Compressive Sensing of Time Series for Human Action Recognition

Oscar Perez Concha, Richard Yi Da Xu, Massimo Piccardi
School of Computing and Communications
University of Technology, Sydney (UTS)
PO Box 123 Broadway, NSW 2007, Australia
Oscar.PerezConcha, YiDa.Xu, Massimo.Piccardi@uts.edu.au

Abstract

Compressive Sensing (CS) is an emerging signal processing technique where a sparse signal is reconstructed from a small set of random projections. In the recent literature, CS techniques have demonstrated promising results for signal compression and reconstruction [9, 8, 1]. However, their potential as dimensionality reduction techniques for time series has not been significantly explored to date. To this aim, this work investigates the suitability of compressive-sensed time series in an application of human action recognition. In the paper, results from several experiments are presented: (1) in a first set of experiments, the time series are transformed into the CS domain and fed into a hidden Markov model (HMM) for action recognition; (2) in a second set of experiments, the time series are explicitly reconstructed after CS compression and then used for recognition; (3) in the third set of experiments, the time series are compressed by a hybrid CS-Haar basis prior to input into HMM; (4) in the fourth set, the time series are reconstructed from the hybrid CS-Haar basis and used for recognition. We further compare these approaches with alternative techniques such as sub-sampling and filtering. Results from our experiments show unequivocally that the application of CS does not degrade the recognition accuracy; rather, it often increases it. This proves that CS can provide a desirable form of dimensionality reduction in pattern recognition over time series.

1. Introduction

Human action recognition from camera videos has been one of the most popular research topics within the computer vision and pattern recognition communities in recent years, with applications to recognition of primitive actions, sport actions, human-computer interaction, movie annotation, and others [16, 21, 23, 15, 13, 14].

Recently, we have witnessed a continuous increase in the size of typical feature sets in an attempt at providing "richer" descriptions of the actions which, in turn, could lead to improved recognition accuracy. For example, in [12], the author proposes a spatio-temporal feature detector and uses histograms of the extracted features as the feature set for activity recognition. A similar idea is also proposed in [22] where the Speeded Up Robust Features (SURF) [2] are used to construct local spatio-temporal features. The volumetric space-time shapes of [10] are also very rich in attributes. On the other hand, one may want to collect such feature sets over very short time spans in order to capture the action dynamics at finer levels of detail. The combined effect of large feature sets and high sampling frequencies leads to "fat" representations of an action instance, potentially in the order of tens or hundreds of kilobytes each. While this is not an issue in isolation, it may become such for processing, storage and communication, especially in scenarios such as the rapidly widespreading large camera networks where the information extracted from each camera is communicated across the network to infer across-camera knowledge [3].

At the same time, the novel field of Compressive Sensing (CS) techniques has provided a new approach to the compression and reconstruction of signals at a rate significantly below that of Nyquist sampling and have hence raised much attention in the signal processing community [9, 8, 1]. The compactness of the CS representation makes it a very appealing technique also for the compression of time series in distributed pattern recognition applications. Apart from offering good compression capability, the relevance of CS to pattern recognition lies in its potential as a dimensionality reduction technique for series of sampled signals. Differently from techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and many others [4], compressive sensing is not learned from a training set and therefore does not suffer from limited generalisation. On the other hand, other conventional compression techniques such as the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT) do not provide a fixed, lower-dimensional feature set suitable for pattern recognition since the positions of the non-negligible coefficients occur at different locations for each compressed sample. In this paper we show that not only compressive sensing of time series does not deteriorate recognition accuracy, but it can actually lead to improved accuracy compared to recognition from the original time series.

The rest of this paper is organized as follows: in section 2, we describe the theoretical background of compressive sensing. In section 3, we present the principles of applying compressive sensing to multivariate time series. In section 4, we describe the action dataset and the feature set used for the experiments. In section 5, we present the results on four sets of experiments: 1) recognition in the compressive-sensed domain; 2) recognition from the reconstructed time series; 3) recognition in a hybrid compressed domain; 4) recognition from time series reconstructed after hybrid compression. Other techniques such as subsampling and averaging are also compared for a broader comparative analysis. In section 6, we summarise our results and presents the conclusions.

