Compressive Sensing based Face Detection without Explicit Image Reconstruction using Support Vector Machines

Filipe Magalhães[1,2], Ricardo Sousa[3], Francisco M. Araújo[1], and Miguel V. Correia[1,2]

[1] INESC TEC (formerly INESC Porto), Optoelectronics and Electronic Systems Unit
[2] DEEC, Faculdade de Engenharia, Universidade do Porto, Portugal
[3] Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto, Portugal
filipe.magalhaes@inescporto.pt

Abstract. The novel theory of compressive sensing takes advantage of the sparsity or compressibility of a signal in a specific domain allowing the assessment of its full representation from fewer measurements. In this work we tailored the concept of compressive sensing to assess the intrinsic discriminative capability of this method to distinguish human faces from objects. Afterwards we enrolled through a feature selection study to empirically determine the minimum amount of measurements required to properly detect human faces. This work was concluded with a comparative experiment against the SIFT descriptor. We determined that using only 40 measurements conducted by compressing sensing one is capable of capturing the relevant information that enable one to properly discriminate human faces from objects.

Keywords: Compressive Sensing, Feature Selection, Pattern Recognition, Face Detection

1 Introduction

Compressive Sensing (CS) is a recent and revolutionary technique that states that under certain conditions a signal or an image can be obtained from fewer measurements than those dictated by the fundamental Nyquist-Shannon sampling theorem. CS relies on the empirical observation that many types of signals or images can be well approximated by a sparse expansion in terms of a suitable basis, that is, by only a small number of non-zero coefficients. This is the key aspect of many lossy compression techniques such as JPEG and MP3, where compression is achieved by simply storing only the largest basis coefficients. It was in 2004 that an explosion of interest in this field occurred, when Emmanuel Candès was working on a problem in magnetic resonance imaging. He was surprised to discover that he could reconstruct a test image exactly even though the available data seemed insufficient according to the Nyquist-Shannon criterion [1, 2]. Many interesting contributions in very different areas derived from

this novel theory [5, 10], being the most representative example the single-pixel camera developed at the Rice University [14].

Face detection which nowadays has become a widespread tool due to the proliferation of mobile devices and social networks could eventually also represent another application of CS. In this work we raised the hypothesis of detecting sensible information such as human faces from images sampled in a compressive manner without requiring reconstruction of those images. For the detection process, a Support Vector Machine has been trained with the typical incoherent projections of CS. As redundancy may derive from the sampling process, which would bring an additional complexity to this framework, a feature selection (FS) algorithm was used to discard the correlated measurements and to identify the most relevant ones in a single step [12].

The rest of the paper is organized as follows: in Section 2 a brief introduction to the CS theory is presented followed by the description of the adopted FS approach (Section 3.1). For self-contained purposes, in Section 3, we present some background knowledge concepts. The created dataset and the methodology of the study here conducted are presented in Section 4. Finally, in Section 5 the experimental study is described and the conclusions are drawn in Section 6.

2 Compressive Sensing

With CS the information assumes a compressed form since the moment it is measured and it is enabled the possibility of reconstructing a sparse n-dimensional signal from $m < n$ measurements [3]. CS assumes that a signal of interest can be sparsely represented in a known basis. Discrete Cosines or wavelets are just two examples of typical bases used in imaging. These bases are chosen because they can compactly describe a signal with few coefficients, thus causing the signal to be considered sparse in that domain. Using the foundations of single-pixel imaging by CS we can shortly describe the principle of measurement behind our work. Consider an image with N pixels that is arranged in a $N \times 1$ column. Consider that image to be sparse when expressed in a conventional basis $\Psi = \{\psi_l\}_{l=1}^{N}$. Then, we can express $x = \Psi s$, where Ψ is a $N \times N$ matrix that has the vectors $\{\psi_l\}$ as columns and s is the $N \times 1$ vector composed of the expansion coefficients, being nonzero only a small portion of them. In order to determine x, the projections of the image are registered on a basis of M intensity patterns $\phi_m, m = 1, \ldots, M$ and this measurement process can be expressed as $y = \Phi x = \Phi(\Psi s) = \Theta s$ where y is the $M \times 1$ column containing the measured projections and Φ is the $M \times N$ measurement matrix. Each row of Φ defines an intensity pattern ϕ_m and the product of Φ and Ψ gives the $M \times N$ matrix Θ acting on s. CS theory states that x can be recovered with high probability from a $M < N$ random subset of coefficients in the Ψ domain. As the number of samples taken is smaller than the number of coefficients in the full image or signal, converting the information back to the intended domain would involve solving an underdetermined matrix equation. Thus, there would be an infinite number of candidate solutions and, as a result, a strategy to select the spars-

est solution must be found. This decoding or reconstruction problem can be seen as an optimization problem and be efficiently solved using the l1-norm [2]. Another relevant aspect of CS is incoherence which states that the rows of Φ (the measurement matrix) cannot describe the columns of Ψ in a sparse way (and vice-versa). In other words, the correlation between both shall be small. A particular aspect of interest, which will be used in this work, is that random matrices are largely incoherent with any fixed basis.

