Multimodal Dictionary Learning and Joint Sparse Representation for HEp-2 Cell Classification

Ali Taalimi[1], Shahab Ensafi[2,3], Hairong Qi[1], Shijian Lu[2], Ashraf A. Kassim[3], and Chew Lim Tan[4]

[1] University of Tennessee-Knoxville
[2] Institute for Infocomm Research, A*STAR, Singapore
[3] Electrical and Computer Engineering Dept., National University of Singapore
[4] School of Computing, National University of Singapore

Abstract. Use of automatic classification for Indirect Immunofluorescence (IIF) images of HEp-2 cells is increasingly gaining interest in Antinuclear Autoantibodies (ANAs) detection. In order to improve the classification accuracy, we propose a multi-modal joint dictionary learning method, to obtain a discriminative and reconstructive dictionary while training a classifier simultaneously. Here, the term 'multi-modal' refers to features extracted using different algorithms from the same data set. To utilize information fusion between feature modalities the algorithm is designed so that sparse codes of all modalities of each sample share the same sparsity pattern. The contribution of this paper is two-fold. First, we propose a new framework for multi-modal fusion at the feature level. Second, we impose an additional constraint on consistency of sparse coefficients among different modalities of the same class. Extensive experiments are conducted on the ICPR2012 and ICIP2013 HEp-2 data sets. All results confirm the higher level of accuracy of the proposed method compared with state-of-the-art.

1 Introduction

Application of automated Computer Aided Diagnosis (CAD) system to support clinicians in the field of Indirect Immunofluorescence (IIF) has been increased in recent years. Use of CAD system enables test repeatability, lowers costs and results in more accurate diagnosis. IIF imaging technique is applied to Human Epithelial Cells type 2 (HEp-2 cells), where antibodies are first stained in a tissue and then bound to a fluorescent chemical compound. In case of Antinuclear Antibodies (ANAs), the antibodies bound to the nucleus demonstrate different visual patterns which can be captured and visualized within microscope images [5]. These patterns can be used for cell classification and for assisting diagnosis. Image quality variations makes interpretation of fluorescence patterns, very challenging. To make the pattern interpretation more consistent, automated methods for classifying the cells are essential.

Recently, there has been an increasing interest in sparse coding in computer vision and image processing research for reconstructive and discriminative tasks [9,7,1]. In sparse coding the input signal is approximated by a linear combination of a few atoms of the dictionary. The state-of-the-art method in HEp-2 cell classification problem is proposed in [2], where the SIFT and SURF features are extracted as the input features to learn a dictionary followed by Spatial Pyramid Matching (SPM) [8] to provide the sparse representation of the input cell images. Then a Support Vector Machine (SVM) is learned to classify the test images.

All above mentioned approaches use unsupervised dictionary learning where the dictionary is obtained purely based on minimizing the reconstruction error. However, in supervised scheme, minimization of misclassification and reconstruction errors results in a dictionary which is adapted to a task and data set [9,10] and leads to a more accurate classification compared with unsupervised formulation. In some supervised approaches the sparse codes obtained in training are not used for classifier training and test signal is classified only based on reconstruction error [10]. Although [13,1] exploit sparse codes to train classifier; it is done independent of dictionary learning. We intend to estimate the dictionary and classifier, jointly so that generated sparse codes by dictionary are more discriminative, leading to better classification result.

The majority of existing dictionary learning methods, supervised or unsupervised, can handle only single source of data [7]. Fusion of information from different sensor modalities can be more robust to single sensor failure. The information fusion happens in feature level or classifier level [14]. In feature fusion different types of features are combined to make one representation to train a classifier while in classifier fusion, for each modality one classifier is trained independent of others and their decisions would be fused. In Bag-of-Words, feature fusion is imposed by concatenating all of features in one vector. The dimension of this vector is high and suffers from curse-of-dimensionality while it does not even contain the valuable information of correlation between feature types.

We propose a supervised algorithm similar to [7] to learn a compact and discriminative dictionary in all-vs-all fashion for each modality. This method can combine information from different feature types and force them to have common sparsity patterns for each class, which is presented in Fig. 1.

