# Multiple Instance Dictionary Learning using Functions of Multiple Instances

Changzhe Jiao
Electrical & Computer Engineering
University of Missouri
Email: cjr25@mail.missouri.edu

Alina Zare
Electrical & Computer Engineering
University of Missouri
Email: zarea@missouri.edu

***Abstract*—Dictionary Learning Functions of Multiple Instances (DL-FUMI) is proposed to address target detection problems with inaccurate training labels. DL-FUMI is a multiple instance dictionary learning method that estimates target atoms that describe distinctive and representative features of the target class and background atoms that account for the shared features found across both target and non-target data points. Experimental results show that the target atoms estimated by DL-FUMI are more discriminative and representative of the target class than comparison methods. DL-FUMI is shown to have improved performance on several detection problems as compared to other multiple instance dictionary learning algorithms.**

## I. Introduction

Obtaining accurate training label information is often time consuming, expensive, and/or infeasible for large data sets. Furthermore, annotators may be inconsistent during labeling providing inherently imprecise labels. Thus, in many applications, one has access only to inaccurately or weakly labeled training data.

Sparse coding and dictionary learning methods, whose low-rank data representations generally reduce redundancy and improve discrimination ability, have been successfully applied to many applications [1]. DL-FUMI leverages the benefits of discriminative dictionary learning for target detection applications given only inaccurate labels. This is accomplished through the use of a novel model that assumes each target data point is a *mixture* of both target and background atoms whereas non-target data points are composed of only background atoms. In other words, unlike the majority of discriminative dictionary learning methods, DL-FUMI does not learn a separate dictionary for each class. Instead, DL-FUMI introduces a shared background dictionary that is used in reconstruction of both target and non-target points. The advantage of this model over class-specific dictionaries is that the target atoms only need to account for the unique characteristics of target (and do not need to address any shared background variability) resulting in more discriminative, representative target atoms.

Furthermore, the target atoms estimated by DL-FUMI can be examined to uncover what discriminates target data points. Since most approaches estimate class-specific dictionaries, each dictionary must characterize both the class-specific characteristics and the characteristics shared among all data points. Thus, in these methods, it is often difficult to pin down what is unique about each class without prior insight since the class-specific features are mixed with background features and spread across the atoms. In contrast, DL-FUMI provides that insight by pulling out the unique target characteristics and identifying which atoms contain those characteristics. In summary, DL-FUMI advances discriminative dictionary learning by (1) addressing multiple instance learning problems and (2) using a shared background model resulting in improved target characterization and discrimination.

### A. Multiple-instance learning (MIL):

MIL [2] is a variation on supervised learning for problems with inaccurate label information. In particular, training data is segmented into positive and negative *bags*. A bag is defined to be a multi-set of data points. In the case of target detection, the MIL problem requires that a positive bag contains at least one instance from the target class and negative bags are composed of entirely non-target data. Given training data of this form, the overall goal can be to predict either unknown instance-level or bag-level labels on test data. MIL methods are effective for problems where accurately labeled training data is unavailable.

Most MIL approaches focus on learning a classification decision boundary to distinguish between positive and negative instances/bags [3], [4]. Although these decision boundary approaches are effective at training classifiers given inaccurate labels, they do not provide an intuitive description or *representative concept* that characterizes the salient and discriminative features of the target class. The approaches that estimate target representatives [5], [6], [7] often only find a single target concept and are, thus, unable to account for large variation in the target class. To address this, DL-FUMI learns a set of target atoms (and background atoms) to characterize target variation.

### B. Supervised Dictionary Learning:

Sparse coding refers to the task of decomposing a signal into a sparse linear combination of dictionary atoms [8], [9]. Of particular relevance are supervised (i.e., task-driven or discriminative) dictionary learning methods [10], [11]. However, among supervised dictionary learning methods, there are only a few approaches that address the problem given inaccurate MIL labels. These include MMDL [12] that trains many linear SVM classifiers and views the estimated parameters as dictionary atoms and DMIL [13], [14] that learns class-specific

