SPARSE REPRESENTATION BASED ACTION AND GESTURE RECOGNITION

Sushma Bomma, Neil M. Robertson and Paolo Favaro†*

Heriot-Watt University, Edinburgh, UK
visionlab.eps.hw.ac.uk

ABSTRACT

In this paper we present a solution to the problem of action recognition using sparse representations. The dictionary is modelled as a simple concatenation of features computed for each action class from the training data, and test data is classified by finding sparse representation of the test video features over this dictionary. Our method does not impose any explicit training procedure on the dictionary. We experiment our model with two kinds of features, by projecting (i)Gait Energy Images (GEIs) and (ii)Motion-descriptors, to a lower dimension using Random projection. Experiments have shown 100% recognition rate on standard datasets and are compared to the results obtained with widely used SVM classifier.

Index Terms— sparse representation, action recognition, gesture recognition, trained dictionaries, convex optimization, gait energy images, motion-descriptors

1. INTRODUCTION

Action recognition is an important area in computer vision because of its numerous applications like video surveillance, human-computer interface, video indexing etc. Recognizing ongoing activities from an unknown video is a challenging task due to unavoidable hitches like occlusions, view-point variations of cameras, anthropometric differences and so on. There has been an influential research in this area so far. A detailed report on the progress in activity recognition in past decade can be found in [1]. The performance of an activity recognition system largely relies on the type of the features used and the classification strategy employed. The features must encapsulate a given action effectively so that they can be handled by the classification algorithms easily. The classification methods must be robust against corrupted or insufficient data. Sparse approximation techniques have found their wide use in signal and image processing applications as a powerful reconstruction tool. Sparse representation of a signal is a linear combination of few elements or atoms from a *dictionary*. Mathematically, it can be expressed as $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $\mathbf{y} \in \mathbf{R^m}$ is a signal of interest, $\mathbf{D} \in \mathbf{R^{m \times n}}$ is a *dictionary* and $\mathbf{x} \in \mathbf{R^n}$ is the sparse representation of $\mathbf{y}$ in $\mathbf{D}$. Typically $\mathbf{m} \ll \mathbf{n}$ resulting in an overcomplete or redundant *dictionary*. The *dictionary* is a collection of either predefined (eg. Wavelets, Fourier) or learned (from the data itself) atoms. The solution to the underdetermined system of equations $\mathbf{y} = \mathbf{D}\mathbf{x}$ can be found by either greedy algorithms or convex algorithms.

Primarily developed for robust reconstruction of signals, sparse representations are currently adopted in computer vision and pattern recognition problems where the goal is to find a meaningful representation besides being compact. Proper selection of dictionary for the purpose is essential. The structure of the dictionary has been transformed from standard basis or bases to more *intelligent* ones which learn from available training data. Learned or data-dependent dictionaries are apt for computer vision and pattern recognition problems. The learning procedure will train a dictionary to adapt to a particular class of data starting from a initial guess. The most popular algorithms for dictionary training are K-SVD [2] and Method of Optimal Directions (MOD)[3].

1.1. Theory

Sparse representations, to a large extent were successfully applied to face recognition [5], image classification [6], object detection [7] problems. These were first used for signal classification by Huang and Aviyente in [4] but was considered only as a reconstruction tool and the discriminative property was enhanced by Fisher's discrimination criteria. The fact that the sparse representations are discriminative by themselves was highlighted by Wright et al in their most celebrated paper [5]. An overcomplete dictionary was constructed with the training data: each column is the feature vector of one of the training samples. Under the assumption that adequate training data is available, a single test input can be represented as a linear combination of the atoms of the corresponding class only. Let $\mathbf{y}_i$ for $i = 1,\ 2,\ \ldots\ k$ be k training samples in $c = 1,\ 2,\ \ldots\ .\ m$ classes each and let $\mathbf{D}_m$ be the set of training samples in m^{th} class, then $\mathbf{D}_m = [\mathbf{y}_{1m}\ \mathbf{y}_{2m}\ \ldots\ \mathbf{y}_{km}]$. Including all m classes, the overall dictionary would be $\mathbf{D} = [\mathbf{D}_1\ \mathbf{D}_2\ \ldots\ \mathbf{D}_m]$. Any test input $\mathbf{y}_q$ can be classified by solving the underdetermined

*Neil M. Robertson is supported by EU FP7 LOCOBOT contract 260101.