2. Background on Compressive Sensing

The general theory behind Compressive Sensing can be summarized as follows: let us assume that we are given a discrete signal, f, in $\Re^N$ and a $N \times N$ matrix, Ψ, whose columns are a set of orthogonal basis vectors. This matrix is called a sparsifying matrix since, when multiplied by signal f, it produces a representation $\omega = \Psi^T f$ which is a version of f in the Ψ domain. For many naturally occurring signals such as images and audio and an appropriate choice of Ψ, vector ω offers a sparse representation in the sense that only K of its elements have values significantly different from zero; the other N - K elements of ω are either zero or very close to zero. In such a case, f is said to be K-sparse. Typical choices for Ψ are the discrete cosine transform (DCT) matrix and the wavelet basis matrix, both widely used in image compression applications.

In addition to Ψ, a linear measurement matrix or *sampling matrix* Φ of size $M \times N$ is also introduced, with $M \ll N$ and M only marginally larger than K such that:

$$v = \Phi f = \Phi\Psi\omega = \Theta\omega \tag{1}$$

Vector v contains the measurements in $\Re^M$ which we can access directly. When Θ satisfies the so-called *restricted isometry property* (RIP) [9, 8], we can reconstruct ω (hence f) given v exactly. A thorough verification of RIP for Θ is prohibitive and requires computations in every $M \times K$ submatrix of Θ, which involves $\binom{N}{K}$ combinations [1]. Yet, it can be shown that if Φ is chosen to be a Gaussian random matrix and condition $M \geq cKlog(N/K)$ is satisfied for some constant c, then it is highly probable that we will be able to reconstruct ω exactly [9].

From (1), it is also immediate that matrix Φ (and hence Θ) has more columns than rows. Therefore, ω lies in a solution space and cannot be reconstructed uniquely. However, thanks to the sparsity assumption, we can recover signal $\hat{\omega}$ by minimizing its l_1 norm [8] as follows:

$$\hat{\omega} = \arg\min_{\omega}(\|\omega\|_1), \quad \textit{s.t.} \quad v = \Theta\omega \tag{2}$$

where notation $\|.\|_p$ defines the l_p norm. Several other reconstruction algorithms have been presented in recent years including Denoising Basis Pursuit and the Lasso [7]. In addition to the estimation of the underlying signal $\hat{\omega}$, recent works such as [18] have also attempted to incorporate CS reconstruction in a Bayesian framework where a full posterior for $\hat{\omega}$ is to be estimated. For example, in [18] the authors assume the likelihood function to be a Gaussian distribution:

$$p(v|\omega, \Sigma) \sim \mathcal{N}(\Phi\omega, \Sigma) \tag{3}$$

and the prior distribution of ω to be a product of Gaussians on its components, ω_i:

$$p(\omega|\alpha) = \prod_{i=1}^{N} \mathcal{N}(\omega_i|0, \alpha_i^{-1}). \tag{4}$$

By using the solution offered by the Relevance Vector Machine (RVM) framework [20] and carefully defining distributions for all hyper-parameters involved [18], it is possible to derive the solution for the full posterior $p(\omega|v)$.