3 Background Knowledge

3.1 Feature Selection

Nowadays, it is relatively easy to solve pattern recognition problems with millions of instances, each of them with a reasonable number of features. However, it is common to have access to datasets with significantly higher number of features (in our work, measurements) than instances leading to the well-known problem of "the curse of dimensionality". Feature selection (FS) techniques provide the means to overcome this issue by identifying the most valuable features so that robust and simple class discrimination models can be obtained. There are three types of feature selection algorithms: Filter, wrapper and embedded. The former is independent of the classifier being usually done before the learning phase. Wrapper algorithms iteratively select subsets of features and assess the learning models performance to determine how useful those sets of features are, whereas embedded algorithms automatically select features during the model construction [12].

In [12] the authors define the FS task as a one-step process through quadratic programming. The quadratic term (Q in Equation (1)) captures the redundancy whereas the linear term (F in Equation (1)) captures the relevance. This is translated as follows:

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}(1-\alpha)\mathbf{x}'Q\mathbf{x} - \alpha F'\mathbf{x} \right\} \tag{1}$$

where the α constant can be considered as a trade-off between redundancy and relevance; and the $\mathbf{x}$ magnitude values show how important each feature is to the problem. To measure this correlation authors in [12] use either the Pearson or Mutual Information (MI) to measure linear or non-linear relations, respectively. For more information, the interested reader is advised to consult reference [12].

3.2 SIFT Descriptor

Scale Invariant Feature Transform (SIFT) is a well-known technique for feature detection and description with a different number of applications already reported in the literature [7, 9, 13]. In general, much of the works encompass the development of new SIFT-based descriptors especially designed for capturing facial structures. Nevertheless, for the problem of face vs. object detection, a standard SIFT is robust and general enough.

After applying an interest point detector, such as a LoG (Laplacian of Gaussian) or a DoG (Difference of Gaussian), for instance, the SIFT descriptor is built upon the measurements of the gradient in 8 different orientations in an image patch of size 4×4. As a result, it is obtained a feature vector of size 128. After the construction of the descriptor, an unsupervised technique is employed to construct the vocabulary that will represent our dataset, which is then concluded by a state-of-the-art bag-of-features technique [11].

3.3 Support Vector Machine

Support Vector Machine (SVM) is a popular learning mechanism. In its simplest form, SVM uses a linear separating hyperplane to create a binary classifier with a maximal margin. In cases where data cannot be linearly separable, data are transformed to a higher dimension than the original feature space. Such is done by choosing a given kernel function, representing the inner product in some implicit higher dimension space. More formally, given the training set $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ with input data $\mathbf{x}_i \in \mathbf{R}^p$ and corresponding binary class labels $y_i \in \{-1, 1\}$, the maximum-margin hyperplane is defined by $g(\mathbf{x}) = \mathbf{w}^t\varphi(\mathbf{x}) + b$ (where $\varphi(\mathbf{x})$ denotes a fixed-feature space transformation and b a bias parameter), $\mathbf{x}$ is assigned to class 1 if $g(\mathbf{x}) > 0$ or to -1 if $g(\mathbf{x}) < 0$. The maximization of the margin is equivalent to solving:

$$\begin{aligned} \min_{w,b}\ & \tfrac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{N}\xi_i \\ s.t.\ & y_i[\mathbf{w}^t\varphi(\mathbf{x}) + b] \geq 1 - \xi_i, i = 1, \ldots, N \\ & \xi_i \geq 0 \end{aligned} \tag{2}$$

The slack variables ξ_i, $i = 1, \ldots, N$ are introduced to penalyze incorrectly classified data points. Knowing that there is still much more to be covered, the interested reader is advised to consult, for example, reference [15] for more information.