†Paolo Favaro is Professor at Institute of Computer Science and Applied Mathematics, University of Bern

system of linear equations

$$\mathbf{y}_q = \mathbf{D}\,\mathbf{x}_q \tag{1}$$

where $\mathbf{x}_q$, the sparse representation of unknown $\mathbf{y}_q$, reveals the identity of $\mathbf{y}_q$. Convex optimization tools were used to compute the sparse solution and classification was made based on the *structure* of the sparse solution. The complete algorithm is named as *'Sparse representation for Classification'*(SRC) algorithm.

1.2. Related Work

Classification with sparse representation has been explored in depth in applications dealing with images, but this is an emerging trend with videos. This idea has recently evinced positive results for the problem of action recognition also [8],[9].

In [8], the authors proposed action recognition using sparse representation where the features(motion context descriptors) were computed for each frame in training and test videos and the classification is made by solving (1). For each frame in the test video the SRC algorithm is repeated and a label is given to that frame. The class of the test video is determined by the majority of the labels. The main drawback of this work, is that the proposed motion context descriptors are computationally intensive. Also, simultaneous sparse approximation algorithms would have been a better option instead of evaluating sparse solution for each frame in the sequence separately.

A different approach is used in [9]. Instead of global features as in [8], authors propose local features known as local motion pattern(LMP) descriptors. For each training video, LMP descriptors were extracted, dimensionally reduced by random-projection and a dictionary is trained using K-SVD algorithm. Given a test video, the sparse solution is computed over the trained dictionary for classification. The authors use Orthogonal Matching Pursuit(OMP) algorithm which is one of the *greedy methods*, to solve an underdetermined system for sparse solution. Dictionaries must be trained explicitly on the computed features before solving for the sparse solution. OMP can be substituted with a convex optimization algorithm to improve the performance.

1.3. Contributions

Having seen the different approaches of using sparse representations for recognition, we propose to solve the problem of action recognition using sparse representations with global features, random projection and convex optimization as the main tools. Our dictionary is formed by simple concatenation of the computed features and does not require additional training as in [9]. Hence, extending to new classes will be easy. We now explain our choice of tools used in our work.

Global Features: Among the variety of global and local features available in literature, we chose the former because they take into account the whole body rather than some keypoints and hence can capture all the detailed action. In our experiments we extract two types of features from silhouette-based (shape-based) Gait Energy Images and optical flow based Motion-descriptors further explained in Section 2.

Random projections: The computed features in images and videos are generally high-dimensional. We have to apply some dimensionality reduction methods like PCA in order to reduce the computational complexity while preserving the information encoded in the features. We use random projection for projecting the high-dimensional data onto a lower dimensional subspace [10]. According to popular *Johnson-Lindenstrauss* lemma, the distances between points in high-dimensional space are preserved when projected to points in much lower dimension using random projection.
We extract GEIs and motion-descriptors first from sequence of images for a given video sample and compute much lower dimensional features using Random projection.

Convex optimization: The exact problem of finding a sparse representation can be formulated as

$$\mathbf{min_x} \| \mathbf{x} \|_\mathbf{0} \text{ subject to } \mathbf{y} = \mathbf{Dx} \tag{2}$$

where $\mathbf{y}$, $\mathbf{x}$, $\mathbf{D}$ are defined in Section 1. This problem is NP-hard and are solved by either greedy methods which are basically approximation methods. Matching Pursuit(MP) and Orthogonal Matching Pursuit(OMP) are the two predominant greedy techniques. A more convenient way to represent the same problem is to replace the l_0 norm in (2) to its nearest convex function which is l_1 norm. The modified formulation

$$\mathbf{min_x} \| \mathbf{x} \|_\mathbf{1} \text{ subject to } \mathbf{y} = \mathbf{Dx} \tag{3}$$

can be solved by convex optimization methods based on convex programming. An error tolerance is generally introduced for the discrepancy between $\mathbf{y}$ and $\mathbf{Dx}$. It was shown in [11] that both above formulations (2) and (3) result in same sparse solution. Availability of reliable estimation algorithms makes the optimization task more efficient with (3). We use the popular *l1-ls solver* with GEIs and *M-BP* algorithm from [19] with motion-descriptors for computing sparse solutions.