2 Method

Notation. Let C represent the number of classes, $\mathcal{M} \triangleq \{1 \cdots M\}$ be a set of M different feature modalities, $\{\mathcal{Y}_{i,c}\}_{i=1}^{N}$, $c \in \{1, \cdots, C\}$ as N training samples where each sample belong to c-th class and has M feature modalities as $\mathcal{Y}_{i,c} = \{Y_i^m \in R^{n^m \times S} | m \in \mathcal{M}\}$ where n^m is the dimension of the m-th feature modality and S is the number of interest points in the image which is the same for all modalities. The binary matrix $H_i \in \mathcal{R}^{C \times S}$ is an identifier for the label of $\mathcal{Y}_{i,c}$. Given $\mathcal{Y}_{i,c}$ from c-th class, the c-th row of H_i is one and all other rows are zero. Also, consider Y^m as set of m-th feature modality of all training samples $Y^m \in \mathcal{R}^{n^m \times K} = [Y_1^m \cdots Y_N^m]$ where $K = N \times S$ is the total number of samples in m-th

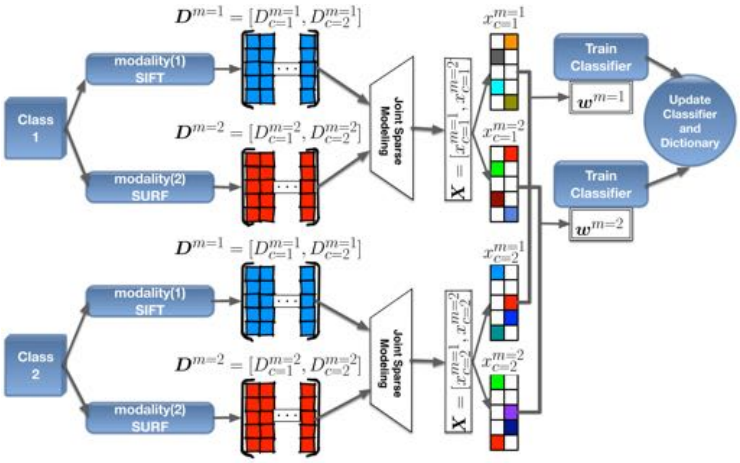

Fig. 1. Multi-modal supervised dictionary learning where two classes and two modalities for each class are assumed. We expect $X_{c=1}^{m=1}$ and $X_{c=1}^{m=2}$ have same sparsity pattern.

modality. The label matrix of Y^m is $H=[H_1 \cdots H_N]$. Corresponding dictionary of m-th modality $D^m \in \mathbb{R}^{n^m \times p}$ has p atoms. D^m is composed of class-specific sub-dictionaries D_c^m as $D^m=[D_1^m \cdots D_C^m]$. Also, assuming w^m as parameters of m-th modality classifier, W is set of all classifiers, $W = \{w^m | m \in \mathcal{M}\}$.

2.1 Supervised Dictionary Learning

Supervised dictionary learning can be done in one-vs-all scheme by training an independent dictionary for each class or in all-vs-all setting where the dictionary is shared between classes. We adopt all-vs-all scheme which allows feature sharing among the classes to obtain modality-based dictionary D^m.

Assuming sample $\mathcal{Y}_{i,c}$ from c-th class, we define binary matrix $Q_i \in \mathcal{R}^{p \times S} = [q_1 \cdots q_S]$. Each column q_i is zero everywhere except for indices of atoms which belong to the c-th class. The relation between labels of Y^m and labels of atoms in D^m is determined by matrix $Q = [Q_1 \cdots Q_N]$. The so called label consistency constraint is applied using Q so that each sample is reconstructed from atoms that belong to the same class as the sample.

