Reconstruction-free action inference from compressive imagers

Kuldeep Kulkarni, Pavan Turaga

Abstract—Persistent surveillance from camera networks, such as at parking lots, UAVs, etc., often results in large amounts of video data, resulting in significant challenges for inference in terms of storage, communication and computation. Compressive cameras have emerged as a potential solution to deal with the data deluge issues in such applications. However, inference tasks such as action recognition require high quality features which implies reconstructing the original video data. Much work in compressive sensing (CS) theory is geared towards solving the reconstruction problem, where state-of-the-art methods are computationally intensive and provide low-quality results at high compression rates. Thus, reconstruction-free methods for inference are much desired. In this paper, we propose reconstruction-free methods for action recognition from compressive cameras at high compression ratios of 100 and above. Recognizing actions directly from CS measurements requires features which are mostly nonlinear and thus not easily applicable. This leads us to search for such properties that are preserved in compressive measurements. To this end, we propose the use of spatio-temporal smashed filters, which are compressive domain versions of pixel-domain matched filters. We conduct experiments on publicly available databases and show that one can obtain recognition rates that are comparable to the oracle method in uncompressed setup, even for high compression ratios.

Index Terms—Compressive Sensing, Reconstruction-free, Action recognition

1 Introduction

Action recognition is one of the long standing research areas in computer vision with widespread applications in video surveillance, unmanned aerial vehicles (UAVs), and real-time monitoring of patients. All these applications are heavily resource-constrained and require low communication overheads in order to achieve real-time implementation. Consider the application of UAVs which provide real-time video and high resolution aerial images on demand. In these scenarios, it is typical to collect an enormous amount of data, followed by transmission of the same to a ground station using a low-bandwidth communication link. This results in expensive methods being employed for video capture, compression, and transmission implemented on the aircraft. The transmitted video is decompressed at a central station and then fed into a action recognition pipeline. Similarly, a video surveillance system which typically employs many high-definition cameras, gives rise to a prohibitively large amount of data, making it very challenging to store, transmit and extract meaningful information. Thus, there is a growing need to acquire as little data as possible and yet be able to perform high-level inference tasks like action recognition reliably.

• *K. Kulkarni and P. Turaga are with the School of Arts, Media and Engineering and School of Electrical, Computer and Energy Engineering, Arizona State University. Email: kkulkar1@asu.edu, pturaga@asu.edu.*

Recent advances in the areas of compressive sensing (CS) [1] have led to the development of new sensors like compressive cameras (also called single-pixel cameras (SPCs)) [2] greatly reduce the amount of sensed data, yet preserve most of its information. More recently, InView Technology Corporation applied CS theory to build commercially available CS workstations and SWIR (Short Wave Infrared) cameras, thus equipping CS researchers with a hitherto unavailable armoury to conduct experiments on real CS imagery. In this paper, we wish to investigate the utility of compressive cameras for action recognition in improving the tradeoffs between reliability of recognition and computational/storage load of the system in a resource constrained setting. CS theory states that if a signal can be represented by very few number of coefficients in a basis, called the sparsifying basis, then the signal can be reconstructed nearly perfectly even in the presence of noise, by sensing sub-Nyquist number of samples [1]. SPCs differ from conventional cameras in that they integrate the process of acquisition and compression by acquiring a small number of linear projections of the original images. More formally, when a sequence of images is acquired by a compressive camera, the measurements are generated by a sensing strategy which maps the space of $P \times Q$ images, $I \in \mathbb{R}^{PQ}$ to an observation space $Z \in \mathbb{R}^{K}$,

$$Z(t) = \phi I(t) + w(t), \tag{1}$$

where ϕ is a $K \times PQ$ measurement matrix, $w(t)$ is the noise, and $K \ll PQ$. The process is pictorially shown