$$F = \frac{1}{2}\sum_{i=1}^{N} w_i \left\| (\mathbf{x}_i - z_i \sum_{t=1}^{T} \alpha_{it}\mathbf{d}_t^+ - \sum_{k=1}^{M} \alpha_{ik}\mathbf{d}_k^-) \right\|_2^2 + \lambda \sum_{i=1}^{N} w_i \left\| \begin{bmatrix} z_i \boldsymbol{\alpha}_i^+ \\ \boldsymbol{\alpha}_i^- \end{bmatrix} \right\|_1 + \sum_{k=1}^{M}\sum_{t=1}^{T} \gamma_{kt} \langle \mathbf{d}_k^-, \mathbf{d}_{t_{\text{old}}}^+ \rangle \tag{3}$$

$$E[F] = \sum_{z_i \in \{0,1\}} P(z_i | \mathbf{x}_i, \boldsymbol{\theta}^{(l-1)}) \left[ \frac{1}{2}\sum_{i=1}^{N} w_i \left\| \mathbf{x}_i - z_i \sum_{t=1}^{T} \alpha_{it}\mathbf{d}_t^+ - \sum_{k=1}^{M} \alpha_{ik}\mathbf{d}_k^- \right\|_2^2 + \lambda \sum_{i=1}^{N} w_i \left\| \begin{bmatrix} z_i \boldsymbol{\alpha}_i^+ \\ \boldsymbol{\alpha}_i^- \end{bmatrix} \right\|_1 \right] + \sum_{k=1}^{M}\sum_{t=1}^{T} \gamma_{kt} \langle \mathbf{d}_k^-, \mathbf{d}_{t_{\text{old}}}^+ \rangle \tag{4}$$

dictionaries by maximizing the noisy-OR model in such a way that the all negative instances are poorly represented by the estimated target dictionary. As outlined in Sec I, DL-FUMI is unique from these existing methods through the use of a shared background dictionary.

## II. DL-FUMI

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ be training data where $d$ is the dimensionality of an instance, $\mathbf{x}_i$, and $N$ is the total number of training instances. The data is grouped into $K$ *bags*, $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$, with associated binary bag-level labels, $L = \{L_1, \ldots, L_K\}$ where $L_j \in \{0, 1\}$ and $\mathbf{x}_{ji} \in \mathbf{B}_j$ denotes the $i^{th}$ instance in bag $\mathbf{B}_j$. Given training data in this form, DL-FUMI models each instance as a sparse linear combination of target and/or background atoms $\mathbf{D}$, $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i$ is the sparse vector of weights for instance $i$. Positive bags (*i.e.*, $\mathbf{B}_j$ with $L_j = 1$, denoted as $\mathbf{B}_j^+$) contain at least one instance composed of some target:

$$\text{if } L_j = 1, \exists \mathbf{x}_i \in \mathbf{B}_j^+ \text{ s.t.}$$
$$\mathbf{x}_i = \sum_{t=1}^{T} \alpha_{it}\mathbf{d}_t^+ + \sum_{k=1}^{M} \alpha_{ik}\mathbf{d}_k^- + \boldsymbol{\varepsilon}_i, \alpha_{it} \neq 0, \tag{1}$$

where $\boldsymbol{\varepsilon}_i$ is a noise term. However, the number of instances in a positive bag with a target component is unknown.

If $\mathbf{B}_j$ is a negative bag (*i.e.*, $L_j = 0$, denoted as $\mathbf{B}_j^-$), then this indicates that $\mathbf{B}_j^-$ does not contain any target:

$$\text{if } L_j = 0, \forall \mathbf{x}_i \in \mathbf{B}_j^-, \mathbf{x}_i = \sum_{k=1}^{M} \alpha_{ik}\mathbf{d}_k^- + \boldsymbol{\varepsilon}_i \tag{2}$$

