RECOGNITION OF FACES AND FACIAL ATTRIBUTES USING ACCUMULATIVE LOCAL SPARSE REPRESENTATIONS

Domingo Mery[⋆] *Sandipan Banerjee*[†]

[⋆] Department of Computer Science – Pontificia Universidad Catolica de Chile
[†]Department of Computer Science and Engineering – University of Notre Dame

ABSTRACT

This paper addresses the problem of automated recognition of faces and facial attributes by proposing a new general approach called Accumulative Local Sparse Representation (ALSR). In the learning stage, we build a general dictionary of patches that are extracted from face images in a dense manner on a grid. In the testing stage, patches of the query image are sparsely represented using a *local dictionary*. This dictionary contains similar atoms of the general dictionary that are spatially in the same neighborhood. If the sparsity concentration index of the query patch is high enough, we build a descriptor by using a sum-pooling operator that evaluates the contribution provided by the atoms of each class. The classification is performed by maximizing the sum of the descriptors of all selected patches. ALSR can learn a model for each recognition task dealing with more variability in ambient lighting, pose, expression, occlusion, face size, etc. Experiments on three popular face databases (LFW for faces, AR for gender and Oulu-CASIA for expressions), show that ALSR outperforms representative methods in the literature, when a huge number of training images is not available.

Index Terms— Sparse representation, face recognition, facial attributes recognition, biometrics, computer vision.

1. INTRODUCTION

Recognition of faces and facial attributes have been a relevant research area in computer vision with many important contributions in the last decades. In recent years, we have witnessed tremendous improvements by using complex deep neural network architectures trained with millions of face images, *e.g.*, in face recognition [1]. Methods based on deep learning have become fundamental in this area. Nevertheless, in order to achieve satisfactory results, an enormous number of correctly labeled training images are required. In our work, we have focused on recognition tasks when a huge number of training images is not available.

We believe that algorithms based on sparse representations can be used for this task given that in many computer vision applications (including face recognition), and under the assumption that natural images can be represented using sparse decomposition, state-of-the-art results have been significantly improved [2]. In addition, in comparison with deep-learning techniques, sparse representation approaches do not require thousands or millions of images in order to learn a model. Thus, training complexity is significantly reduced.

Face recognition algorithms based on sparse representation have been widely used over the last decade [3]. In the sparse representation approach, a dictionary is constructed from the gallery images, and matching is undertaken by reconstructing the query image using a sparse linear combination of the dictionary. The identity of the query image is assigned to the class that has the least reconstruction error. Several variations of this approach were recently proposed. In [4], structured sparsity is proposed for dealing with the problem of occlusion and illumination. In [5], a new dictionary is constructed by the discriminative common vector per class. In [6], the dictionary is assembled by the class centroids and sample-to-centroid difference. In [7], the sparse representation is extended by incorporating the low-rank structure of data representation. In [8] and [9], sparse representations of patches distributed in a grid-like manner are used. In [10] for faces and in [11] for face attributes, patches that do not provide information (*e.g.*, occluded parts) are automatically filtered out in the recognition process.

Reflecting on the problems confronting recognition of faces and facial attributes, we believe that there are some key ideas that should be present in new proposed solutions. First, face parts that do not provide any information in this task (*e.g.*, sunglasses), should not be considered by the recognition algorithm. Second, parts of the face that are more relevant than other parts (*e.g.*, the mouth when recognizing happiness), should be class-dependent, and could be found using unsupervised learning. Third, feature extraction in face images should not be in fixed positions in order to consider misalignments. Fourth, rather than holistic approaches it would be helpful to search for similar face parts in all images of the gallery instead of similar gallery images.

Inspired in theses key-ideas, we propose in this paper a new method for face recognition that is able to deal with less constrained conditions. The contributions of our approach, called Accumulative Local Sparse Representation (ALSR), are the following two:

1) A new representation for the training images based on a dictionary of patches and the location in the face of each patch (similar to [12]). It corresponds to a rich collection of representation of relevant parts of the faces that are selected in the testing stage using closeness and similarity criteria.
2) A new strategy for the testing stage based on accumulative sparse contributions according to location and relevance criteria. With this criteria we select automatically patches that provide discriminative information avoiding patches from occluded parts por example.

The rest of the paper is organized as follows. In Section 2, the proposed method ALSR is explained in further details. In Section 3, the experiments and results are presented. Finally, in Section 4, the concluding remarks are given.