The dictionary D^m can be estimated by minimizing $\mathcal{L}_u(X^m, D^m)$ using elastic-net formulation [16] as $\mathcal{L}_u(.) \triangleq \min \|Y^m - D^m X^m\|_2^2 + \lambda_1 \|X^m\|_1 + \lambda_2 \|X^m\|_F^2$ where λ_1 and λ_2 are regularization parameters. $\mathcal{L}_u$ is an *unsupervised* reconstruction loss function and is small if D^m is successful in finding sparse representation of Y^m. Given X^m obtained by elastic-net, *supervised* loss function $\mathcal{L}_{su}^m$, for dictionary learning and classifier training for modality m is formulated as [7]:

$$\underset{w^m, A^m}{\operatorname{argmin}} \mathcal{L}_{su}^m(D^m, Y^m, w^m, H, A^m, Q) + \frac{\nu_1}{2}\|w^m\|_F^2 + \frac{\nu_2}{2}\|A^m\|_F^2 \qquad (1)$$

where ν_1, ν_2 are regularization parameters. The supervised loss function of m-th modality is defined as $\mathcal{L}_{su}^m \triangleq \mu\|Q - A^m X^m\|_2^2 + (1-\mu)\|H - w^m X^m\|_2^2$ with μ as a regularization and A^m as a linear transformation matrix. The so called label consistency prior $\|Q - A^m X^m\|_F^2$ allows sparse code X^m to be different from Q up to a linear transformation A^m; hence it forces sparse representation of different classes to be discriminative. The classification error in $\mathcal{L}_{su}^m$, $\|H - w^m X^m\|_F^2$ shows that how well H can be predicted by the linear classifier with parameter w^m.

We want that multi-modal sparse representation $X_c^1, \cdots, X_c^M$ of data of c-th class, $\mathcal{Y}_{i,c}$, share same sparsity pattern. We propose multi-modal supervised dictionary learning and joint sparse modeling as:

$$X_c = \underset{X_c=[X_c^1,\cdots,X_c^M]}{\operatorname{argmin}} \sum_{m=1}^{M} \mathcal{L}_{su}^m(D^m, w^m, A^m, X^m) + \eta \|X_c\|_{1,2} \tag{2}$$

each sub-matrix X_c^m is sparse representation for data reconstruction of m-th modality and c-th class. Collaboration between $X_c^1, \cdots, X_c^M$ is imposed by $\|X_c\|_{1,2}$ in (2) and is defined as $\|X\|_{1,2} = \sum_{r=1}^{p} \|x_r\|_2$; where x_r are rows of X_c. The $l_{1,2}$ regularization $\|X\|_{1,2}$ promotes solution with sparse non-zero rows x_r; hence, sparse representations share the consistent pattern across all the modalities of the same class.

Optimization. As suggested in [9], the modality-based dictionary D^m is trained over Y^m using elastic-net [16]. This is done for each modality, independently to obtain multi-modal dictionaries $D = \{D^m | m \in \mathcal{M}\}$. We expect the data of c-th class to be reconstructed by atoms that belong to the c-th class. Given multi-modal dictionaries D, the joint sparse representation of $\mathcal{Y}_i$) is calculated using (2) and solved by proximal algorithm [12]. Then, we make modality-based sparse codes of m-th modality as $X^m = [X_1^m, \cdots, X_C^m]$. Assuming $\tilde{X} = (X^m)^T$, multivariate ridge regression model with quadratic loss and l_2 norm regularization are adopted to estimate initial values of w^m and A^m:

$$A^m = Q\tilde{X}(\tilde{X}^T\tilde{X} + I)^{-1}, \quad w^m = H\tilde{X}(\tilde{X}^T\tilde{X} + I)^{-1} \tag{3}$$

where I is identity matrix. The final values of D^m and w^m is obtained using stochastic gradient descent scheme proposed in [9,7]. The proposed algorithm is summarized in Algorithm (1).