Our approach differs from SVMs(Support Vector Machines), where the training has to be repeated on the entire new dataset every time a new class is added to the training data.

2. EXPERIMENTS AND RESULTS

We performed experiments with two datasets. Below we give brief details on the datasets used, GEIs and motion-descriptors.

Weizmann Dataset: This dataset [15] is one of the standard datasets which is used to assess the performance of a new approach towards action recognition. This set consists of 90 videos of 10 actions performed by 9 actors. The actions include bend, jump, jump in place, jumping jack, run, skip, side-walk, walk, wave.

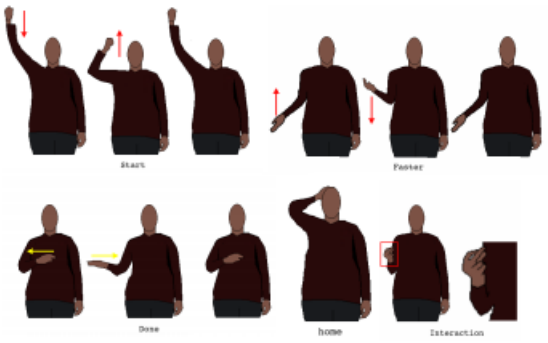

Fig. 1: Sample gestures of HWU dataset

HWU Dataset: This dataset was proposed by Baratteni et al for the control of industrial collaborative robots [12]. This is basically a gesture dataset consisting 10 gestures performed by 5 actors. The gestures in the dataset are done, faster, follow me, home, identification, interaction, ok, slower, start, stop. The data is recorded by Kinect. The recordings are available in 3 forms (i) Skeleton (ii) Depth frames and (iii) RGB frames. We used RGB images of the recordings in our experiments. Figure 1 shows the details of this dataset.

Gait Energy Images: Gait energy images(GEIs) were first

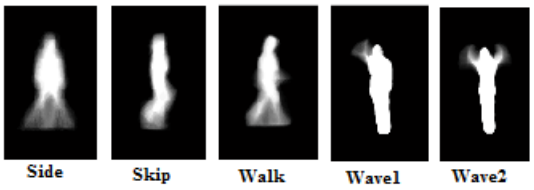

Fig. 2: GEIs for some actions in Weizmann dataset [13]

proposed by Han and Bhanu in [14] for individual identification. These are obtained by averaging the binary silhouettes of entire frames in a particular sequence. Sample GEIs for Weizmann dataset are shown in Figure 2.

Motion-descriptors: These are optical flow based fea-

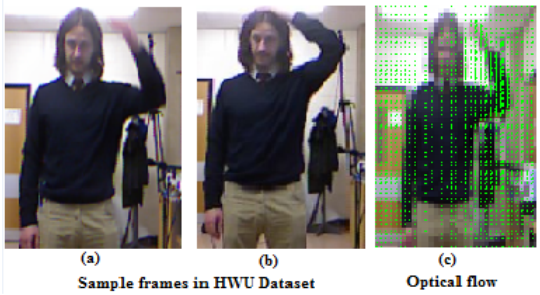

Fig. 3: Optical flow computed for sample two frames in HWU dataset

tures first proposed by Efros et al in [16]. We extract same spatio-temporal motion-descriptors for two of our experiments. Optical flow is computed between two frames, the obtained vector field is split into horizontal and vertical directions, each direction is half-wave rectified to obtain four non-negative channels. These channels are blurred with a Gaussian filter. Figure 3 shows the optical flow computed between sample frames in HWU dataset.