4 Methods

As demonstrated by Duarte et. al [4], CS can be further extended to statistical inference related tasks, such as detection, classification and recognition since the signal reconstruction is not explicitly required, but only the relevant statistics for the problem at hand. Following this idea, in this work, it was envisioned and analyzed the possibility of detecting faces in images without explicit reconstruction. For this, the idea was to acquire incoherent projections of images, of faces and different objects, with Hadamard based random codes (built on-the-run, thus avoiding waste of memory) and, then, use those projections to train a classifier. For that, two sets of images were created, one containing 200 images of faces in an upright frontal position, with 2 images per person, and other containing 50 images of different objects/animals, with 2 images per object/animal. The faces' images were obtained from the "FEI Face Database" [8] and the objects/animals'

Fig. 1: Examples of images belonging to the created sets. Images of faces of two persons in upright frontal positions and images of two different objects/animals.

images were obtained from the "Caltech 101" database [6]. Images were cropped and resized to 256 × 256 pixels (see Fig. 1). For each image, it was created a vector comprising the incoherent projections produced with the Hadamard based codes. Each incoherent projection or measurement was obtained with a different code and corresponded to the sum of the pixel values of the image resulting from that product. As the images exhibited different intensities, each of the vectors was posteriorly normalized to values between 0 and 1. To reduce the computational complexity, the used images and codes were scaled down to 32 × 32 pixels, therefore, producing vectors with 1024 elements. Fig. 2 shows an example of a vector obtained for an image of a face.

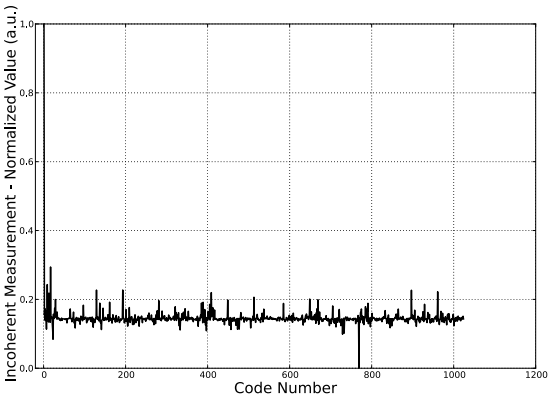

Fig. 2: Plot of a vector with the 1024 incoherent measurements obtained for an image of a face.

For the discrimination task a SVM [15] was used with all the features to establish a baseline. The next step in our study was to train the SVM with the features selected by the FS algorithm described in Section 3.1 to establish a comparison. Following the compressive sensing approach, as presented in Section 2, by performing different measurements for each image presented in our data set, this new data representation can encompass redundant information. In our experiments, the feature selection algorithm employed a wrapper strategy (See [12] for further details). Furthermore, in our study, only the Pearson correlation coefficient was used.

For comparative purposes, it was also assessed the classification performance with the information obtained with the SIFT descriptors. Fig. 3 shows the interest points detected for a set of images from the used data set. The vectors

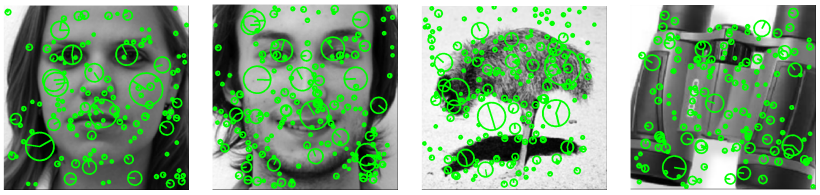

Fig. 3: Examples of images belonging to the created sets with over imposed blobs obtained with the SIFT descriptor. Images of faces of two persons in upright frontal positions and images of two different objects/animals.

containing the incoherent measurements were split into training and test sets to be used by the classifier. The training was performed with 40% of the data, being the respective remainder used as testing data. The splitting of the data into training and test sets was repeated 10 times in order to assess the variability of the obtained performances. The best parameterization of each model was found by a 'grid-search' based on a 5-fold cross validation scheme conducted on the training set. In our experiments we have used a RBF kernel given by: $k(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2), \gamma \geq 0$, and the grid search was performed over the parameters C and γ with the following values: C $= 2^{-3}, \ldots, 2^{10}$ and $\gamma = 2^{-5}, \ldots, 2^{3}$.

5 Results

We started our analysis by assessing the detection capability when all the compressive sensing incoherent measurements (1024) were used. For this case, the SVM yielded a performance rate of 97%. As mentioned, our comparison was performed against SIFT, which produced a performance rate of 80%. From these results it is possible to state that with a CS-based approach the classifier performance was better than that obtained with SIFT descriptors. One aspect that may support this difference is that the CS-based approach considers multiple measurements of the whole image. To understand the influence of each individual measurement we conducted a feature relevance study by comparing the FS approach described above with the performance provided by a random choice of features. We have also analysed the impact that the sequential aggregation of the measurements by their natural order had on the performance. These results are depicted in Fig. 4 under the 'sequential' label. It should be referred that in this study each feature corresponded to a measurement obtained with a specific code. In Fig. 4 one can observe the performance of the SVM classifier trained with 40% of the data. From the results presented in Fig. 4 it is clear the increased performance of the classifier with the feature selection algorithm. With