3. Compressive sensing of multivariate time series

In general terms, we can assume that a K-dimensional feature set can be extracted from each frame of a video depicting an action. Such a feature set can be as varied in nature as the actor's pixel map, a set of shape descriptors, histograms of special interest points, or others. We note the sequence of the extracted feature sets as $\mathbf{O}_{1:T} = \{o_1, ..., o_t, ..., o_T\}$, with each o_t being a K-variate random variable, commonly continuous in value. The main idea for applying compressive sensing to such a time series is to partition it into contiguous windows and compress the samples in each window by way of a sampling matrix. The sampling matrix, Φ, is $M \times N$ in size, with $M << N$, and its application over each window transforms N univariate samples into a single, M-variate sample. Each feature in the feature set is compressed independently of the others. As a result, an original action sequence with T frames and K dimensions, i.e.,

$$\mathbf{O}_{1:T}(1:K) = \left| \begin{array}{ccc} o_1(1) & ... & o_T(1) \\ ... & ... & ... \\ o_1(K) & ... & o_T(K) \end{array} \right|.$$

is transformed into a new sequence with T/N frames and $K * M$ dimensions:

$$\mathbf{O}'_{1:\frac{T}{N}}(1:K{\times}M) = \left| \begin{array}{ccc} o'_1(1) & ... & o'_{\frac{T}{N}}(1) \\ ... & ... & ... \\ o_1(K \times M) & ... & o_{\frac{T}{N}}(K \times M) \end{array} \right|.$$

In other words, a reduction in the length of the time series by a factor N is obtained at the expense of an increase in the feature set dimensionality of a factor M. The overall size of the time series decreases from $K{*}T$ to $K{*}T{*}M/N$. The intrinsic, overall dimensionality of the time series may or may not vary as a consequence of this manipulation; in general, it is not obvious how to assess the extent of the conditional dependencies between time samples and estimate the intrinsic dimensionality along the time dimension [5]. In the distributed cameras scenario described in the Introduction, the time series are computed at the local camera level, communicated in compressed form, and recognition then performed at the received end, either from the compressed time series directly or after reconstruction via inverse algorithms.

4. Dataset and feature set

For all experiments described in this paper we have used the Weizmann dataset[1], a simple dataset depicting 10 different primitive actions performed by 9 different subjects. Figure 1 shows examples of an actor's foreground masks in the Weizmann dataset. In the near future, we plan to extend our results to other popular datasets such as KTH and MuHAVi [19, 11] where segmentation is probing. As feature set, we decided to use a set of five region centroids extracted from the foreground mask, hoping that they would prove sufficiently action-discriminative. While more sophisticated articulated motion models could be fit on the foreground masks, the chosen feature set is adequate for comparison and for proving the point of this paper. Figure 2 shows examples of the feature set.

To extract the feature set, we model the location x_i of every pixel in the foreground mask by using a Gaussian mixture model (GMM), where each x_i is given probability density:

$$p(x_i|\alpha, \mu, \mathbf{\Sigma}) = \sum_{l=1}^{L} \alpha_l \mathcal{N}(x_i; \mu_l, \Sigma_l). \quad (5)$$

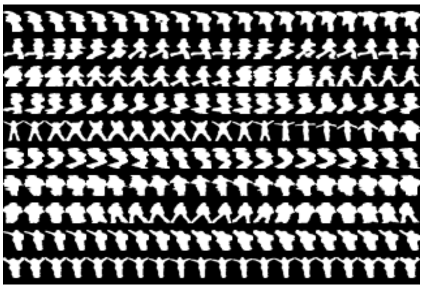


Figure 1. Examples of foreground masks from the Weizmann dataset (reduced to 16 × 16 for illustration). From up to bottom, actions are 'Bend', 'Run', 'Walk', 'Skip', 'Jumping Jack', 'Jump Forward On Two Legs', 'Jump In Place On Two Legs', 'Gallop Sideways', 'Wave With Two Hands' and 'Wave With One Hand' for one of the actors ('daria').

where $\alpha = \{\alpha_1, ...\alpha_l, ...\alpha_L\}$ is the set of the components' priors; $\mu = \{\mu_1, ...\mu_l, ...\mu_L\}$ are the components' means, with each μ_l a two-dimensional mean vector; and $\mathbf{\Sigma} = \{\Sigma_1, ...\Sigma_l, ...\Sigma_L\}$ are the components' covariances, with each Σ_l a 2×2 covariance matrix. For the purpose of this paper, we set the number of Gaussian components L to five. Future work will address training of an optimal L to give the best discriminative results for the entire dataset.

