Compressive sensing based privacy for fall detection

Ronak Gupta[1], Prashant Anand[1], Santanu Chaudhury[1,2] Brejesh Lall[1], and
Sanjay Singh[3]

[1] Department of Electrical Engineering, Indian Institute of Technology Delhi
ronakgupta143@gmail.com
[2] Indian Institute of Technology Jodhpur
[3] Cognitive Computing Group, CSIR-CEERI, Pilani, India

Abstract. Fall detection holds immense importance in the field of health-care, where timely detection allows for instant medical assistance. In this context, we propose a 3D ConvNet architecture which consists of 3D Inception modules for fall detection. The proposed architecture is a custom version of Inflated 3D (I3D) architecture, that takes compressed measurements of video sequence as spatio-temporal input, obtained from compressive sensing framework, rather than video sequence as input, as in the case of I3D convolutional neural network. This is adopted since privacy raises a huge concern for patients being monitored through these RGB cameras. The proposed framework for fall detection is flexible enough with respect to a wide variety of measurement matrices. Ten action classes randomly selected from Kinetics-400 with no fall examples, are employed to train our 3D ConvNet post compressive sensing with different types of sensing matrices on the original video clips. Our results show that 3D ConvNet performance remains unchanged with different sensing matrices. Also, the performance obtained with Kinetics pre-trained 3D ConvNet on compressively sensed fall videos from benchmark datasets is better than the state-of-the-art techniques.

Keywords: Fall detection · Human Privacy · Compressive sensing · 3D Convolutional Neural Network · Human Activity recognition

1 Introduction

As per WHO report [12], India is the second most populous country in the world with more than 75 million people lying in the age group of more than 60 years. Human fall is a serious problem concerning people with this age group and is considered as one of the "Geriatric Giants" [12]. Therefore, to address this issue, the need for intelligent monitoring system of the elderly people has risen over the past years. The precise objective for these systems is to automatically detect falls while minimizing false negatives and then to intimate the caregivers/family members.

Several deep learning based fall detection techniques [17,8,3,19] have been presented and for generalization few depend on large action recognition datasets

for pre-training. In [17] authors proposed a scheme for fall detection through ambient camera, where they employed 3D convolutional neural network (3D CNN) to obtain coarse spatio-temporal features, This was followed by Long short-term memory (LSTM) based visual attention mechanism to extract the motion information encoded within the region of interest from coarse spatio-temporal features of the video sequence. The kinetic database Sports-1M which does not have fall data was used for training the 3DCNN. In [3] fall events are detected as a series of sequential change in human pose and these different poses are recognized using CNN. They tried different input image combinations of RGB, Depth, background subtracted RGB to name a few as input to the CNN. Their focus was on human silhouette extracts for recognizing human pose for fall detection.

In this paper, we propose 3D ConvNet architecture which consists of 3D Inception modules for the task of fall detection. The architecture takes spatio-temporal input in compressed domain, rather than spatio-temporal input in image domain as done in Inflated 3D (I3D) architecture. The compressive sensing captures the measurements which are then used for performing classification as a fall or other daily activities (labelled as non fall). In visual systems, while training the fall data is usually generated by simulated falls under a variety of circumstances, that makes it difficult to obtain large quantity of training instances and thus trained classifier has high chance of overfitting the training data. Also, since both the fall dataset used for experiments do not have sufficient training samples, we pre-train the architecture on action recognition datasets for learning better representation of the input videos. This significantly improves the generalization of the deep neural network by giving good detection rates [26,5].

The authors adopt compressive sensing step in the recognition framework which render the compressive samples visually imperceptible. This is essential in circumstances where one might prefer a system which doesn't disclose their identity and capturing all personal activities/details via visual systems/cameras used for detecting falls poses a serious threat to one's privacy. Compressive sensing demonstrates that a signal that is K-sparse in one basis called sparsity basis can be recovered or classified from K linear projections onto a second basis. The latter is called measurement basis which is incoherent with the first. While the measurement process is linear, the reconstruction or classification process has to be done through non linear transformations. It is also a well known fact that the compressive samples of images/video frames containing personal information can essentially be used to achieve privacy. This is because CS transformation is viewed as a symmetric cipher resulting in computational secrecy when the secret sensing matrix is unknown to the adversary [20,21,10].