3 Experiments and Results

We evaluate proposed method on two publicly available HEp-2 image datasets, referred to as ICPR2012[1] and ICIP2013[2]. Fig. 2 shows the ICPR2012 that contains 1445 cells in six categories and divided to train and test sets by the organizers. Fig. 3 shows the ICIP2013 that has 13650 cells in six categories for the training set but the test set is not publicly available. Also, each cell image is

[1] http://mivia.unisa.it/datasets/biomedical-image-datasets/hep2-image-dataset/
[2] http://mivia.unisa.it/icip-2013-contest-on-hep-2-cells-classification/

Algorithm 1: Multi-modal Dictionary and Classifier Learning

Input: $Y^m \; \forall m \in \{1 \cdots M\}, Q, H, \mu, \eta$ and T =number of iterations
Output: $D^m, w^m \;\; \forall m \in \{1 \dots M\}$
begin
  foreach *modality* $m \in \{1, \cdots, M\}$ **do**
    foreach *class* $c \in \{1, \cdots, C\}$ **do**
      Obtain D_c^m from $\mathcal{Y}_{i,c}$ using elastic-net;
    Find initial value of modality-based dictionary $D_0^m = [D_1^m, \cdots, D_C^m]$;
    Estimate D^m by applying elastic-net on Y^m given D_0^m
  Solve joint sparse coding problem (2) to find X_c using proximal method [12];
  Initialize w^m and A^m using (3)
  foreach *modality* $m \in \{1, \cdots, M\}$ **do**
    for *iter* $= 1 \cdots T$ **do**
      foreach *mini-batch samples of* Y^m **do**
        Update learning rate, D^m, A^m and w^m by a projected gradient step following [9];
end

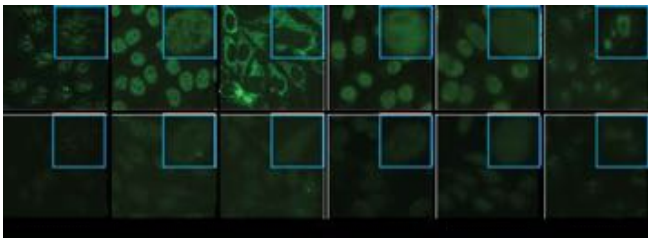

Fig. 2. ICPR2012 dataset. Positive (Top) and Intermediate (Bottom) images.

assigned to one of the two types of intensity patterns: positive or intermediate, which can be used as a prior information.

To prove the effectiveness of the proposed joint sparse representation we report our performance for four scenarios: sift (OnlySIFT), surf (OnlySURF), concatenation of sift and surf features (SIFTSURF) and joint sift and surf (Joint).

3.1 Implementation Details

Choosing the Parameters. To reduce the burden of required cross validation to set regularization parameters λ_1, λ_2 (elastic-net parameters), ν_1, ν_2 in (1), η in (2) and p (number of atoms in dictionary), we follow generally accepted heuristics proposed in [9]. To promote sparsity similar to [9,7] we set λ_2=0 and choose λ_1 by cross-validation in the set $\lambda_1 = 0.15 + 0.025k$ with $k \in \{-3, \cdots, 3\}$ and set it to $\lambda_1 = 0.5$. We observed that increasing number of atoms, p, usually leads to a better performance at the cost of higher computational complexity. We try the values p from $\{30, 60, 100, 150\}$. Our experiments on ν_1, ν_2 confirms observations in [9,7] that when p is smaller than number of normalized training patches, ν_1 and

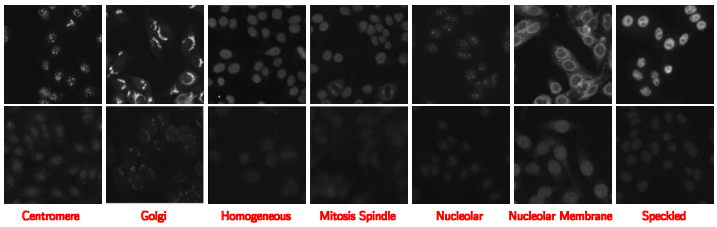

Fig. 3. ICIP2013 dataset. Positive (Top) and Intermediate (Bottom) images.

