network architecture is able to extract features across them.

Two-stream Networks We extend our single-stream network to two-stream by adding a similar architecture but with optical flow fields as the input. Since motion features between consecutive low resolution video frames are often quite small, our model benefits from the optical flow to learn pixel-level correspondences of temporal features.

3. Experiments

Table 1 shows results of the evaluation. Our full model featuring pre-trained C3D networks, the bi-directional GRU network, and the fully-coupled two-stream architecture with sum fusion achieves 44.96% accuracy on the low resolution HMDB51 dataset and 73.19% on the low resolution DogCentric dataset. As shown in Tables 1 and 2, our results beat the state of the art on low resolution video, including Pooled Time Series (PoT) (which uses a combination of HOG, HOF, CNN features) [6], Inverse Super Resolution (ISR) [5], Semi-coupled Two-stream Fusion ConvNets [1], and Multi-Siamese Embedding CNNs [4]. Our best result outperforms these methods by 7.2% (3.3% without pre-training) on HMDB51 and 3.7% on DogCentric.

4. Acknowledgments

This work was supported by the Midea Corporate Research Center University Program, the National Science Foundation under grants CNS-1408730 and CAREER IIS-1253549, and the IU Office of the Vice Provost for Research, the College of Arts and Sciences, and the School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project "Learning: Brains, Machines, and Children." We thank Katherine Spoon, as well as the anonymous reviewers, for helpful comments and suggestions on our paper drafts.

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 *WACV*, 2017.

[2] J. Dai, B. Saghafi, J. Wu, J. Konrad, and P. Ishwar. Towards privacy-preserving recognition of human activities. In *ICIP*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] M. S. Ryoo, K. Kim, and H. J. Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. *AAAI*, 2018.

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

[6] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *CVPR*, 2015.

[7] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 2008.

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv:1412.0767*, 2014.

[9] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang. Studying very low resolution recognition using deep networks. In *CVPR*, 2016.

[10] M. Xu, A. Sharghi, X. Chen, and D. Crandall. Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition. In *WACV*, 2018.