Given this problem formulation, the goal of DL-FUMI is to estimate the dictionary[1] $\mathbf{D} = \begin{bmatrix} \mathbf{D}^+ & \mathbf{D}^- \end{bmatrix} \in \mathbb{R}^{d \times (T+M)}$, where $\mathbf{D}^+ = \begin{bmatrix} \mathbf{d}_1^+, \cdots, \mathbf{d}_T^+ \end{bmatrix}$ are the $T$ target atoms and $\mathbf{D}^- = \begin{bmatrix} \mathbf{d}_1^-, \cdots, \mathbf{d}_M^- \end{bmatrix}$ are the $M$ background atoms. This is accomplished by minimizing (3) which is proportional to the complete negative data log-likelihood, where $\boldsymbol{\alpha}_i^{l+}$ and $\boldsymbol{\alpha}_i^{l-}$ are subsets of $\boldsymbol{\alpha}_i$ corresponding to $\mathbf{D}^+$ and $\mathbf{D}^-$, respectively. The first term in (3) computes the squared residual error between each instance and its estimate using the dictionary. In this term, a set of hidden binary latent variables $\{z_i\}_{i=1}^N$ that indicate whether an instance is or is not a target (*i.e.*, $z_i = 1$ when $\mathbf{x}_i$ contains target) are introduced. For all points in negative bags, $z_i = 0$. For points in positive bags, the value of $z_i$ is unknown. Also, a weight $w_i$ is included where $w_i = 1$ if $\mathbf{x}_i \in \mathbf{B}_j^-$ and $w_i = \psi$ if $\mathbf{x}_i \in \mathbf{B}_j^+$ where $\psi$ is a fixed parameter. This weight helps balance terms when there is a large imbalance between the number of negative and positive instances.

The second term is an $l_1$ regularization term to promote sparse weights. It also includes the latent variables, $z_i$, to account for the uncertain presence of target in positive bags.

The third term is a robust penalty term that promotes discriminative target atoms (and inspired by a term presented in [15]). Instead of using a fixed penalty coefficient, we introduce an adaptive coefficient $\gamma_{kt}$ defined as:

$$\gamma_{kt} = \Gamma \frac{\langle \mathbf{d}_k^-, \mathbf{d}_t^+ \rangle}{\|\mathbf{d}_k^-\| \|\mathbf{d}_t^+\|} = \Gamma \cos\theta_{kt}, \tag{5}$$

where $\theta_{kt}$ is the vector angle between the $k^{th}$ background atom and the $t^{th}$ target atom. Since $sign(\gamma_{kt}) = sign(\langle \mathbf{d}_k^-, \mathbf{d}_t^+ \rangle)$, this discriminative term is always positive and will add large penalty when $\mathbf{d}_k^-$ and $\mathbf{d}_t^+$ have similar shape. Thus, this term encourages a discriminative dictionary by promoting background atoms that are orthogonal to target atoms. In implementation, $\gamma_{kt}$ is updated once per iteration using $\mathbf{d}_{k_{old}}^-$ and $\mathbf{d}_{t_{old}}^+$ which are the dictionary values from the previous iteration.

## III. DL-FUMI Optimization

Expectation-Maximization is used to optimize (3) and estimate $\mathbf{D}$. During optimization, the fact that many of the binary latent variables $\{z_i\}_{i=1}^N$ are unknown is addressed by taking the expected value of the log likelihood with respect to $z_i$ as shown in (4). In (4), $\boldsymbol{\theta}^l = \left\{\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^N\right\}$is the set of parameters estimated at iteration $l$ and $P(z_i|\mathbf{x}_i, \boldsymbol{\theta}^{(l-1)})$ is the probability that each instance is or is not a true target instance. During the E-step of each iteration, $P(z_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(l-1)})$ is computed as:

$$P(z_i|\mathbf{x}_i, \boldsymbol{\theta}^{(l-1)}) = \begin{cases} e^{-\beta \left\| \mathbf{x}_i - \sum_{k=1}^{M} \alpha_{ik}\mathbf{d}_k^- \right\|_2^2} & \text{if } z_i = 0, L_j = 1 \\ 1 - e^{-\beta \left\| \mathbf{x}_i - \sum_{k=1}^{M} \alpha_{ik}\mathbf{d}_k^- \right\|_2^2} & \text{if } z_i = 1, L_j = 1 \\ 0 & \text{if } z_i = 1, L_j = 0 \\ 1 & \text{if } z_i = 0, L_j = 0 \end{cases} \tag{6}$$

where $\beta$ is a fixed scaling parameter. If $\mathbf{x}_i$ is a non-target instance, then it should be characterized by the background atoms well, thus $P(z_i = 0|\mathbf{x}_i, \boldsymbol{\theta}^{(l-1)}) \approx 1$. Otherwise, if $\mathbf{x}_i$ is a true target instance, it will not be characterized well using only the background atoms and $P(z_i = 1|\mathbf{x}_i, \boldsymbol{\theta}^{(l-1)}) \approx 1$.