2. PROPOSED METHOD

The proposed method (ALSR) consists of two stages - 1) Learning and 2) Testing. In the learning stage, we build a general dictionary of patches that are extracted from training face images in a dense manner on a grid. In the testing stage, patches of the query image are extracted in the same way. For each query patch a local dictionary is built by selecting similar atoms of the general dictionary that are spatially in the same neighborhood. Using this local dictionary, each query patch is sparsely represented. If the sparsity concentration index of the query patch is high enough, we build a k-element descriptor (where k is the number of the classes) by using a sum-pooling operator that evaluates the contribution provided by the atoms of each class. The classification of the face image is performed by maximizing the sum of the descriptors of all selected patches.

2.1. Learning Stage

In the learning stage, we build dictionary $\mathbf{D}$ that contains patches of the classes of the gallery. We call this dictionary the *general dictionary*. The process starts with a set of n face images of k classes, where $\mathbf{I}_j^i$ denotes image j of class i (for $i = 1 \dots k$ and $j = 1 \dots n$). In each image $\mathbf{I}_j^i$, m patches $\mathcal{P}_{jp}^i$ of size $w \times w$ pixels (for $p = 1 \dots m$) are extracted in a grid manner centered in (x_{jp}^i, y_{jp}^i). The grid has m_v patches in vertical direction and m_h in horizontal direction, *i.e.*, $m = m_v \times m_h$.

In this work, a patch $\mathcal{P}$ is defined as *i)* vector $\mathbf{z} \in \mathcal{R}^d$, that is a descriptor of patch $\mathcal{P}$ (in our work $d = w \times w$, and the descriptor corresponds to the gray values of the patch given by stacking its columns); and *ii)* image coordinates (x, y) of the center of patch. Descriptor $\mathbf{z}$ is described using a vector normalized to unit length. All extracted patches are described as $\mathbf{y}_{jp}^i = f(\mathcal{P}_{jp}^i)$. Thus, for class i an array with the description of all patches is defined as $\mathbf{Y}^i = \{\mathbf{y}_{jp}^i\} \in \mathcal{R}^{d \times nm}$ (for $j = 1 \dots n$ and $p = 1 \dots m$). The general dictionary $\mathbf{D}$ is built by concatenating the arrays of all k classes $\left\{\mathbf{Y}^i\right\}_{i=1}^k$. Optionally, non-discriminative patches can be removed from our visual dictionary $\mathbf{D}$ using a *stop list* [11, 13].

2.2. Testing Stage

In testing stage, the task is to determine the class of query image $\mathbf{I}^t$ given the model learned in previous Section. From test image, m patches $\mathcal{P}_p^t$, for $p = 1 \cdots m$, are extracted as done in learning stage: the size of the patches is $w \times w$ pixels, the patches are extracted in a grid manner with $m = m_v \times m_h$, and the patches are described in the same way. Testing stage has three steps: *i)* construction of local dictionary, *ii)* sparse representation, and *iii)* analysis of contributions.

2.2.1. Local dictionary

For each extracted test patch, we have its description $\mathbf{y}_p^t$ and the coordinates of the center of the patch given by (x_p, y_p). For simplicity in this Section, we use patch $\mathbf{y}$ and coordinate (x, y).

In our method, we attempt to compute a sparse representation of $\mathbf{y}$ using dictionary $\mathbf{D}$, however, this approach requires a huge dictionary for reliable performance, *i.e.*, each sparse representation process would be very time consuming. This problem can be remedied by using only a part of the dictionary adapted to patch $\mathbf{y}$. Thus, the whole dictionary $\mathbf{D}$ can be reduced into a *local dictionary* by removing atoms of $\mathbf{D}$ that are not relevant, and only the selected (relevant) atoms can be used to compute the sparse representation of the patch.

Local dictionary is computed in two steps: closeness and similarity. In first step, using location information (x, y), we select from general dictionary $\mathbf{D}$ only those atoms that have been extracted close to (x, y). We call this new dictionary $\mathbf{C}$. In second step, using intensity information $\mathbf{y}$, we select from dictionary $\mathbf{C}$ only those atoms that are similar to $\mathbf{y}$. That means, we select the most similar patches from $\mathbf{C}$. We call this new dictionary $\mathbf{A}$.