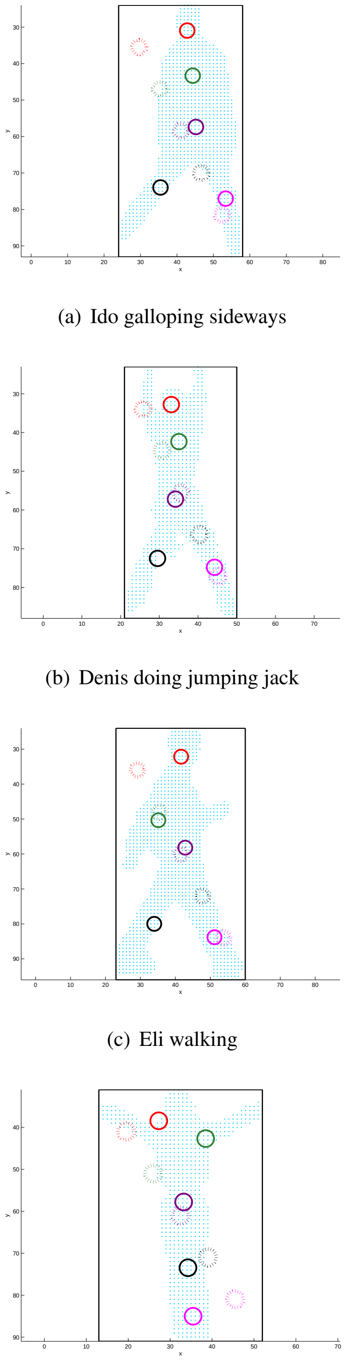
The initial set of parameters, $\alpha^{(0)}, \mu^{(0)}, \Sigma^{(0)}$, is obtained in a heuristic way: we first obtain a bounding box containing the foreground mask. Then, the initial set of means $\mu^{(0)}$ is obtained from equally-spaced positions along the diagonal of the bounding box, shown as the dotted circles in Figure 2. The initial covariance matrices set, $\mathbf{\Sigma}^{(0)}$, is chosen as: $\Sigma_1^{(0)} = \Sigma_2^{(0)} = ... = \Sigma_L^{(0)} = \sqrt{A_{bound}/L} \times \mathcal{N}_{2\times2}$, where A_{bound} is the area of the bounding box, and $\mathbf{I}$ is the identity matrix.

After the application of an expectation-maximization (EM) training algorithm, the set of parameters converges to $\alpha^{(f)}$, $\mu^{(f)}$ and $\mathbf{\Sigma}^{(f)}$. The final set of means, $\mu^{(f)}$, is also plotted in Figure 2 as solid circles. Finally, $\mu^{(f)}$ is then used as the feature set for every video frame, with a dimensionality of $K = dim(o_t) = 2 \times L = 10$ in our case.

5. Experiments

In order to test the effectiveness of compressive-sensed time series for human action recognition, we have conducted several comparative experiments. As time-series classifier, we have used the hidden Markov model for its flexibility and easy application [17]. A hidden Markov model is a latent state model defined by its set of parameters, λ, which provides a density value, $p(\mathbf{O}_{1:T}|\lambda)$, for

[1] http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html



(a) Ido galloping sideways

(b) Denis doing jumping jack

(c) Eli walking

(d) Ira waving with her two hands

Figure 2. The initial and final set of means are plotted, where the corresponding means are drawn with the same color. Original means, in the principal diagonal, are displayed by dotted circles, whereas the final means are displayed by solid circles

time series $\mathbf{O}_{1:T}$. To prepare the classifier, a hidden Markov model, λ_c, is first trained for each of the C classes of interest, $c = 1..C$, from examples from the class. After training of the class models, maximum-likelihood classification for time series $\mathbf{O}_{1:T}$ is simply provided by:

$$c_{ML} = \underset{c}{\operatorname{argmax}}(p(O_{1:T}|\lambda_c)), \quad c = 1..C. \qquad (6)$$