[1] $\begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}$ are the concatenation of arrays $\mathbf{A}$ and $\mathbf{B}$ horizontally and vertically, respectively.

**Algorithm 1** DL-FUMI EM algorithm

```
Initialize $\boldsymbol{\theta}^0 = \left\{\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^N\right\}$, $l = 1$
repeat
    E-step: Compute $P(z_i|\mathbf{x}_i, \boldsymbol{\theta}^{(l-1)})$
    M-step:
    Update $\mathbf{d}_t^+$ using (9), $\mathbf{d}_t^+ \leftarrow \frac{1}{\|\mathbf{d}_t^+\|_2}\mathbf{d}_t^+, t = 1, \cdots, T$
    Update $\mathbf{d}_k^-$ using (8), $\mathbf{d}_k^- \leftarrow \frac{1}{\|\mathbf{d}_k^-\|_2}\mathbf{d}_k^-, k = 1, \cdots, M$
    for $q \leftarrow 1$ to $iter$ do
      Update $\{\boldsymbol{\alpha}_i\}_{i=1}^{N^+}$ for $\mathbf{x}_i \in \mathbf{B}_j^+$ using (11), (12)
      Update $\{\boldsymbol{\alpha}_i\}_{i=1}^{N^-}$ for $\mathbf{x}_i \in \mathbf{B}_j^-$ using (13)
    end for
    $l \leftarrow l + 1$
until Convergence
return D, $\{\boldsymbol{\alpha}_i\}_{i=1}^N$
```

*DL-FUMI code can be found at: https://github.com/TigerSense/FUMI

The *M-step* is performed by iteratively optimizing (4) for each of the desired parameters. The dictionary $\mathbf{D}$ is updated atom-by-atom using a block coordinate descent scheme [16], [17]. The sparse weights, $\{\boldsymbol{\alpha}_i\}_{i=1}^N$, are updated using an iterative shrinkage-thresholding algorithm [18], [19]. For readability, the derivation of update equations are described in Sec. VII. The method is summarized in Alg. 1.

## IV. Classification using Estimated Dictionary

Given $\mathbf{D}$, a confidence that the $i^{th}$ instance is target can be computed using a ratio of the reconstruction errors given the target and background atoms, $\mathbf{D}$, vs. background atoms, $\mathbf{D}^-$:

$$c_i = \frac{\left\|\mathbf{x}_i - \boldsymbol{\alpha}_i^- \mathbf{D}^-\right\|^2}{\left\|\mathbf{x}_i - \boldsymbol{\alpha}_i \mathbf{D}\right\|^2}, \tag{7}$$

where $\boldsymbol{\alpha}_i^-$ are the sparse weights for the non-target atoms for the $i^{th}$ instance. If the numerator has a large error and the denominator has a low error, then the target atoms are needed to reconstruct instance $i$.

## V. Experiments

DL-FUMI is evaluated on two MIL AR Face data [20], [21] recognition problems and an MIL USPS hand-written digits [22], [23] recognition problem. In all of our experiments, target atoms were initialized by computing mean of $T$ random subsets drawn from the union of all positive bags. $K$-means was applied to the union of all negative bags and the $M$ cluster centers were set as the initial background atoms.

### A. AR Face Recognition

The AR-face data set consists of frontal-pose images with 26 images/person (2 sessions, 13 per session) corresponding to different expressions, illuminations and occlusions. Pre-processed and cropped imagery of 50 male and 50 female subjects provided by Martinez and Kak [21] was used. Each image was down-sampled to 83 × 60 pixels and the raw gray scale values were used as features.