Although, several privacy based intelligent systems for fall detection have been designed in the past [18]. These systems employ action recognition algorithms which run directly on the camera monitoring the person thus enhancing privacy. Their deployment is done in such a manner that only the fall alarms are transmitted but the the video frames are not. Other popular systems [19] are usually based on thermal heat- maps although capable of masking the persons

identity effectively but are an expensive option. The earlier in-house implementation will be problematic to update when new instances are available [18]. In contrast to the aforementioned approaches, compressive sensing field suggests that a small group of linear projections of a compressible signal contains enough information for reconstruction, classification and processing [15,14,28,27,9,6,13].

2 Related works

Existing non-deep learning fall detection techniques depends on extracting the person (foreground) first, which is highly influenced by image noise (background), illumination variation and occlusion. In [23] authors presented the fall detection by quantifying human shape deformation. For human shape change analysis, they extract and compare two consecutive silhouettes of a person. The landmarks/edge points extracted from silhouette are then matched through video sequence to quantify the silhouette deformation. They compare the mean matching cost of silhouette landmarks and the full Procrustes distance [7] as body shape deformation measures. Based on these shape deformation measures during the fall followed by a lack of significant movement after the fall are fed to Gaussian Mixture Model (GMM) to classify the different activities as fall or not. In [18] the authors presented a fall detection system that uses silhouette area as a feature. Their approach works irrespective of the direction of the movement of the person with respect to the camera. They present a mathematical analysis to confirm the relation between silhouette area and a fall event. The classification is done separately based on the variations of silhouette area as features for SVM classifier.

In [8] authors have proposed a spatial-temporal fall detection method, which can present specific spatial and temporal locations of fall events in complex scenes. In their method, an object detector YOLO v3 [22] is used for person detection, later a deep learning based method for multi-object tracking is used. The features from the tracker are fed to an attention guided LSTM model to detect specific fall events. In [19] the authors presented the use of thermal camera for fall detection which is privacy preserving as it effectively masks the identity of those being monitored. They formulated the fall detection problem as an anomaly detection problem and used Convolutional LSTM Autoencoders to identify unseen falls.

In compressive sensing, random Gaussian matrix or random Bernoulli matrix has been widely used to generate linear measurements of natural images, frames of video, etc. [9]. In practice there are several problems with GRM such as GRM is non-sparse and complicated, and hence highly computational complex and highly difficult in hardware implementation. The other issue is that the measurements generated by GRM are random, neither are data-driven nor adjacent measurements have enough correlation. In literature other measurement matrices have been proposed to solve the above issues. In [6], the authors proposed structural measurement matrix (SMM) to achieve a better Rate-Distortion performance in CS based image coding, in which the image is

sampled by small blocks for better measurement coding while CS recovery can be performed in large blocks for better quality of recovered images. Their method of measurement coding with SMM, helps exploit the spatial correlation in measurement domain, which is represented by directional pixel behaviour (i.e object edges), that improves measurement prediction scheme and reconstructed with large blocks spliced from small correlated blocks improves CS recovery. In [9], the authors proposed a novel local structural measurement matrix (LSMM) for block-based CS coding of natural images by utilizing the local smooth property of images. Their proposed LSMM is a highly sparse matrix and the adjacent measurement elements generated by LSMM have high correlation that has been shown to improve the coding efficiency of spatial information.

Outline of the paper is as follows: Section 3 introduces methodology to solve the problem and the proposed architecture. Section 4 presents experimental results to show the effectiveness of the framework and Section 5 concludes the paper.