We carried out a *leave-one-actor-out* cross-validation so that the same actor will not be used for training and validation. Every actor in turn is used for validation. With the purpose of making the comparisons consistent, in all the experiments we have set the number of hidden states in the HMM to five, and each emission probability was modelled as a Gaussian mixture model with two components. In addition, the same set of initial HMM parameters was used throughout the experiments. The techniques we compared are divided into the following categories: 1) compressive sensing of the original time series; 2) averaging or sub-sampling of the original time series; 3) reconstruction of the signals from the compressed domain; 4) *hybrid* compression of the original time series by way of a mixed Haar/compressive sensing compression. Results are compared with recognition accuracy from the original signal. Each of the following sub-sections describes a category of techniques.

5.1. Compressed domain

For compressing the original time series, we have used an $M \times N$ sampling matrix, Φ, generated from a Gaussian process which is likely to ensure the required restricted isometry property [9]. Although not reported in the paper, we have experimented with several instances of Φ and generally reported results delivering the highest accuracy. Two cases were tested in this experiment, where the first one was set to have $M = 4$ and the second one $M = 2$, for a compression ratio of 50% and 25%, respectively. The size of the time window, N, like for all other experiments in this work, was set to 8. This experiment is labelled as *number 1*.

5.2. Averaging and sub-sampling

Another set of experiments was accomplished to compress the time series by deterministic techniques: averaging over the N samples or sub-sampling from them. These two processes were applied in two ways: a) over the original signal (experiments labelled as *numbers 2 and 3*) or b) over the sparse version of the original signal (experiments labelled as *numbers 4 and 5*). For the latter, we used the Haar wavelet as the $N \times N$ orthonormal basis. In the case of sub-sampling, we also tried to sub-sample at even and odd indices in the window. For all combinations, the number of retained samples were set to both $M = 4$ and $M = 2$.

Averaging and sub-sampling, too, can be represented as a compression matrix, with the only difference that Φ is no longer a random matrix, but instead set to take fixed values. For conciseness, we describe such values as:

1. Sub-sampling:
   - for M = 2:
     $\Phi_o = [1 \quad 0 \quad 0 \quad 0]$, $\Phi_e = [0 \quad 1 \quad 0 \quad 0]$
   - for M = 4:
     $\Phi_o = [1 \quad 0]$, $\Phi_e = [0 \quad 1]$
2. Averaging:
   - for M = 2:
     $\Phi = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$
   - for M = 4:
     $\Phi = [0.5 \quad 0.5]$

The averaged or sub-sampled signals were then fed into the classifier. Therefore, the classification was directly carried out in the compressed domain.

5.3. Reconstruction

In the next set of experiments, the input to the classifier was obtained by reconstructing the time series after CS compression. Two approaches were used:

- we first converted the time series into a sparse representation by applying a Haar matrix, Ψ, and then compressed it by applying Φ. The subsequent reconstruction is achieved by minimizing the l_1 norm of the solution by a Lasso-type algorithm. We labelled this experiment as *number 6*.
- in addition, we directly applied Φ to the original signal. The reconstruction step follows Eq. (2). We labelled this experiment as *number 7*.

The reconstructed signals, which have the same dimensionality of the original ones, were so fed into the classifier.

5.4. Hybrid compression

In the last set of experiments, we applied *hybrid* CS, a technique which compresses the signal by a combination of sparse coefficients and compressive-sensed coefficients. In this approach, the original time series are first converted into the sparse basis; then, a small amount, n, of low-order coefficients is retained unchanged; eventually, the remaining $N - n$ coefficients are compressed by compressive sensing. The rationale for this approach is given by the empirical observation that the coarse-scale wavelet coefficients are important for effective signal reconstruction [6]. For the reconstruction process, we proceeded in an inverse way, that is, reconstructing the compressed part, or highest frequencies of the signal, with the Lasso algorithm, concatenating the solution with the n unaltered low-order sparse coefficients, and then applying Ψ^T to reconstruct the signal in its original domain. The selected window was also $N = 8$ frames.