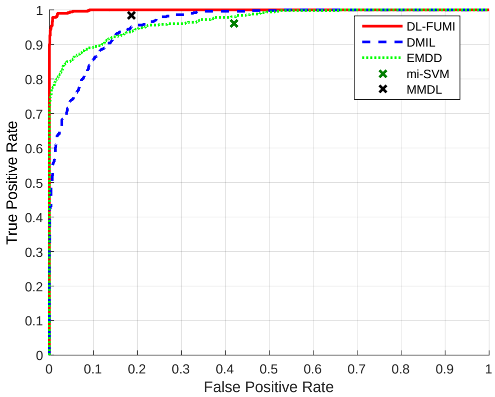


Fig. 1. ROC analysis for sun-glasses detection using AR face database comparing DL-FUMI, DMIL [13] and EM-DD [6] (code from [24]). True Positive Rate vs. False Positive Rate of mi-SVM [3] and MMDL [12] (code from author's website) are also plotted.

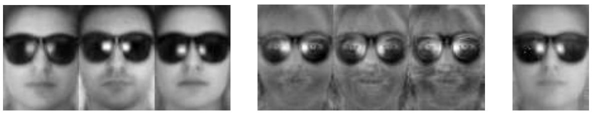
For the first AR Face experiment, sun-glasses were the target concept. Specifically, 50 positive training bags of 10 instances each were created. Each positive bag contained only two instances of randomly selected images of people wearing sun-glasses; the other eight were randomly chosen from images of people without sun-glasses. 50 negative bags were constructed by randomly selecting 10 instances per bag of images of individuals not wearing sun-glasses. Test data included all imagery that was not used for training.

The parameters for DL-FUMI for this experiment were set to $T = 3$, $M = 8$, $\Gamma = 0.001$, $\beta = 30$ and $\lambda = 0.001$. After dictionary estimation, the target confidence was computed for each test instance following Sec. IV. Receiver operating characteristic (ROC) curve analysis was conducted. Fig. 1 shows one of the 10 ROCs obtained by DL-FUMI, DMIL and EM-DD where the TPR vs FPR obtained by mi-SVM and MMDL were also plotted. The average TPRs of DL-FUMI, EM-DD and DMIL over 10 runs are shown in Table I at FPRs 1%, 18.4% and 41.9%, where 18.4% and 41.9% are average FPRs by two classification algorithms MMDL and mi-SVM, respectively. Fig. 2 and Fig. 3 show estimated target and background atoms by DL-FUMI and comparison methods, respectively. To estimate the DMIL background atoms, we flipped the sign of positive and negative bags (*i.e.*, swapped the target and background classes) and re-trained the dictionary. This was done since, as stated in [14], DMIL does not learn a set of background atoms simultaneously. As shown, DL-FUMI target atoms are very discriminative and representative of the target class, *e.g.*, there are male and female sun-glasses atoms and variation in light reflections. Finally, the overall dictionary set estimated by DL-FUMI is qualitatively more smooth which will help to reduce error in classification.

For the second AR Face experiment, Woman No. 10 was selected as the positive target class. Two positive training bags containing 50 instances each were created. The first positive bag contained 6 images from Woman No. 10 set 1 and the second positive bag contained the remaining 7

TABLE I
AVERAGE TPR AT FPRS OVER 10 RUNS

| Algorithm | TPR(%)<br>FPR=1% | TPR(%)<br>FPR=18.4% | TPR(%)<br>FPR=41.9% |
|---|---|---|---|
| mi-SVM [3] | - | - | 96.2 |
| EM-DD [6] | 78.2 | 93.8 | 98.1 |
| MMDL [12] | - | 98.0 | - |
| DMIL [13] | 60.2 | 95.2 | 99.5 |
| **DL-FUMI** | 97.5 | 100 | 100 |

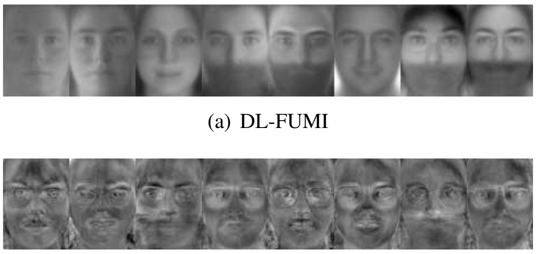


(a) DL-FUMI (b) DMIL (c) EMDD

Fig. 2. Plot of estimated dictionary atoms for sun-glasses. (a): DL-FUMI. (b): DMIL. (c): EM-DD.

images from Woman No. 10 set 1, and the rest of the instances in each positive bag were randomly selected from other individuals. Three negative bags with 200 instances per bag were constructed by randomly selecting images from the data set excluding those from Woman No. 10. Given this, there are only 13 positive training instances and 687 negative training instances. This is a more difficult problem than sunglass detection. The test data contained the 13 images of Woman No. 10 from set 2 and 100 images randomly selected from images that are not Woman No. 10. There is no overlap between the training and testing data.