2.2.2. Sparse Representation

With local dictionary $\mathbf{A}$, we look for a sparse representation of $\mathbf{y}$ using the ℓ_1-minimization approach:

$$\hat{\mathbf{x}} = \text{argmin}||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}. \tag{1}$$

In addition, we define vector $\mathbf{g}_i$, for $i = 1 \cdots k$, as a vector whose elements are the entries in $\hat{\mathbf{x}}$ corresponding to class i. The contribution of the atoms of class i in the sparse representation of the patch is defined as the sum of the absolute values of $\mathbf{g}_i$:

$$s_i = ||\mathbf{g}_i||_1. \tag{2}$$

Thus, for a k-class problem, the contribution vector of the patch is defined as a k-element vector:

$$\mathbf{s} = [s_1 \cdots s_k]. \tag{3}$$

In order to evaluate how the sparse coefficients of $\hat{\mathbf{x}}$ are distributed, we use the *sparsity concentration index* (SCI) of the patch [3], that is defined by

$$\text{SCI}(\mathbf{y}) = \frac{k \max(s_i)/||\mathbf{s}||_1 - 1}{k-1}. \quad (4)$$

SCI value is between 0 and 1, if patch $\mathbf{y}$ is discriminative enough its SCI is expected to be closer to 1.

2.2.3. *Analysis of Contributions*

In this step, we analyze the contribution of the patches. The key-idea of this step is that not all query patches are relevant, *i.e.*, some patches of the face do not provide any discriminative information of the class, *e.g.*, the nose is not important when recognizing facial expressions and sunglasses are not relevant when identifying a subject. This problem can be addressed by *i)* removing fixed patched (*e.g.*, the patches of the nose in expression recognition), and *ii)* by selecting automatically the query patches according to a score value. The first selection can be performed by using a mask over the face image (see examples in Fig. 1). Thus, only certain contributions of the face image will be used. The second selection is done automatically by not considering the contribution of a patch if its SCI value is below a threshold. Therefore, all elements s_i, for $i = 1 \cdots k$, are set to zero if SCI$< \theta_{\text{SCI}}$. In case the patch fulfills the SCI criteria, we normalize the contributions s_i by its maximal value: $\bar{s}_i = s_i / \max(s_i)$, and if $\bar{s}_i < \theta_c$ then $\bar{s}_i$ is set to zero. Thus, noise contributions can be removed. Finally, a patch $\mathbf{y}$ can be represented as a k-element vector of contributions:

$$\bar{\mathbf{s}}(\mathbf{y}) = [\bar{s}_1 \cdots \bar{s}_k] \quad (5)$$

Now, considering the whole query image, for each test patch $\mathbf{y}_p^t$, for $p = 1 \cdots m$, we obtain a normalized contribution vector $\bar{\mathbf{s}}(\mathbf{y}_p)$ given by (6). By summing all vectors, we achieve a rich representation of test image $\mathbf{I}^t$:

$$\mathbf{z}(\mathbf{I}^t) = \sum_{p=1}^{m} \bar{\mathbf{s}}(\mathbf{y}_p), \quad (6)$$

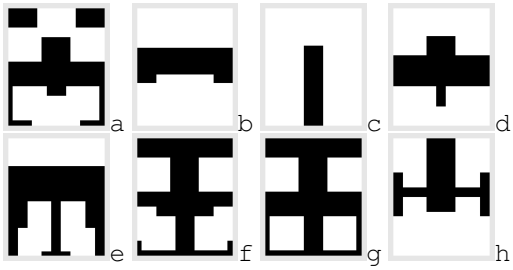

Fig. 1. Used face image masks: only patches that are centered in white zones are analyzed. The masks were defined manually by giving more relevance to certain parts of the face using superimposed rectangular regions.

that is a k-element vector $[z_1 \cdots z_k]$. Test image $\mathbf{I}^t$ will then assigned to class i if z_i is maximal.

3. EXPERIMENTAL RESULTS

We performed experiments with our method on three databases: LFW for face recognition [14], AR for gender recognition [15] and Oulu-CASIA for expression recognition [16]. For gender and expression recognition, we follow an evaluation protocol where training and testing subsets are subject-disjoint [17], *i.e.*, subjects that are present in training subset are not allowed to be in testing subset. In this section, we report the experiments in each dataset and the details of the implementation.

Face recognition in LFWa: The 'Labeled Faces in the Wild' (LFW) dataset [14] contains real-life images taken under unconstrained conditions, collected from the web. We used the deep funneled version ('LFWa' [18]). We followed hold-out experimental protocol given by [19]: subjects that have at least 10 images each were selected (giving a total of 143 subjects), and the first 10 images are used for training and the rest for testing. LFW face images have a large amount of intra-class variability, due to factors such as pose, background, expression and lighting. The recognition results on LFWa can be found in Table 1. In this table, we do not compare our algorithm with other deep learning methods that require millions of training images (for the sake of truth, VGG-F [1] in this experiment achieves 97.7%). An interesting result is shown in Fig. 2, where the robustness of ALSR against occlusion is demonstrated. In this example (with subject #98), the query image has sunglasses. Our method does not consider in the classification step those patches where SCI value is below a threshold ($\theta_{\text{SCI}} = 0.1$). Thus, the majority of the patches that are centered in the region of the sunglasses are filtered out. In addition, the figure shows the patches where the maximal