2.1. Experiments

Feature Extraction:
In our experiments we used previously extracted GEIs from [13] for Weizmann dataset and compute features by downsampling and random-projection to a lower dimension.Each GEI was approximately of size 120 x 80.
For extracting spatio-temporal motion descriptors to be used in our experiments, first, using mean-shift tracker from [17], figure-centric frame sequences are obtained for all the data. The size of each frame was set to 60x40. Optical flow is computed between two frames with a temporal extent of 7 frames for both datasets using the default algorithm from [18]. The size of descriptor for each frame is 4 times the size of frame due to resulting four channels. For both datasets, the size of each feature vector was 9600. Random projection is used to reduce the dimension to 256. The entries in the random matrix are normal distributed with zero mean.
In all experiments, testing is done based on leave-one-out cross validation strategy and all results are compared with that of SVM classifier with linear kernel.

Random projection vs Downsampling: Our first experi-

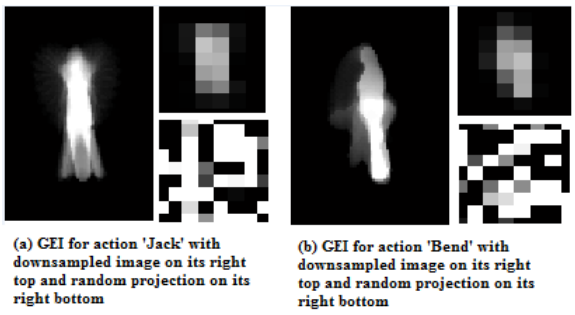

Fig. 4: Example of Downsampled and Random projected GEIs

ment highlights the significance of Random projection as the dimensionality reduction tool. For this we take GEIs for all the videos in the Weizmann dataset. We reduced the dimensions by (a)simple downsampling using matlab function and (b)random projection, to form a dictionary of 72 x 80 excluding one set of samples for testing. Figure 4 shows the sample GEIs with its downsampled and random-projected representations. Sparse representation is computed using *l1-ls* solver. The recognition rate with (a) was 75.49% while with (b) was **78.79%**. There is an improvement in the recognition rate with Random projection, though much behind than the state-of-the-art. This is because with each GEI as feature vector, the dimension of the feature vector is much higher than the

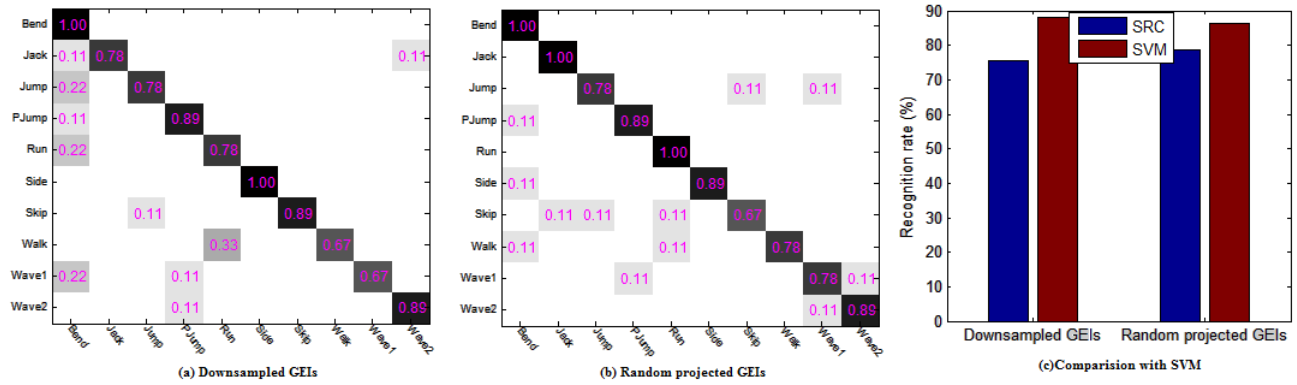

Fig. 5: Results from Random projection vs Downsampling experiment

number of training samples. Results are shown in Figure 5.