The parameters used in DL-FUMI for this experiment were $T = 3$, $M = 20$, $\Gamma = 0.001$, $\beta = 50$ and $\lambda = 0.001$. One of 10 ROCs is shown in Fig. 4. The average TPRs of DL-FUMI, EM-DD and DMIL are shown in Table II at FPRs 2.9%, 5% and 12.0%, where 2.9% and 12.0% are average FPRs by two classification algorithms MMDL and mi-SVM, respectively. Table II and Fig. 4 clearly show that DL-FUMI outperforms all the comparison algorithms. In order to further show that the estimated target dictionary by DL-FUMI is effective at characterizing the target class, the subspace adaptive cosine estimator (subACE) target detection algorithm [25] was applied for detection using the target dictionary estimated by DL-FUMI directly. One of the subACE ROCs using the DL-FUMI dictionary shows a 100% TPR with 0% FPR in Fig. 4. Since subACE is a target detection algorithm that relies on target signatures that encompass the distinguishing characteristics of a target class, this further emphasizes that target dictionary estimated by DL-FUMI is highly representative of the target class. Fig. 5(a) - 5(c) show the target atoms estimated by DL-FUMI, DMIL and EM-DD for Woman No. 10, where it can be seen that the target dictionary atoms estimated by DL-FUMI are more discriminative as it captures different distinct features of the positive class (different occlusions, expressions, *etc.*). Fig. 6(a) and 6(b) show the background atoms estimated for Women No. 10 and it can be seen that the background dictionary estimated by DL-FUMI has better representative quality.



(a) DL-FUMI

(b) DMIL

Fig. 3. Plot of estimated dictionary atoms for background. (a): DL-FUMI. (b): DMIL.

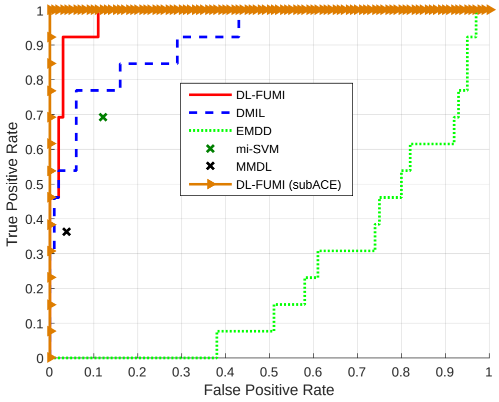
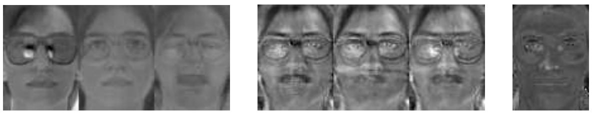
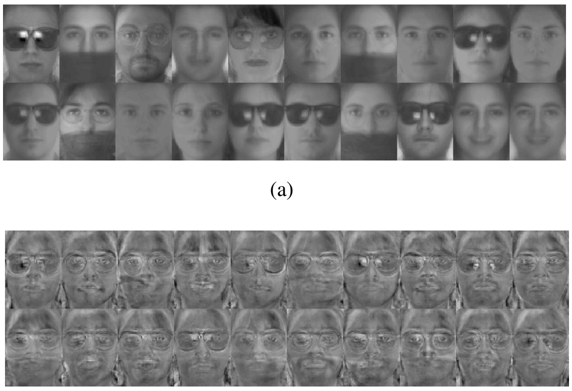


Fig. 4. Woman No. 10 detection on AR face database



(a) (b) (c)

Fig. 5. Plot of estimated dictionary atoms for Woman No. 10. (a): DL-FUMI. (b): DMIL. (c): EM-DD



(a)

(b)

Fig. 6. Plot of estimated background dictionary atoms for Woman No. 10.

