Learning local spatio-temporal features for activity recognition

Graham W. Taylor, Chris Bregler
Courant Institute of Mathematical Sciences, New York University
{gwtaylor,bregler}@cs.nyu.edu

Introduction

Recognition of human activity from video data is a challenging problem that has received an increasing amount of attention from the computer vision community in recent years. The ability to parse high-level visual information has wide-ranging applications that include surveillance and security, the aid of people with special needs and the understanding of human non-verbal communication. Most of the methods proposed for human activity recognition have borrowed ideas from another task: object recognition, which has been dominated by methods that represent each image as a "bag of features" using hand-crafted descriptors applied to image patches. Therefore the majority of methods proposed follow a similar trajectory: detect local interest points, compute a representation of a window of pixels around each of these points using an engineered descriptor, quantize the local space-time features, represent each sequence as a spatio-temporal "bag of features", and then feed to a classifier. Most of the effort has been made in designing space-time interest-point detectors [1, 2, 3] and descriptors [4, 5, 3] and up to this point, learning has not played much of a role in advancing the field.

There is much evidence that learning feature detectors in a supervised setting [6], unsupervised setting [7, 8, 9], or semi-supervised [10] can improve performance in vision tasks, including object recognition. However, other than [11], we know of no methods that attempt to use learning at the level of feature-detectors to improve human activity recognition. This may be due to the prohibitive computational cost of learning descriptors on video. Standard datasets for activity recognition (e.g. [12, 13]) contain typically an order of magnitude more pixels than common datasets for object recognition. However, with the advent of general-purpose GPU computing, and its growing popularity in learning features from images [14], it is now worth considering large-scale feature learning on these datasets.

The first contribution of this paper is to address the problem of learning feature detectors for use in human activity recognition. Specifically, we focus on a recently proposed type of conditional random field called the gated Restricted Boltzmann Machine (GRBM) [15] which learns distributed, domain-specific representations of image transformations. Fundamental to all of the leading activity recognition methods is a step where spatio-temporal descriptors are quantized to generate a "bag-of-words" codebook (typically using K-means). In an attempt to easily obtain compact video representations, much of the potentially useful discriminative power of the descriptors is lost. A second contribution of this paper is to to address this issue. We argue and demonstrate that sparse, overcomplete *distributed* representations are more appropriate for video analysis.

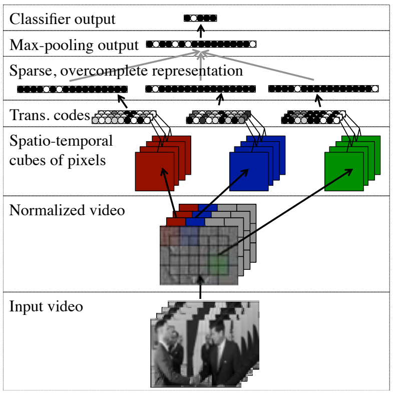


Figure 1: Our proposed architecture. A video is decomposed into spatio-temporal cubes. These are processed by two layers of learned feature detectors and a max-pooling layer before classification.

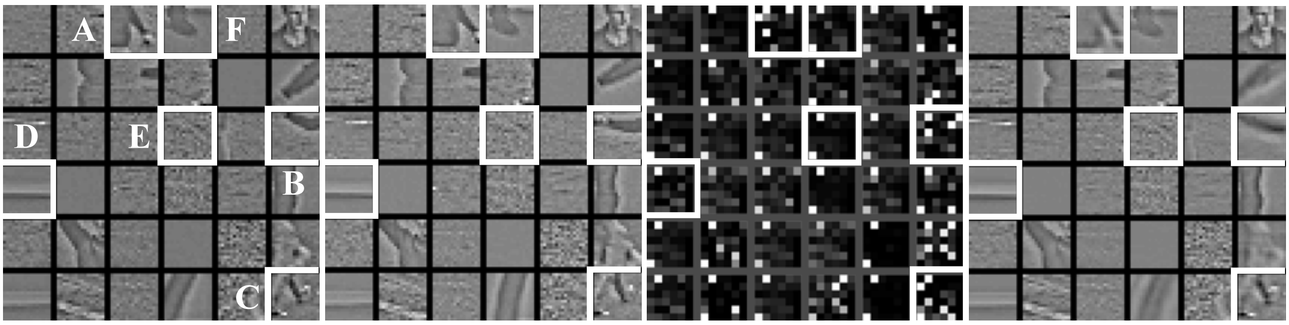
Method

Our proposed architecture is shown in Fig. 1. We first apply local contrast normalization to each of the pixels in the input video. Then we extract local space-time "cubes" from the normalized video. In the first phase of unsupervised learning we estimate so-called latent "transformation-codes" that model local appearance and motion of the space-time cubes (see Fig. 2). Given low-level codes, we infer longer-term/mid-level encodings with a sparse dictionary learning method [16]. Finally max-pooling is used to find a sparse, vector representation for the video which is then fed to an SVM classifier.