For this experiment, we considered two cases: 1) retaining $n = 1$ low-order coefficients and compressing the remaining $N - n = 7$ into 3 compressed samples and 2) retaining $n = 2$ low-order coefficients and compressing the remaining $N - n = 6$ into 2 compressed samples, for an overall compression ratio of 50% in both cases. This experiment is labelled as *number 8*. Eventually, we tested retaining $n = 1$ low-order coefficients and compressing the remaining $N - n = 7$ into 1 compressed sample, for an overall compression ratio of 25% (experiment *number 9*). In experiment *number 8*, case $n = 3$, we also attempted explicit reconstruction of the time series from its hybrid representation.

5.5. Results

The results for all the experiments are displayed in Table 1. The second column of the table indicates if the processing was carried out over the original signal (O) or its sparse version (S) obtained by the application of a Haar transform. We report accuracies as the best out of several trials to mollify the risk that they depend on unfortunate parametrisation of the experiment.

The immediate, stunning result from the table is that recognition with two different flavours of compression achieved an accuracy of 95.5% compared to an accuracy of 77.8% from the original data. This shows at once that the conversion of some time complexity into feature complexity seems to actually achieve a regularisation in the actual, intrinsic dimensionality of the time series. Namely, the equal-highest accuracies are achieved by compressive sensing and by sub-sampling of the Haar coefficients. In the following, we address detailed comments to the individual experiments.

For recognition directly *in the CS compressed domain* (experiment *number 1*), we obtained accuracies of 95.5% and 92.2% at a compression rate of 8×2 and 8×4, respectively. Such accuracies were the highest across the entire range of compared techniques. Surprisingly, the accuracy proved even higher (by 3.3%) for the more reduced size, 8×2. For this reason, we report a further analysis of the impact of the size of the CS compressed data, M, later in this section.

Recognition by *averaging and sub-sampling the original signal* (experiments *number 2* and *number 3*) mildly improved the accuracy compared to recognition from the original data, with values ranging between 74.4% and 83.3% (compared to 77.8% of the original data). Sub-sampling achieved slightly better results than averaging: however, its results seem to significantly depend on the arbitrary choice

Experiment #	O (original) / S (sparse)	Technique	Compression	Accuracy (%)
0	O	Original	N/A	**77.7**
1	O	CS	**8x4** **8x2**	**92.2** **95.5**
2	O	Averaging	8x4 8x2	81.1 74.4
3	O	Sub-sampling	8x4, Φ_o 8x4, Φ_e 8x2, Φ_o 8x2, Φ_e	83.3 78.8 83.3 76.6
4	S	Averaging	8x4 **8x2**	82.2 **90.0**
5	S	Sub-sampling	8x4, Φ_o 8x4, Φ_e 8x2, Φ_o 8x2, Φ_e	63.3 **95.5** 42.2 **95.5**
6	S	Reconst. CS	8x4 8x2	71.1 61.1
7	O	Reconst. CS	8x4 **8x2**	90.0 **95.5**
8	S	Hybrid CS	6x2 7x1	78.8 71.1
9	S	Hybrid Reconst.	7x3	73.3

Table 1. Accuracy (%) for the 9 experiments carried out: compressive sensing over the original signal (experiment 1), averaging and sub-sampling over the original and sparse signal (experiments 2-5), reconstructed CS signal (experiments 6 and 7) and hybrid CS and hybrid CS-reconstruction (experiments 8 and 9). The second column shows if the compression has been carried over the original (O) or the sparse (S) signal. In the case of experiments *number 3 and 5*, the two values in the accuracy column show the results when the first and fifth (result on the left) and second and sixth (result on the right) coefficients are sampled over a window of 8 frames

of which samples are retained: by retaining samples at odd indices in the window, accuracy reached 83.3%; by retaining samples at even indices, accuracy ranged between 76.6%-78.8%. Such a dependence is certainly not desirable as it makes hard to generalise results.