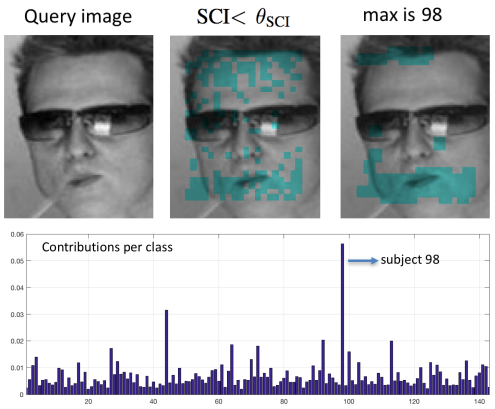

Fig. 2. Top: Patches selected in the faces with sunglasses. Bottom: Contribution per class (vector $\mathbf{z}$ see (7).

Table 1. Face Recognition in LFWa

Method	Reference	η [%]
LC-KSVD	[21]	66.0
DLSI	[22]	73.8
FDDL	[20]	74.8
LDL	[19]	77.2
JNPDL	[23]	78.1
SADL	[24]	78.4
ASR	[10]	78.5
ALSR	(ours)	**80.4**

Table 2. Gender Recognition in AR

Method	Reference	η [%]
LC-KSVD	[21]	86.8
LDL	[19]	95.3
FDDL	[20]	95.4
LGBP+SRC	[25]	97.7
ASR+	[11]	97.6
ALSR	(ours)	**98.9**

Table 3. Expression Recognition in Oulu-CASIA

Method	Reference	η [%]
UDCS	[26]	49.5
GoogLeNet	[27]	66.6
ALSR	(ours)	**68.2**

contribution is given by atoms corresponding to class #98. In the final contribution per class, it is clear that this query image is classified as subject #98.

• **Gender recognition in AR:** The images in the 'AR' dataset [15] were taken from 100 subjects (50 women and 50 men) with different facial expressions, illumination conditions, and occlusion with sun glasses and scarf (we used the cropped version). For gender recognition, we followed the protocol from [20] that uses the non-occluded subset (14 images per subject). In this experiment, the first 25 males and 25 females were used for training and the last 25 males and 25 females were used for testing. See results in Table 2, where sample male and female images are shown.

• **Expression recognition in Oulu-CASIA:** For expression recognition we use the Oulu-CASIA dataset [16]. In this dataset, face images were taken with six different facial expressions (surprise, happiness, sadness, anger, fear and disgust) under normal illumination from 80 subjects (59 males and 21 females) ranging from 23 to 58 years in age. The dataset contains 480 sequences. We used the protocol suggested in [26], where the first 9 images of each sequence are not considered, the first 40 individuals are taken as training subset and the rest as testing. See results in Table 3, where the six expressions are shown in different sample subjects. We exclude method PPDN [27] from Table 3 in which the accuracy was 72.4%, because the training stage of PPDN, instead of single images, includes pairs of images (one for the peak expression and another one for the non-peak expression).

The significance of our results is twofold: *i)* ALSR is a general recognition algorithm that can be used in different facial attribute analysis with few number of parameters. *ii)* Results show that ALSR deals well with unconstrained conditions in every experiment, achieving a high recognition performance in many complex conditions and obtaining better performance in comparison with other representative methods.

For the experiments, we used the implementation of k-means and sparse coding from [28] and [29] respectively. The rest of the algorithms were implemented on MATLAB. In our experiments, the total training time / the testing time per query image in seconds for each experiment was: 500/20 (LFWa), 3000/3 (AR) and 450/15 (Oulu-CASIA). The experiments were carried out on a iMac OS X 10.12.4 with a 3.7 GHz Quad-Core Intel Xeon E5 processor and 12GB memory (12GB RAM 1866 MHz DDR3). The code of the MATLAB (version R2016a) implementation and the detail of the used parameters are available on our webpage[1].

4. CONCLUSION

In this paper we presented a new algorithm that is able to recognize faces and facial attributes automatically from face images captured under less constrained conditions including some variability in ambient lighting, pose, expression, size of the face and distance from the camera. The robustness of our algorithm is due to two reasons: *i)* The dictionary used in the recognition corresponds to a rich collection of representations of relevant parts which were selected using closeness and similarity criteria. *ii)* The testing stage is based on accumulative sparse contributions according to location and relevance criteria. Combining these ideas, the algorithm deals with unconstrained conditions very well achieving high recognition performance in many complex conditions outperforming the other tested algorithms. We believe that this new approach can be used to solve other kind of computer vision problems in which there are similar unconstrained conditions and a huge number of training images is not available. In the future, we will train our own deep learning network to obtain a better description of the patches, and we will learn the face image masks from training data, instead of manual selection.