Experiments

We evaluated our method on two publicly available datasets intended for benchmarking human activity recognition. Using NVIDIA GPU hardware, it takes about a day to train our low-level feature detectors on millions of image patches.

KTH actions The KTH actions dataset [12] is the most commonly used dataset in evaluating human action recognition. It consists of 25 subjects performing six actions, under 4 scenarios. Each sequence is further divided into shorter "clips" for a total of 2391 sequences. We use the original training and test split so our results are directly comparable to the recent survey by Wang et al. [17]. Our approach using the GRBM descriptors and sparse coding with 4000 dictionary elements, gives a mean accuracy of 89.1±0.4 (averaged over 10 runs). This is, to the best of our knowledge, the best performance on KTH amongst methods that do not use interest-point detection. The currently best performing method [5] uses the STIP interest-point detector and and HOG/HOF or HOF descriptors (91.8 and 92.1%, respectively). All of the descriptors to which we compare use some sort of explicit local geometry (e.g. computing the histograms in a fixed number of spatial regions). All local geometrical information used by our descriptor is captured implicitly in the learned weights.



(a) Input, $\mathbf{x}$ (patch $t-1$) (b) Output, $\mathbf{y}$ (patch t) (c) Encoding, $\mathbf{z}$ (d) Reconstruction, $\hat{\mathbf{y}}$

Figure 2: Random pairs of normalized image patches extracted from the KTH actions dataset: a) Patches at frame $t-1$. b) Patches at frame t. c) Given each pair of patches, we can infer the corresponding transformation code. The codes are vectors, but we have reshaped them to 2-d for easier viewing. Note that that the features are most active when there is motion (compare patches A-C which contain body movement, compared to patches D-F which are either background or stationary body parts.) d) The model's reconstruction of the output, given input and transformation code.

Hollywood2 The Hollywood2 dataset [13] consists of a collection of video clips containing 12 classes of human action extracted from 69 movies. It totals approximately 20.1 hours of video and contains approximately 150 samples per action. It provides a more realistic and challenging environment for human action recognition by containing varying spatial resolution, camera zoom, scene cuts and compression artifacts. Performance is evaluated as suggested by Marszalek et al. [13]: by computing the average precision (AP) for each of the action classes and reporting mean AP over all actions. We achieve an AP of 46.6±0.8% using dense sampling, learned GRBM low-level features and sparse coding with 4000 elements. The best known score (47.4%) uses dense sampling with HOG/HOF features and quantization [17]. Our result outperforms all other known published results, including popular methods like Cuboids [1] (45.0%) and Willems et al. [3] (38.2%).

Conclusion

We present an architecture for human activity recognition that differs from the majority of existing methods in two major ways: 1) Instead of using hand-crafted spatio-temporal interest-point detectors and descriptors we densely sample videos and *learn* our descriptors using millions of patches; and 2) Rather than mapping descriptors to a spatio-temporal "words", we represent them as sparse linear combinations of atoms from a trained dictionary. Our learned feature detectors are sensitive to motion and perform a kind of implicit interest-point detection that is data-adaptive.

Among dense sampling methods, we achieve the current best score on the KTH actions dataset (methods that employ explicit interest-point detectors still outperform our approach). We are able to scale our architecture to perform competitively against state-of-the-art approaches on the challenging Hollywood2 dataset that contains over 20 hours of high-resolution video. Concerning computational cost: although it takes about a day to train the first-layer, at test-time our model is roughly equivalent to other state-of-art methods that employ dense sampling.

Our results support the so-called "deep learning" philosophy which advocates learning multiple layers of distributed representations. We intend to scale our method up to use additional sparse, overcomplete layers. Our long-term goal is to abandon a patch-based approach and train our model convolutionally, employing spatial and temporal pooling layers to permit scale and shift invariance. Finally, as more labeled video data becomes available, we plan to investigate training the lower layers on multiple datasets and global training of all layers.

References

[1] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[2] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.

[3] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. pages 650–663, 2008.

[4] Alexander Kläser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 995–1004, 2008.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[7] G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Comput*, 18(7):1527–1554, July 2006.

[8] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *ICML*, pages 473–480, 2007.

[9] M-A. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *NIPS*, pages 1137–1144, 2006.

[10] V. Nair and G. Hinton. 3D object recognition with deep belief nets. In *NIPS*, pages 1339–1347, 2009.

[11] T. Dean, G. Corrado, and R. Washington. Recursive sparse spatiotemporal coding. In *Proc. IEEE Int. Workshop on Mult. Inf. Proc. and Retr.*, 2009.

[12] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. *ICPR*, pages 32–36, 2004.

[13] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *CVPR*, pages 2929–2936, 2009.

[14] Rajat Raina, Anand Madhavan, and Andrew Ng. Large-scale deep unsupervised learning using graphics processors. In *ICML*, pages 873–880, 2009.

[15] R. Memisevic and G. Hinton. Unsupervised learning of image transformations. In *CVPR*, 2007.

[16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, pages 689–696, 2009.

[17] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, pages 127–138, 2009.

Topic: visual processing and pattern recognition Preference: oral