GEIs vs Motion-descriptors: Next we compare the shape-based features and flow-based features to see how the choice of features accounts for recognition rate. For the Weizmann dataset, we use the random projected motion-descriptors as features. Each video sample now is represented by a set of these features. Since the length of each video sample is different due to different action types and repetitions, the number of descriptors for every video is not same. We considered first 14 descriptors only for each video in dictionary construction resulting in a dictionary of size 256 x 1120. Unlike the previous experiment, a video sample is represented with a set of vectors(matrix). So we used simultaneous sparsity algorithm(M-BP) to solve the problem. Confusion matrices are shown in figures for the two experiments.The obtained recognition rate was **100%** which can be compared to that of GEIs.

SRC vs SVM: Finally, we compare our approach i.e. action recognition using SRC with SVM using random-projected motion-descriptors from both datasets. The dictionary for HWU dataset is formed in a similar fashion as explained above for Weizmann dataset. The size of this dictionary is 256 x 560. The recognition rates on HWU dataset was **98%** and 88.88% with our approach and SVM respectively. The results are shown in Figure 6.

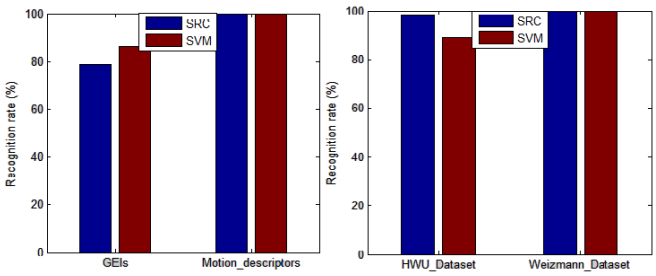

Fig. 6: Results from GEIs vs Motion-descriptors and SRC vs SVM experiments

3. CONCLUSION

From the experiments conducted we observe that main points. First, random projections can be used to obtain reduced dimension feature descriptors. Second, the flow-based features are more suitable for action recognition than shape-based features. Third, sparse representations can be successfully applied for the problem of action or gesture recognition if we have well structured dictionary.

In this paper, we proposed a classification strategy for action recognition based on sparse representation. Experimentally we have shown that the classification can be simply obtained by computing sparse representation of the test data without explicit training of the dictionary or applying any computationally intensive dimension reduction methods as in previous works. This being an initiation, we would like to explore further with corrupted data with *cross and bouquet* structure of the dictionary as in [5] which actually is an extension to the proposed dictionary structure.

Acknowledgements: We would like to thank Tenika Whytock for sharing GEIs data and SVM code.

4. REFERENCES

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[2] M. Elad and M. Aharon, "Image denoising via learned dictionaries and sparse representation," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 895–900.

[3] K. Engan, S.O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, vol. 5, pp. 2443 –2446 vol.5.

[4] K. Huang and S. Aviyente, "Sparse representation for signal classification," *Advances in neural information processing systems*, vol. 19, pp. 609, 2007.

[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[6] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries," Tech. Rep., DTIC Document, 2008.

[7] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," *Computer VisionECCV 2002*, pp. 97–101, 2006.

[8] C. Liu, Y. Yang, and Y. Chen, "Human action recognition using sparse representation," in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*. IEEE, 2009, vol. 4, pp. 184–188.

[9] T. Guha and R.K. Ward, "Learning sparse representations for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 8, pp. 1576–1588, 2012.

[10] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[11] D.L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[12] P. Barattini, C. Morand, and N.M. Robertson, "A proposed gesture set for the control of industrial collaborative robots," in *RO-MAN, 2012 IEEE*, sept. 2012, pp. 132 –137.

[13] T. Whytock, A. Belyaev, and N. Robertson, "GEI + HOG for action recognition - presented at the 2012 4th UK Computer Vision Student Workshop (British Machine Vision Workshop)," .

[14] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 316–322, 2006.

[15] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[16] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 726–733.

[17] R.T. Collins, "Mean-shift blob tracking through scale space," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. IEEE, 2003, vol. 2, pp. II–234.

[18] D. Sun, S. Roth, and M.J. Black, "Secrets of optical flow estimation and their principles," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2432–2439.

[19] A. Rakotomamonjy, "Algorithms for multiple basis pursuit denoising," in *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.