Acknowledgments

This work was supported by Fondecyt Grant No. 1161314 from CONICYT, Chile.

[1] See `http://dmery.ing.puc.cl/index.php/material/`.

5. REFERENCES

[1] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC 2015)*. Springer, 2016, p. 1.

[2] Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang, "A survey of sparse representation: algorithms and applications," *IEEE access*, vol. 3, pp. 490–530, 2015.

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

[4] Kui Jia, Tsung-Han Chan, and Yi Ma, "Robust and practical face recognition via structured sparsity," in *European Conference on Computer Vision (ECCV 2012)*, pp. 331–344. Springer, 2012.

[5] Ying Wen, "A novel dictionary based src for face recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2582–2586.

[6] Weihong Deng, Jiani Hu, and Jun Guo, "In defense of sparsity based face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, 2013, pp. 399–406.

[7] Jie Chen and Zhang Yi, "Sparse representation for face recognition by discriminative low-rank matrix recovery," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 763–773, 2014.

[8] Yi Chen, Thong T Do, and Trac D Tran, "Robust face recognition using locally adaptive sparse representation," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1657–1660.

[9] Tomas Larrain, John S Bernhard, Domingo Mery, and Kevin W Bowyer, "Face recognition using sparse fingerprint classification algorithm," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1646–1657, 2017.

[10] Domingo Mery and Kevin Bowyer, "Face recognition via adaptive sparse representations of random patches," in *IEEE International Workshop on Information Forensics and Security (WIFS2014)*. IEEE, 2014, pp. 13–18.

[11] Domingo Mery and Kevin Bowyer, "Automatic facial attribute analysis via adaptive sparse representation of random patches," *Pattern Recognition Letters*, vol. 68, pp. 260–269, 2015.

[12] Chun-Guang Li, Jun Guo, and Hong-Gang Zhang, "Local sparse representation based classification," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 649–652.

[13] J Sivic and A Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision (ICCV 2003)*, 2003, pp. 1470–1477.

[14] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[15] A Martinez and R Benavente, "The AR face database," June 1998, CVC Tech. Rep, No. 24.

[16] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[17] D. Mery, Y. Zhao, and K. Bowyer, "On accuracy estimation and comparison of results in biometric research," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–8.

[18] G B Huang, M A Mattar, H Lee, and E G Learned-Miller, "Learning to Align from Scratch.," *NIPS*, 2012.

[19] Meng Yang, Dengxin Dai, Lilin Shen, and Luc Van Gool, "Latent Dictionary Learning for Sparse Representation Based Classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014, pp. 4138–4145.

[20] Meng Yang, D Zhang, and Xiangchu Feng, "Fisher Discrimination Dictionary Learning for sparse representation," in *IEEE International Conference on Computer Vision (ICCV 2011)*, 2011, pp. 543–550.

[21] Zhuolin Jiang, Zhe Lin, and L S Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.

[22] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*. IEEE, 2010, pp. 3501–3508.

[23] Weiyang Liu, Zhiding Yu, and Meng Yang, "Jointly learning non-negative projection and dictionary with discriminative graph constraints for classification," *arXiv preprint arXiv:1511.04601*, 2015.

[24] Heyou Chang, Meng Yang, and Jian Yang, "Learning a structure adaptive dictionary for sparse representation based classification," *Neurocomputing*, vol. 190, pp. 124–131, 2016.

[25] Cunjian Chen and Arun Ross, " Local gradient Gabor pattern (LGGP) with applications in face recognition, cross-spectral matching, and soft biometrics," *SPIE Defense, Security, and Sensing*, vol. 8712, pp. 87120R, May 2013.

[26] Mingliang Xue, Wanquan Liu, and Ling Li, "The uncorrelated and discriminant colour space for facial expression recognition," in *Optimization and Control Techniques and Applications*, vol. 86 of *Springer Proceedings in Mathematics & Statistics*, pp. 167–177. Springer, 2014.

[27] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan, "Peak-piloted deep network for facial expression recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 425–442.

[28] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008, http://www.vlfeat.org/.

[29] J Mairal, F Bach, J Ponce, G Sapiro, R Jenatton, and G Obozinski, *SPAMS: SPArse Modeling Software*, INRIA, 2014, Software available on http://spams-devel.gforge.inria.fr.