However, recognition by *averaging and sub-sampling the Haar coefficients* (experiments *number 4* and *number 5*) reported much higher accuracies in some cases and a top accuracy equal to that of compressive sensing. These results call for more discussion due to the variations between different cases. The average of neighbouring Haar coefficients tries to retain an average value between coefficients of similar order; for instance, in the 8×2 case, the lowest four coefficients are averaged into one and so are the highest four coefficients. This simple approach appears effective as it reaches accuracies of 82.2%-90.0%. However, such values are lower than those obtained with the CS transform. On the other hand, sub-sampling selects a sub-set of Haar coefficients to retain and achieves even more pronounced accuracies, up to 95.5% on a par with CS. The downside of this approach is that the accuracy drastically depends on which coefficients are arbitrarily retained, reaching worst scores as low as 42.2%.

In the next set of experiments, we performed recognition from the *explicitly reconstructed series after CS compression* (experiments *number 6* and *number 7*). The most interesting outcome of this experiment is the remarkable difference between compressive sensing from the original signal, $v = \Phi f$, and compressive sensing from its sparse representation $v = \Phi \Psi^T f = \Phi \omega$. In the theory of compressive sensing these two approaches are equivalent, as they both can lead to the reconstruction of ω via constrained L1-norm minimization. However, in practice the choice has an impact on the reconstruction results. Accuracy results from reconstruction from the sparse basis were significantly worse (with accuracy as low as 61.1%). Instead, accuracy from reconstruction from the original signals were almost equivalent to those achieved by operating directly in the CS compressed domain (90.0%-95.5%).

Eventually, the *hybrid CS/Haar experiments* (*number 8* and *9*) did not report noteworthy accuracy, in the range of or worse than the recognition accuracy from the original time series.

Given that results with compressive sensing and sub-sampling of Haar coefficients led to equal best accuracy, we conducted a further experiment to explore the stability of these results over the parameters. For compressive sensing (experiment *number 1*), we computed the accuracy at the increase of the number of compressed samples, from 1 to 10. Please note that when the number of compressed samples becomes greater or equal than the length of the time window, $N = 8$, the signal can always be reconstructed without error. However, the original signal is not conducive to the highest accuracies, and the test is therefore worthwhile. Figure 3 shows the accuracy we obtained after applying consecutive CS transformation to the original signal, from $\Phi_{8\times1}$ to $\Phi_{8\times10}$, including $\Phi_{8\times8}$. We can draw several comments from this experiment: the first is that we unexpectedly achieved the best result with only $M = 2$ samples,

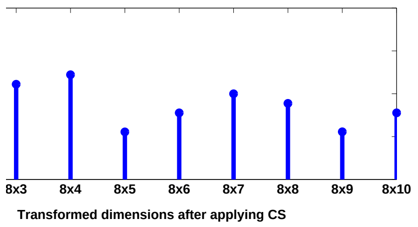


Figure 3. Accuracy after applying consecutive CS transformations to the original signal

for a very favourable compression ratio of 25%. The second is that the accuracy with $M = 8$ samples is higher by 11% (88.8%) than that based on the original signal. In this case, the size of the two time series is identical, with the only difference that in the CS case we have converted a certain extent of time complexity into an equivalent extent of feature set complexity. The better performance of CS may be conclusive evidence that CS performs an implicit, desirable data regularisation which has invariably led to higher recognition accuracy than when recognition is attempted directly from the original signal.