ν_2 can be arbitrarily set to small value. We try ν_1 and ν_2 from $\{10^{-1}, \cdots, 10^{-8}\}$ and choose $\nu = \nu_1 = \nu_2$ for both data sets. The regularization parameter η is selected by cross-validation in the set $\{0.001, 0.01, 0.05, 0.1, 0.2, 0.3\}$.

We extract 2-modalities of SIFT and SURF from each cell image. Each of these modalities are extracted from patches of 16×16 that are densely sampled using a grid with step size of 6 pixels. Then, spatial pyramid represents each feature-type using three grids size 1×1, 2×2 and 4×4 and codebook with $k = 900$ atoms [8]. The vector quantization codes of all spatial subregion of the spatial pyramid are pooled together to construct a pooled feature. The final spatial pyramid feature of each cell image is obtained by concatenating and l_2 normalization of the pooled features originated from subregions.

We train D_c^m from $\mathcal{Y}_{i,c}^m$ using elastic-net. Then, the initial value for dictionary of m-th modality, D_0^m is obtained by concatenating $D_c^m|_{c\in\{1\cdots C\}}$. This way we know the class label of each atom in D^m. The D^m is tuned by running elastic-net once more on training data of m-th modality Y^m given initial value D_0^m. Unlike all other methods of HEp-2 classification an explicit corresponding is made between labels of atoms in D^m and labels of data in Y^m; hence the estimated sparse codes are more distinctive. This leads to high accuracy classification result while D^m has a few atoms. We consider $p = 100$ and $p = 150$ atoms for dictionary of each cell class; hence modality-based dictionary D^m has 600 and 900 atoms for all six cell classes for ICIP2012 and ICIP2013, respectively.

The evaluation for the ICPR2012 is performed on the provided test set. Since the test set is not publicly available for ICIP2013 dataset, we follow [6] to design train and test. Training set includes 600 samples from each class except Golgi which has 300 cell samples. The remaining samples belong to the test data. In both datasets, we report performance of our method on each intensity level separately and final result is the average of classification results. As suggested by the competition organiser we evaluate our method based on Mean Class Accuracy (MCA): $MCA = \frac{1}{C}\sum_{c=1}^{C} CCR_c$; where CCR_c is correct classification rate of c-th class.

We report the performance of the proposed method for different values of ν and μ when η is changed from 0.1 to 0.7 for ICPR2012 in Fig. 4. For each ν we report the accuracy once with considering label consistency constraint as dotted line (μ=0) and once with the label consistency involved ($\mu = 0.3$). The performance is always better with label consistency constraint. Fig. 4 agrees the

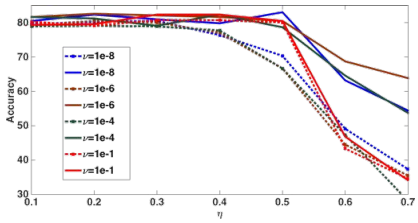

Fig. 4. The effect of changing parameters on ICIP2012 positive samples. $\mu = 0.3$ and $\mu = 0$ for the straight and dotted lines, respectively for different η values.

Table 1. The MCA accuracy on test set of ICPR2012 dataset and Comparison with state-of-the-art on ICIP2013.

ICPR2012	**OnlySIFT**	**OnlySURF**	**SIFTSURF**	**Joint**	[1]	[4]	[3]	[11]	[15]	[6]
Positive	74	72	76	**82**	81	62	63	74	69	78
Intermediate	67	66	69	**79**	62	41	60	35	48	48
Average Accuracy	70	69	73	**80**	72	52	62	55	59	63

ICPR2013	**OnlySIFT**	**OnlySURF**	**SIFTSURF**	**Joint**	[1]	[6]
Positive	88.4	90.3	90.7	**98.2**	95.8	95.5
Intermediate	76.2	72.5	81.2	**92.1**	87.9	80.9
Average Accuracy	82.3	81.4	85.9	**95.1**	91.9	88.2

observations made by [9,7] that ν should be set to small value when the number of training patches is a lot more than number of atoms. We set $\eta = 0.2, \mu = 0.3$ and $\nu = 1e-4$ in our experiments.