Method	CR = 1	CR = 100	CR =400
Our method ('Type 1' + 'Type 2')	60.86 (2300s) (OM)	54.55 (2250s)	46.48 (2300s)
Recon + IDT	91.2 (FBI)	21.72 (3600s)	12.52 (4000s)
Action Bank [27]	57.9 (FBI)	NA	NA
Jain *et al.*[43]	59.81 (FBI)	NA	NA
Kliper-Gross *et al.*[44]	72.7 (FBI)	NA	NA
Reddy *et al.*[37]	76.9 (FBI)	NA	NA
Shi *et al.*[45]	83.3 (FBI)	NA	NA

TABLE 5

UCF50 dataset: The recognition rate for our framework is stable even at very high compression ratios, while in the case of Recon + IDT, it falls off spectacularly. The mean time per clip (given in parentheses) for our method is less than that for the baseline method (Recon + IDT).

of 100 and above, to perform action recognition, it is better to work in compressed domain rather than reconstructing the frames, and then applying a state-of-the-art method.

Method	CR = 1	CR = 100	CR =400
Our method ('Type 1' + 'Type 2')	22.5 (2200s) (OM)	21.125 (2250s)	17.02 (2300s)
Recon + IDT	57.2 (FBI)	6.23 (3500s)	2.33 (4000s)
Action Bank [27]	26.9 (FBI)	NA	NA
Jain *et al.*[46]	52.1 (FBI)	NA	NA
Kliper-Gross *et al.*[44]	29.2 (FBI)	NA	NA
Jiang *et al.*[47]	40.7 (FBI)	NA	NA

TABLE 7

HMDB51 dataset: The recognition rate for our framework is stable even at very high compression ratios, while in the case of Recon+IDT, it falls off spectacularly.

5 Discussions and Conclusion

In this paper, we proposed a correlation based framework to recognize actions from compressive cameras without reconstructing the sequences. It is worth emphasizing that the goal of the paper is not to outperform a state-of-the-art action recognition system but is to build a action recognition system which can perform with an acceptable level of accuracy in heavily resource-constrained environments, both in terms of storage and computation. The fact that we are able to achieve a recognition rate of 54.55% at a compression ratio of 100 on a difficult and large dataset like UCF50 and also localize the actions reasonably well clearly buttresses the applicability and the scalability of reconstruction-free recognition in resource constrained environments. Further, we reiterate that at compression ratios of 100 and above, when reconstruction is generally of low quality, action recognition results using our approach, while working in compressed domain, were shown to be far better than reconstructing the images, and then applying a state-of-the-art method. In our future research, we wish to extend this approach to more generalizable filter-based approaches. One possible extension is to use motion sensitive filters like Gabor or Gaussian derivative filters which have proven to be successful in capturing motion. Furthermore, by theoretically proving that a single filter is sufficient to encode an action over the space of all affine transformed views of the action, we showed that more robust filters can be designed by transforming all training examples to a canonical viewpoint.

Acknowledgments

The authors would like to thank Henry Braun for their useful suggestions and comments.

References

[1] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, pp. 21 – 30, 2008.
[2] M.B. Wakin, J.N. Laska, M.F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K.F. Kelly and R.G. Baraniuk, "An architecture for compressive imaging," in *IEEE Conf. Image Process.*, 2006.
[3] V. Cevher, A. C. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," in *Euro. Conf. Comp. Vision*, 2008.
[4] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2004.
[5] G. K. M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2003.
[6] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.
[7] I. Laptev, "On space-time interest points," *Intl. J. Comp. Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.
[8] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid *et al.*, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conf.*, 2009.
[9] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conf. Comp. Vision and Pattern Recog*. IEEE, 2011.
[10] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE Intl. Conf. Comp. Vision.* IEEE, 2013, pp. 3551–3558.
[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Comp. Vision and Pattern Recog*. IEEE, 2005, pp. 886–893.
[12] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2009.
[13] A. Klaser and M. Marszalek, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conf.*, 2008.
[14] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, April 2011.
[15] C. Yeo, P. Ahammad, K. Ramchandran, and S. S. Sastry, "High speed action recognition and localization in compressed domain videos," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 18, no. 8, pp. 1006–1015, 2008.
[16] B. Ozer, W. Wolf, and A. N. Akansu, "Human activity detection in MPEG sequences," in *Proceedings of the Workshop on Human Motion (HUMO'00)*, ser. HUMO '00. IEEE Computer Society, 2000, pp. 61–66.
[17] A. C. Sankaranarayanan, P. Turaga, R. Baraniuk, and R. Chellappa, "Compressive acquisition of dynamic scenes," in *Euro. Conf. Comp. Vision*, 2010.
[18] V. Thirumalai and P. Frossard, "Correlation estimation from compressed images," *J. Visual Communication and Image Representation*, vol. 24, no. 6, pp. 649–660, 2013.
[19] R. Calderbank, S. Jafarpour and R. Schapire, "Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain," in *Preprint*, 2009.