3   Methodology

We use 3D ConvNet which includes submodules designed from Inception-V1 network architecture for fall detection. The submodules present in Inception-V1 architecture are inflated as done in I3D Convolutional neural network [5] to construct 3D ConvNet. The inflated Inception-V1 modules are found to be more effective in action recognition compared to VGG-style 3D CNN [5]. There are four inflated Inception submodules in our 3D ConvNet architecture. For fall detection, our 3D ConvNet takes compressed measurements of video sequence as spatio-temporal input, obtained from compressive sensing framework (as shown in Figure 1), rather than video sequence as input, as in the case of I3D convolutional neural networks. Here, the compressed measurements for RGB frames of given video sequence are stacked together along the color (RGB) channel dimension. Figure 2 shows the fall detection architecture.

We adopt a compressive sensing step in the recognition framework which render the compressive samples visually imperceptible, a necessity for privacy. When block based compressive sensing is performed over video frame, we get compressed measurements for the corresponding block. If the dimension of block is $N(= B^2)$ and when it is multiplied with a sensing matrix of size MxN, we get M measurements and the compression ratio is defined as $r = \frac{N}{M}$. The compressed measurement vectors obtained for corresponding blocks in a frame, are arranged across channel dimension as shown in figure before given as input to fall detection architecture. Hence, when compressive sensing is applied to the frame at block level, the output compressed representation will have spatial dimension depending on the number of blocks in video frame and the channel dimension depending on the compression ratio. Similar rearrangement of images or video frame is also performed in the inverse pixel shuffling operation present in sub-convolutional layer of image or video super-resolution frameworks [24]. The difference between their inverse pixel shuffling operation is that it does not

involve dimensionality reduction. Moreover, the linear transformation involved in CS of the video frame blocks into compressed measurements makes rearrangement of the measurements back to the input frame difficult compared to pixel shuffling in sub-convolutional layer.
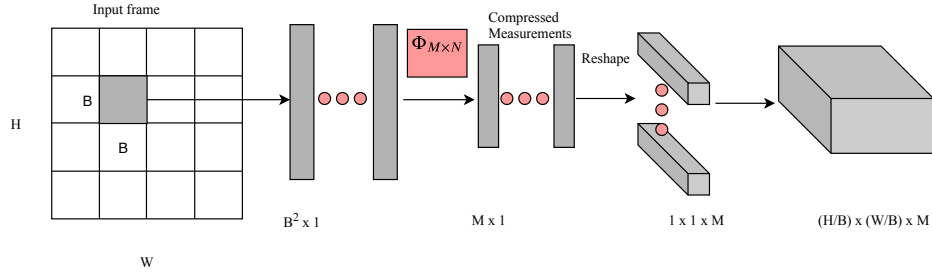
Fig. 1. Compression technique

We show that our CS based privacy for fall detection architecture can work with different compressive sensing matrices. Random Gaussian matrix or random Bernoulli matrix has been used to generate random linear measurements of the video frame blocks. We have also used structural measurement matrix and local structural measurement matrix which exploits intra-block correlation in spatial domain.
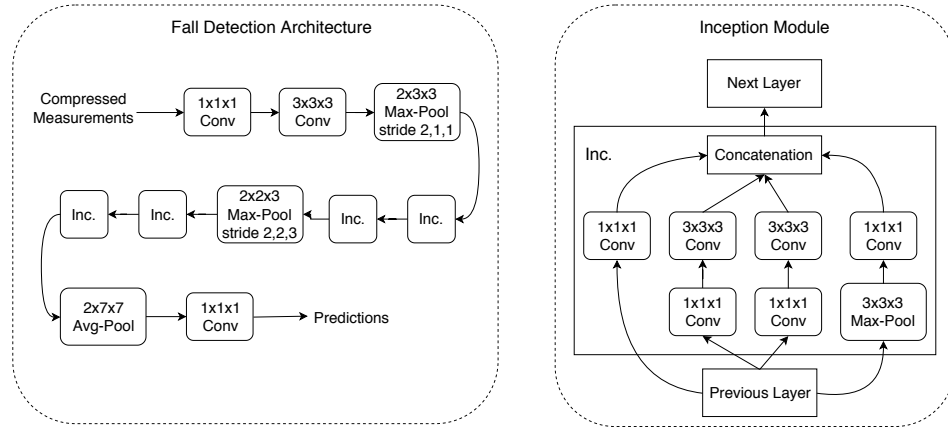

Fig. 2. Fall detection architecture

4   Experimental Results

In this section we report performance of our framework over action recognition and fall datasets with a wide variety of sensing matrices. Once our 3D ConvNet is trained on action recognition dataset, we fine-tune the network for fall detection dataset.