Table 2 shows expanded results for the experiment where the feature set is built by sub-sampling the Haar coefficients (experiment *number 5*). The highest accuracy (95.5%) is obtained when the second and sixth frequencies are kept. However, results vary drastically for different choices of the two retained coefficients, down to 63.3%. If all coefficients are kept - for no reduction in size of the time series - accuracy is still only 80%. This proves that this technique is hard to tune and results not obviously generalisable.

Experiment #	O (original) / S (sparse)	Technique	Compression	ACCURACY (%)
0	O	Original	8x8	**77.7**
5	S	Sub-sampling	8x2 - (1,5)	63.3
			8x2 - (2,6)	95.5
			8x2 - (3,7)	86.6
			8x2 - (4,8)	82.2
			8x8 - (all)	80.0

Table 2. Sub-sampling technique when only two sparse coefficients (out of 8) are kept, starting from the first, second, third and fourth coefficient respectively. The last row refers to retaining all coefficients.

6. Conclusions

This paper has presented a comparison of several methods to compress time series from instances of human actions for action recognition, with special emphasis on the recently proposed Compressive Sensing (CS) techniques. The comparison includes compressing by CS, Haar transforms, hybrid CS-Haar, averaging and sub-sampling, and performing recognition either directly in the compressed domain or over the reconstructed signals. These techniques are flexible and can be applied to any type of features other than those used for this paper's experiments. As time-series classifier, we have used the well-known hidden Markov model with Gaussian mixture outputs.

The main and somehow exciting result stemming from the comparison is that the accuracy of action recognition was improved by the application of compressive sensing to the original time series (95.5% accuracy vs. 77.7%). This means that compressive sensing not only offers the opportunity to significantly reduce the overall size of the time series (down to 25% in this work), but also operates some form of desirable dimensionality reduction which facilitates the recognition of patterns. Compressive sensing was originally proposed only for signal compression, under the unquestioned assumption that the original signal is the ideal target to reconstruct. Results reported in this papers seem to suggest that imperfectly reconstructed signals may enjoy other properties of benefit for pattern recognition. In the near future, we plan to expand this analysis to other datasets and features, and explore the theoretical intertwining between compressive sensing and recognition.

References

[1] R. G. Baraniuk. Compressive sensing. In *Lecture Notes in IEEE Signal Processing Magazine*, volume 24, pages 118–120, 2007.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006.

[3] N. Bellotto, E. Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, C. Fernández, L. Van Gool, and J. Gonzàlez. A Distributed Camera System for Multi-Resolution. 2010.

[4] C. M. Bishop, editor. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] F. Camastra. Data dimensionality estimation methods: A survey. *Pattern Recognition*, 36:2945–2954, 2003.

[6] E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23:969, 2007.

[7] E. J. Candes. Compressive sampling. In *International Congress of Mathematicians*, 2006.

[8] D. L. Donoho. Compressed sensing. In *IEEE Trans. Info. Theory*, volume 52, pages 1289–1306, 2006.

[9] J. R. E. Candes and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete fre-

quency information. In *IEEE Trans. Inform. Theory*, volume 52, pages 489–509, 2006.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[11] D. I. R. C. Kingston-University. The muhavi-mas database, http://dipersec.king.ac.uk/muhavi-mas/, 2008.

[12] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005.

[13] I. Laptev and T. Lindeberg. Space-time interest points. In *the 9th IEEE International Conference on Computer Vision, ICCV 2003*, volume 1, pages 432–439, 2003.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[15] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 863–876, 2006.

[16] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 2009.

[17] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. 77:257–286, 1989.

[18] Y. X. S. Ji and L. Carin. Bayesian compressive sensing. In *IEEE Trans. Signal Processing*, volume 56, pages 2346–2356, 2008.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, 2004.

[20] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. In *Journal of Machine Learning Research*, pages 2346–2356, 2001.

[21] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, September 2008.

[22] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *Computer Vision–ECCV 2008*, pages 650–663, 2008.

[23] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *Proc. Comp. Vis. and Pattern Rec*, pages 379–385, 1992.