We compare performance of our method with state-of-the-art on ICPR2012 in Table 1. Our supervised method has 82% and 79% accuracy on positive and intermediate classification. It increases accuracy of OnlySIFT, OnlySURF more than 10% and enhances SIFTSURF around 7%. It also, outperforms other methods on average accuracy by at least 8%.

In the cell level classification on ICIP2013, Table 1 shows that applying SIFT and SURF jointly using our method enhances accuracy of OnlySIFT and OnlySURF around 13% while getting better result than simple concatenation of SIFTSURF at least 8% on average accuracy. It also outperforms other methods more than 3% on average accuracy. The proposed joint method shows superior results than concatenation of feature modalities in one vector in both datasets.

4 Conclusion

The problem of HEp-2 cell classification using sparsity scheme was studied and a supervised method was proposed to learn the reconstructive and discriminative dictionary and classifier simultaneously. Having label consistency constraint within each modality and applying joint sparse coding between modality-based sparse representations leads to discriminative dictionary with few atoms. The imposed joint sparse prior enable algorithm to fuse information in feature-level

by forcing their sparse codes to collaborate and in decision-level by augmenting the classifier decisions. The result of HEp-2 cell classification experiments demonstrates that our proposed method outperforms state-of-the-art while using common features. It is trivial that our approach will further improve by adding complex and well-designed features [6].

References

1. Ensafi, S., Lu, S., Kassim, A.A., Tan, C.L.: Automatic cad system for hep-2 cell image classification. In: Pattern Recognition (ICPR), 2014 22nd International Conference on. pp. 3321–3326. IEEE (2014)
2. Ensafi, S., Lu, S., Kassim, A.A., Tan, C.L.: A bag of words based approach for classification of hep-2 cell images. In: Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A), 2014 1st Workshop on. pp. 29–32. IEEE (2014)
3. Ensafi, S., Lu, S., Kassim, A.A., Tan, C.L.: Sparse non-parametric bayesian model for hep-2 cell image classification. In: Biomedical Imaging: From Nano to Macro, 2015. ISBI '15. IEEE International Symposium on. IEEE (April 2015)
4. Foggia, P., Percannella, G., Soda, P., Vento, M.: Benchmarking hep-2 cells classification methods. Medical Imaging, IEEE Transactions on 32(10), 1878–1889 (2013)
5. González-Buitrago, J.M., González, C.: Present and future of the autoimmunity laboratory. Clinica chimica acta 365(1), 50–57 (2006)
6. Han, X.H., Wang, J., Xu, G., Chen, Y.W.: High-order statistics of microtexton for hep-2 staining pattern classification. Biomedical Engineering, IEEE Transactions on 61(8), 2223–2234 (Aug 2014)
7. Jiang, Z., Lin, Z., Davis, L.S.: Label consistent k-svd: learning a discriminative dictionary for recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(11), 2651–2664 (2013)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. vol. 2, pp. 2169–2178. IEEE (2006)
9. Mairal, J., Bach, F., Ponce, J.: Task-driven dictionary learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(4), 791–804 (2012)
10. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. pp. 1–8. IEEE (2008)
11. Nosaka, R., Fukui, K.: Hep-2 cell classification using rotation invariant co-occurrence among local binary patterns. Pattern Recognition 47(7), 2428–2436 (2014)
12. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends in optimization 1(3), 123–231 (2013)
13. Ramirez, I., Sprechmann, P., Sapiro, G.: Classification and clustering via dictionary learning with structured incoherence and shared features. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. pp. 3501–3508. IEEE (2010)
14. Ruta, D., Gabrys, B.: An overview of classifier fusion methods. Computing and Information systems 7(1), 1–10 (2000)
15. Wiliem, A., Sanderson, C., Wong, Y., Hobson, P., Minchin, R.F., Lovell, B.C.: Automatic classification of human epithelial type 2 cell indirect immunofluorescence images using cell pyramid matching. Pattern Recognition 47(7), 2315–2324 (2014)
16. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Royal Statistical Society: Series B (Statistical Methodology) 67(2), 301–320 (2005)