4.1   Fall and Action Datasets

| Caméra 1 | Caméra 2 | Caméra 3 | Caméra 4 |
| Caméra 5 | Caméra 6 | Caméra 7 | Caméra 8 |
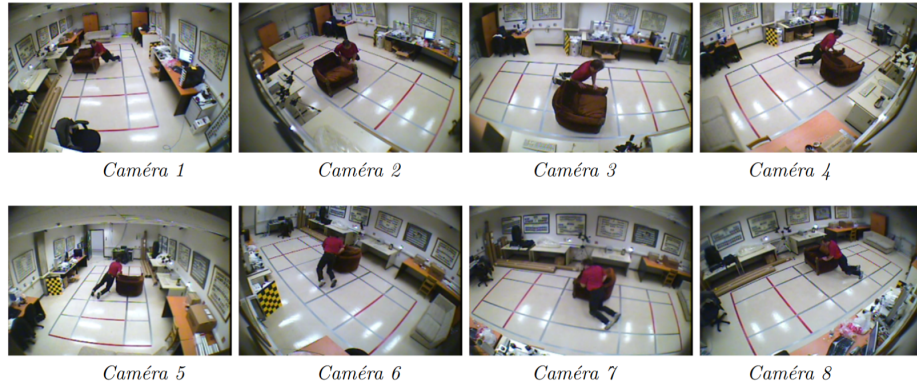
Fig. 3. Fall example from multiple cameras [4]

In [4], the authors collected a dataset of fall and normal activities from a calibrated Multi-camera system, of eight inexpensive IP cameras with a wide angle to cover the whole room. There are 22 scenarios of fall captured by 8 cameras which include sequences of forward falls or backward falls while walking, falls when inappropriately sitting down, loss of balance etc. and 2 scenarios of normal daily activities such as walking in different directions, housekeeping, activities with characteristics similar to falls (sitting down/standing up, crouching down). The fall sequences in dataset are not trimmed action videos as they involve frames containing walking before fall, recovery phase and walking after fall. The temporal annotations of fall is also provided in the dataset which we use to create fall and non-fall sequences. The fall and non-fall video sequences from the first 17 scenarios along with 23rd scenario, are used as training set while the video sequences from 18th to 22nd along with 24th scenario, are used as test set.

In [16], the authors collected dataset containing 70 videos, comprising of 30 fall videos and 40 videos with activities of daily living. Fall and daily activities sequences were recorded with Microsoft Kinect cameras in form of RGB and depth data. Here we create the learning set containing 70 fall and 642 non-fall

sequences with temporal strides. Fall sequences from first 24 fall videos and non-fall sequences from first 32 non-fall videos are used as training set and the rest are used as test set.

For pretraining our 3D CovNet, we create a learning set by randomly selecting 10 classes* from Kinetics-400 dataset [5]. The actions involved in these 10 classes from Kinetics-400 are archery, belly dancing, cheerleading, dodgeball, high jump, playing cello, push up, swimming backstroke, tying tie and washing hair. This subset is composed of around 8K clips of YouTube videos. Each video includes only one actions. The training set, validation set and test set is divided as given in Kinetics-400 dataset.