[20] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: a spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2008.

[21] K. Kulkarni and P. Turaga, "Recurrence textures for activity recognition using compressive cameras," in *IEEE Conf. Image Process.*, 2012.

[22] I. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172 –185, 2011.

[23] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," vol. 31, no. 8, pp. 1415–1428, 2009.

[24] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *IEEE Conf. Comp. Vision and Pattern Recog*. IEEE, 2005.

[25] H. J. Seo and P. Milanfar, "Action recognition from one example," vol. 33, no. 5, pp. 867–882, 2011.

[26] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Efficient action spotting based on a spacetime oriented structure representation," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2010.

[27] S. Sadanand, J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2012.

[28] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly and R. G. Baraniuk, "The smashed filter for compressive classification and target recognition," *Computat. Imag. V*, vol. 6498, pp. 142–153, 2007.

[29] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Conference in Modern Analysis and Probability (New Haven, Conn.)*, 1982.

[30] X. Zhu, S. Liao, Z. Lei, R. Liu, and S. Z. Li, "Feature correlation filter for face recognition," in *Advances in Biometrics*. Springer, 2007, vol. 4642, pp. 77–86.

[31] P.H Hennings-Yeoman, B.V.K.V Kumar and M. Savvides, "Palmprint classification using multiple advanced correlation filters and palm-specific segmentation," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 3, pp. 613–622, 2007.

[32] A. D, "Database-friendly random projections," *Proc. ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, pp. 274–281, 2001.

[33] S. Sims and A. Mahalanobis, "Performance evaluation of quadratic correlation filters for target detection and discrimination in infrared imagery," *Optical Engineering*, vol. 43, no. 8, pp. 1705–1711, 2004.

[34] R. Bracewell, K.-Y. Chang, A. Jha, and Y.-H. Wang, "Affine theorem for two-dimensional fourier transform," *Electronics Letters*, vol. 29, no. 3, p. 304, 1993.

[35] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *IEEE Intl. Conf. Comp. Vision.*, 2005.

[36] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Intl. Conf. Pattern Recog*, 2004.

[37] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[38] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *IEEE Intl. Conf. Comp. Vision.*, 2011, pp. 2556–2563.

[39] D. Needell and J. A. Tropp, "Cosamp: iterative signal recovery from incomplete and inaccurate samples," *Communications of the ACM*, vol. 53, no. 12, pp. 93–100, 2010.

[40] S. Ali and S. Lucey, "Are correlation filters useful for human action recognition?" in *Intl. Conf. Pattern Recog*, 2010.

[41] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *IEEE Intl. Conf. Comp. Vision.*, 2011.

[42] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2013.

[43] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis, "Representing videos using mid-level discriminative patches," in *IEEE Conf. Comp. Vision and Pattern Recog*. IEEE, 2013.

[44] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *Euro. Conf. Comp. Vision*. Springer, 2012.

[45] F. Shi, E. Petriu, and R. Laganiere, "Sampling strategies for real-time action recognition," in *IEEE Conf. Comp. Vision and Pattern Recog*. IEEE, 2013.

[46] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *IEEE Conf. Comp. Vision and Pattern Recog*. IEEE, 2013.

[47] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *Euro. Conf. Comp. Vision*. Springer, 2012.

Kuldeep Kulkarni

Pavan Turaga