Compressive Sensing of Time Series for Human Action Recognition

Oscar Perez Concha, Richard Yi Da Xu, Massimo Piccardi

iNEXT, UTS, Sydney, Australia
{oscarpc, ydxu, massimo@it.uts.edu.au}
In brief

• A short overview of
  – human action recognition
  – compressive sensing

• The proposed method

• Experimental results

• Discussion
Human action recognition

• Extract features from a video of a human subject and classify its action

• Fundamental assumptions:
  – human is (at least roughly) tracked
  – time segmentation (start/end times) is provided

• Not: action detection, online action recognition, scene recognition (a la Hollywood, Hollywood2 etc)
Well-known challenges

- High intra class variance, both in (feature) space and time
- Limited inter class distance for some classes
- Illumination & viewpoint changes, occlusions
Approaches

• **Bag-of-features approaches**
  – Bag-of-features + discriminative classifiers [Schuldt ICPR 2004], [Dollar VS-PETS 2005], [Laptev CVPR 2008] etc

• **Sequential approaches**
  – HMM + ML classification [papers since 1992 up to Ikizler IJCV 2008]
  – Action classification as state decoding in a CRF [S. Wang CVPR 2006], [L. Wang CVPR 2007]

• Others: DTW, Hausdorff distance etc

DICTA 2010, Sydney, December 2010
Sequential approaches

- Sequence of feature vectors (aka observations, measurements, emissions), one per frame:
  \[ O_{1:T} = \{ o_1, \ldots, o_t, \ldots, o_T \} \]

- Each \( o_t \) is a continuous, multivariate random variable of \( K \) dimensionality

- Example of sequential approach: the hidden Markov model, a generative model for \( p(O_{1:T}) \)
Hidden Markov model

- First-order Markov dynamics, mutual independence of observations given their states:

\[
p(O_{1:T}) = \sum_{Q_{1:T}} p(O_{1:T} \cdot Q_{1:T}) = \sum_{Q_{1:T}} \left( p(q_1) \prod_{t=2}^{T} p(q_t \mid q_{t-1}) \prod_{t=1}^{T} p(o_t \mid q_t) \right)
\]

- Each \( q_t \) (hidden state) is discrete and takes values in \( S = \{s_1, \ldots, s_N\} \) symbolic dictionary

- Time-independent (aka stationary, time-homogeneous)
Compressive sensing

• Co-credited to [Candès & Tao 2006], [Donoho 2006]; related work from many other authors

• In plain terms, a signal $f$ in $\mathbb{R}^N$ enjoying a sparse representation, $\omega$, $f = \Psi \omega$

can be effectively sampled by a sampling matrix, $\Phi$, whose number of rows, $M$, is significantly smaller than $N$
$v = \Phi f = \Phi \Psi \omega$
CS recovery algorithms

• A recovery algorithm estimates \( \tilde{\omega} \) from \( \nu \)

• The ideal objective of CS recovery algorithms is to recover \( \tilde{\omega} \) with the smallest number of coefficients different from zero:

\[
\tilde{\omega} = \arg \min \|\omega\|_{L_0} \quad s.t. \quad \nu = \Phi \Psi \omega
\]

• In many cases*, the same value can be obtained from:

\[
\tilde{\omega} = \arg \min \|\omega\|_{L_1} \quad s.t. \quad \nu = \Phi \Psi \omega
\]

* D. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Comm. On Pure and Applied Math, 59(6):797–829, 2006
Conditions on $\Phi$

- $\Phi$ should be *incoherent* with $\Psi$ i.e. small correlation between elements of $\Phi$ and $\Psi$ [Donoho & Huo 2001]

- To be robust against additive noise, $\Phi$ should possess the *restricted isometry property* i.e. the $L_2$ norm of $\Phi f$ be close to that of $f$ [Candès & Tao 2005]