Table 1. Accuracy on test split of Kinetics dataset with different deep learning architectures

Dataset	Method	Accuracy
Kinetics-400	I3D network (ImageNet pre-trained)	71.1%
Kinetics-10*	I3D network (ImageNet pre-trained)	92.3%
Kinetics-10	I3D network (scratch)	79.73%
Kinetics-10	**3D ConvNet (scratch)**	**78.98**%

Table 2. Accuracy on test split of Kinetics-10* with our 3D ConvNet architecture

Sensing Matrix Type	Compression Ratio			
	4	**16**	**32**	**64**
Random Gaussian Matrix	77.07	77.22	**78.48**	**78.26**
Random Bernoulli Matrix	75.50	75.28	77.22	76.99
Structural Measurement Matrix (SMM) [6]	**78.63**	78.11	77.81	75.58
Local Structural Measurement Matrix [9]	74.98	75.13	77.74	76.99
Convolutional CS Measurement Matrix [25]	77.96	**78.78**	76.62	75.23

Table 1, shows the accuracy performance on test split of Kinetics dataset with different deep learning architectures. Table 2 shows the accuracy results over 10 classes of Kinetics dataset with random Gaussian, random Bernoulli, structural measurement matrix, local structural measurement matrix and Convolutional CS measurement matrix at different compression ratios. We train separately, from scratch, the 3D ConvNet for different compression ratios and different measurement matrices. The performance of 3D ConvNet is more or less similar for the reported measurement matrices. If we train I3D [5] network from scratch over the given classes from Kinetics dataset, the performance comes out to be 79.73% and the performance of our 3D ConvNet comes out to be 78.98%. Since there is small difference in performance between I3D and our 3D ConvNet

with compressive sensing, it is safe to say our 3D ConvNet is sufficient to learn actions for the reported action recognition dataset.

Table 3. Performance of various techniques over Multi-camera fall dataset and UR fall dataset

Method	Compression ratio	Pre-trained on Dataset	Multi-camera fall dataset	UR Fall dataset [16]
Full Proscrustes distance [23]	1 (No privacy)	-	96.20%	-
3DCNN [17]	1 (No privacy)	Sports-1M [11]	99.73%	-
Visual Attention Guided 3DCNN [17]	1 (No privacy)	Sports-1M [11]	99.36%	99.27%
Proposed framework (SMM+ 3DConv Inception Network)	1 (No privacy)	Kinetics-10	100%	100%
	4	Kinetics-10	100%	100%
	16	Kinetics-10	100%	100%
	32	Kinetics-10	100%	100%
	64	Kinetics-10	100%	100%

In Table 3, we report the performance on fall detection dataset using pre-trained 3D ConvNet (over reported action recognition dataset) with structural measurement matrix at different compression ratios. Since fall detection is a binary classification problem, we report 100% accuracy with pre-trained 3D ConvNet. We found that our 3D ConvNet architecture performs better than 3D CNN from [17] for fall detection.

4.2   Implementation Details

All action sequences (including fall and non-fall), were resized to 224x320 before compressed using measurement matrix. We train our model using ADAM optimizer with initial learning rate of 10^{-3} which is reduced by a factor of 10 when validation loss doesn't decrease for 10 consecutive epochs and training is terminated when validation loss doesn't decrease for 22 consecutive epochs. We implemented all the models in TensorFlow [2] and trained and evaluated them on nvidia-docker [1] for Tensorflow on NVIDIA DGX-1.

5   Conclusion

A compressive sensing based fall detection framework has been presented in the paper that also enables privacy preserving since it is a huge concern for patients being monitored through regular cameras. Our deep learning architecture performs similar to I3D network [5], when trained from scratch, in accuracy for reported action recognition dataset, even with wide variety of compressive sensing measurement matrices. Experimental results on Multi-camera fall dataset and UR-Fall dataset were presented to show the effectiveness of the framework at different compression ratios.