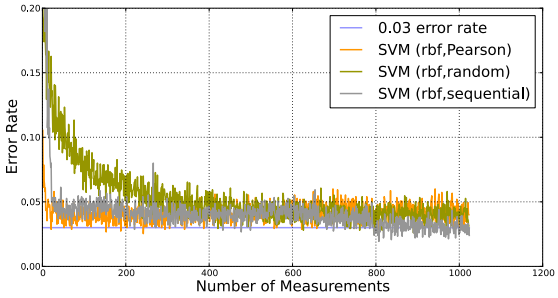

Fig. 4: Performance of the SVM classifier trained with 40% of data with features selected randomly (green trace) and with Pearson correlation as described in Section 3.1 (orange trace). A reference line of 97% performance is presented with a blue trace and the sequential approach with a grey trace.

the feature selection the classifier was able to provide an error rate close to 3% with roughly 40 measurements, while the same performance was reached only after the use of 370 measurements for the case of random feature selection, and roughly 200 measurements for the sequential case. The stabilization of the error rate for the case of the feature selection evidences its capacity to discriminate the relevance and redundancy of the involved features at an early stage. As a reference, it was also obtained the performance of the SVM classifier when all the 1024 measurements have been used. With these results, it can be said that using feature selection methods it was possible to obtain a competitive performance, translated by a smaller error rate, with far less features.

In this study, it was also evaluated which features typically provided the best results. Knowing that information, it could then be combined with compressive sensing theory to design a face detection system that could simultaneously minimize the amount of required measurements and maximize the amount of meaningful information. Evaluating the data, it was then possible to infer that the 16th, 28th, 408th, 512th, 768th and 784th measurements (generated with the corresponding Hadamard based codes) were the most commonly used features when a detection performance of 97% was obtained. In terms of computational effort, this fact can be extremely appealing once only a reduced set of codes has to be computed to detect a human face with such a significant performance.

6 Conclusion

Compressive Sensing (CS) is an interesting and powerful recent research topic finding an increasing number of applications at a fast pace. Its capability to reconstruct data from what appears to be incomplete information has opened a myriad of opportunities. In this work we considered a set of experiments to evaluate the capability of CS coupled with SVM to discriminate human faces from objects. The results obtained with CS were compared to those obtained with the SIFT descriptor and it was possible to determine that CS yielded a

better performance and that it is a promising technique for this discrimination task. Moreover, it was also possible to assess that with a small number of measurements we could obtain a classification performance of 97% for the tested data set. Finally, in face of the obtained results, we strongly believe that this framework has the potential to be further developed in the future and that it would be interesting to apply it to other scenarios such as video surveillance.

References

1. Candes, E., Tao, T.: Decoding by linear programming. Information Theory, IEEE Transactions on 51(12), 4203 – 4215 (2005)
2. Candès, E.J., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory 52(12), 5406–5425 (2006)
3. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory 52, 1289–1306 (2006)
4. Duarte, M., Davenport, M., Wakin, M., Baraniuk, R.: Sparse signal detection from incoherent projections. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. vol. 3, p. III (2006)
5. Duarte, M., Sarvotham, S., Baron, D., Wakin, M., Baraniuk, R.: Distributed compressed sensing of jointly sparse signals. In: Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers. pp. 1537 – 1541 (2005)
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. 106(1), 59–70 (2007), `http://dx.doi.org/10.1016/j.cviu.2005.09.012`
7. Geng, C., Jiang, X.: Face recognition using SIFT features. In: Proceedings of the 16th IEEE international conference on Image processing. pp. 3277–3280. ICIP'09, IEEE Press, Piscataway, NJ, USA (2009)
8. Junior, L., Thomaz, C.: Fei face database (2006), available from: `http://fei.edu.br/~cet/facedatabase.html`
9. Luo, J., Ma, Y., Takikawa, E., Lao, S., Kawade, M., Lu, B.L.: Person-Specific SIFT Features for Face Recognition. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. vol. 2, pp. II–593 –II–596 (2007)
10. Lustig, M., Donoho, D.L., Pauly, J.M.: Rapid MR imaging with "compressed sensing" and randomly Under-Sampled 3DFT trajectories. In: ISMRM (2006)
11. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. Computer Vision–ECCV 2006 pp. 490–503 (2006)
12. Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C.S.: Quadratic programming feature selection. Journal of Machine Learning Research 11, 1491–1516 (2010)
13. Stein, S., Fink, G.: A new method for combined face detection and identification using interest point descriptors. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. pp. 519 –524 (2011)
14. Takhar, D., Laska, J., Wakin, M.B., Duarte, M.F., Baron, D., Sarvotham, S., Kelly, K., Baraniuk, R.: A new compressive imaging camera architecture using Optical-Domain compression. In: Proc. IS&T/SPIE Symposium on Electronic Imaging (2006)
15. Vapnik, V.: Statistical learning theory. Wiley (1998)