- A *Gaussian random matrix* (a matrix whose elements are sampled from $N(0, \sigma^2)$) with a “big enough” $M$ satisfies these conditions
Example

\[ \Phi \quad \Psi \text{ (DCT)} \]
The idea - 1

- Take each sequence of feature vectors, $O_{1:T}$, and partition it into short segments of $N$ (e.g. 8) frames

NB: each dimension of $o_t$ is considered separately
The idea - 2

• Compress each segment by an $M \times N \Phi$

\[
\begin{align*}
\Phi &\uparrow \\
\uparrow & \quad \nu \\
\end{align*}
\]
Example of recovery

\[ e_{L_1} = \frac{\| f - \tilde{f}_{L_1} \|_{L_2}}{\| f \|_{L_2}} = 0.1869 \quad e_{L_2} = \frac{\| f - \tilde{f}_{L_2} \|_{L_2}}{\| f \|_{L_2}} = 0.8276 \]
The idea – 3

• We repeat this for all $K$ dimensions and concatenate all the resulting compressed values

• In this way, we replace a sequence of $N$, $K$-dimensional feature vectors with a single, $K*M$-dimensional feature vector

• The sequentiality of the observations is somehow preserved

• The dimensionality of the feature space increases remarkably
Motivation

• Original motivation: compress the sequences of feature vectors by a simple, non-adaptive operation

• A compact representation for the action which can save storage space and bandwidth

• No signal recovery is required for analysis, and classification is typically faster

• A-posteriori motivation: seemingly, a gain in recognition accuracy
Dataset, feature vector

• The dataset is Weizmann, eased by the availability of the actor’s silhouettes

• The feature vector are the centers of 5 2D Gaussian components fit on the binary pixels
Compared techniques - 1

0. Original signal
1. Compressive sensing (CS) domain
2. Averaging in the time domain
3. Sub-sampling in the time domain
4. Averaging of neighbouring coefficients in the Haar domain
5. Sub-sampling of neighbouring coefficients in the Haar domain
Compared techniques - 2

6. Signal reconstructed from CS domain (sampling $\nu' = \Phi \omega$)

7. Signal reconstructed from CS domain (sampling $\nu = \Phi f$)

8. “Hybrid” compressive sensing [Romberg 2008]

9. Signal reconstructed from hybrid CS domain
Compared techniques - 3

• All techniques (except original) compared at a parity of compressed size
• Classification provided by training an HMM per class (5 states, 2 Gaussian components per state) and ML classification
• Leave-one-actor-out cross validation
• Recovery by the LASSO in Donoho’s SparseLab
• Not interested in accuracy per se, dataset is easy
### Results (see paper)

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

**Note:** The accuracy values are given for different compression ratios and techniques.
Summary

![Graph showing performance comparison between different methods: CS (8 x 2), Reconstructed CS (8 x 2), Sub-sampling w (even coeff), Original signal, and Best of others. The x-axis represents the methods, and the y-axis represents the performance scores, with values ranging from 0 to 100. The scores for each method are visually compared to highlight their relative performance.]
Discussion

• The most notable result is that accuracy is higher after CS compression (no matter the signal is explicitly reconstructed or not)

• \( f = x + \varepsilon \)? no clear model for the noise

• Typical target of recovery algorithms: minimise the relative reconstruction error,

\[
\frac{\| \tilde{f} - f \|_{L_2}}{\| f \|_{L_2}}
\]

• Target of classification: minimise the classification error
Accuracy vs reconstruction error

 Different experiments for different Phi matrix
Accuracy vs coherence

- Accuracy vs max coherence and average coherence [Elad 2007]
Conclusions

• An interesting experimental results showing that compressive sensing in the time domain has provided an increase in action recognition accuracy

• Accuracy depends on actual $\Phi$; no obvious correlation with reconstruction error and (in)coherence

• Still preliminary results to be confirmed over other datasets and features
Thank-you slide

- Thank you
- Easy questions, please