6 Acknowledgment

References

1. Nvidia gpu cloud tensorflow, `https://ngc.nvidia.com/catalog/containers/nvidia:tensorflow`, nVIDIA offers GPU accelerated containers via NVIDIA GPU Cloud (NGC) for use on DGX systems.
2. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), `http://tensorflow.org/`, software available from tensorflow.org
3. Adhikari, K., Bouchachia, H., Nait-Charif, H.: Activity recognition for indoor fall detection using convolutional neural network. In: 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA). pp. 81–84. IEEE (2017)
4. Auvinet, E., Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Multiple cameras fall dataset. DIRO-Université de Montréal, Tech. Rep **1350** (2010)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 4724–4733. IEEE (2017)
6. Dinh, K.Q., Shim, H.J., Jeon, B.: Measurement coding for compressive imaging using a structural measuremnet matrix. In: 2013 IEEE International Conference on Image Processing. pp. 10–13. IEEE (2013)
7. Dryden, I.L.: Shape analysis. Wiley StatsRef: Statistics Reference Online (2014)
8. Feng, Q., Gao, C., Wang, L., Zhao, Y., Song, T., Li, Q.: Spatio-temporal fall event detection in complex scenes using attention guided lstm. Pattern Recognition Letters (2018)
9. Gao, X., Zhang, J., Che, W., Fan, X., Zhao, D.: Block-based compressive sensing coding of natural images by local structural measurement matrix. In: 2015 Data Compression Conference. pp. 133–142. IEEE (2015)
10. Hu, G., Xiao, D., Xiang, T., Bai, S., Zhang, Y.: A compressive sensing based privacy preserving outsourcing of image storage and identity authentication service in cloud. Information Sciences **387**, 132–145 (2017)
11. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
12. Krishnaswamy, B., Usha, G.: Falls in older people
13. Kulkarni, K., Lohit, S., Turaga, P., Kerviche, R., Ashok, A.: Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 449–458 (2016)

14. Kulkarni, K., Turaga, P.: Recurrence textures for human activity recognition from compressive cameras. In: Image Processing (ICIP), 2012 19th IEEE International Conference on. pp. 1417–1420. IEEE (2012)
15. Kulkarni, K., Turaga, P.: Reconstruction-free action inference from compressive imagers. IEEE transactions on pattern analysis and machine intelligence **38**(4), 772–784 (2016)
16. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer methods and programs in biomedicine **117**(3), 489–501 (2014)
17. Lu, N., Wu, Y., Feng, L., Song, J.: Deep learning for fall detection: Three-dimensional cnn combined with lstm on video kinematic data. IEEE journal of biomedical and health informatics **23**(1), 314–323 (2019)
18. Mirmahboub, B., Samavi, S., Karimi, N., Shirani, S.: Automatic monocular system for human fall detection based on variations in silhouette area. IEEE Transactions on Biomedical Engineering **60**(2), 427–436 (2013)
19. Nogas, J., Khan, S., Mihailidis, A.: Fall detection from thermal camera using convolutional lstm autoencoder. In: 2nd Workshop on AI for Aging, Rehabilitation and Independent Assisted Living at IJCAI (07 2018)
20. Orsdemir, A., Altun, H.O., Sharma, G., Bocko, M.F.: On the security and robustness of encryption via compressed sensing. In: MILCOM 2008-2008 IEEE Military Communications Conference. pp. 1–7. IEEE (2008)
21. Rachlin, Y., Baron, D.: The secrecy of compressed sensing measurements. In: 2008 46th Annual Allerton conference on communication, control, and computing. pp. 813–817. IEEE (2008)
22. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
23. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Robust video surveillance for fall detection based on human shape deformation. IEEE Transactions on circuits and systems for video Technology **21**(5), 611–622 (2011)
24. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874–1883 (2016)
25. Shi, W., Jiang, F., Zhang, S., Zhao, D.: Deep networks for compressed image sensing. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 877–882. IEEE (2017)
26. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
27. Wakin, M.B., Laska, J.N., Duarte, M.F., Baron, D., Sarvotham, S., Takhar, D., Kelly, K.F., Baraniuk, R.G.: An architecture for compressive imaging. In: Image Processing, 2006 IEEE International Conference on. pp. 1273–1276. IEEE (2006)
28. Xu, K., Ren, F.: Csvideonet: A real-time end-to-end learning framework for high-frame-rate video compressive sensing. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1680–1688. IEEE (2018)