TABLE II
AVERAGE TPR AT FPRS OVER 10 RUNS

| Algorithm | TPR(%)<br>FPR=2.9% | TPR(%)<br>FPR=5% | TPR(%)<br>FPR=12.0% |
|---|---|---|---|
| mi-SVM [3] | - | - | 69.2 |
| EM-DD [6] | 0 | 0 | 0 |
| MMDL [12] | 31.5 | - | - |
| DMIL [13] | 54.5 | 69.1 | 78.8 |
| **DL-FUMI** (using (7)) | 78.9 | 91.5 | 95.5 |
| **DL-FUMI** (subACE) | 94.6 | 100 | 100 |

### B. *USPS Digit Classification*

DL-FUMI is further evaluated on a multi-class classification task given the USPS[2] data set. The USPS data set contains 9298 images of handwritten digits from 0 to 9. Each image is $16 \times 16$ in size. The raw gray-level pixel values are used as features in this experiment. The training and testing data partitions in this paper mimics the experimental set-up in [14]. Specifically, for each class $c$, 50 positive training bags were generated. Each bag contains 4 instances in total and only one comes from true $c^{th}$ positive class and the other three instances are randomly chosen from other classes. 50 negatively labeled bags were also constructed by randomly selecting 50 instances per bag from classes other than $c$. The testing data contains 2000 samples in total, 200 from each class.

In this experiment, the parameters used were $T = 4$, $M = 15$, $\Gamma = 0.1$, $\beta = 25$ and $\lambda = 0.001$. For instance level classification, the approach described in Sec. IV was applied given a dictionary estimated given each class as the target class. Then, the final class label, for multi-class classification, was assigned by selecting the class with the largest confidence value computed in (7). The classification results of DL-FUMI and comparison algorithms are listed in Table III, where results for GD-MIL are as reported in [14]. Table III shows that DL-FUMI outperforms two multiple instance dictionary learning methods, GD-MIL and MMDL, and two MIL methods mi-SVM and EM-DD. Fig. 7(a) and 7(b) show the estimated DL-FUMI target and non-target dictionary atoms, in these we can see that DL-FUMI is able to learn a set of discriminative target dictionary as well as characteristic background dictionary (*i.e.*, each target dictionary atom looks like the target digit and the background dictionary atoms look like all the other digits).

To get insight into classification errors, Fig. 8(a) - 8(d) show several examples of randomly selected misclassified instances and Fig. 8(e) - 8(h) show the reconstructed images by DL-FUMI. For example, Fig. 8(a) has a true class label of 0, but was misclassified to 6; the reconstructed image is shown in Fig. 8(e). As can be seen, this data point appears to be very similar to the digit 6 and is even difficult for a human to correctly recognize. Similarly, Fig. 8(b) - 8(d) show the other three images and Fig. 8(f) - 8(h) show the corresponding reconstructed images, respectively.

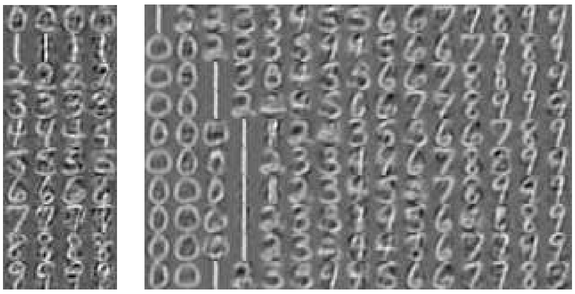
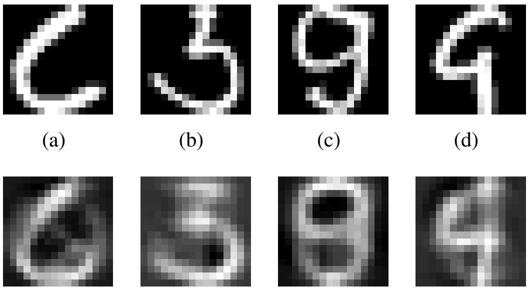

(a) (b)

Fig. 7. USPS dictionary atoms estimated by DL-FUMI. (a): Target atoms. (b): Non-target atoms.


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 8. Examples of misclassified images by DL-FUMI. (a): true class 0 misclassified to 6. (b): true class 5 misclassified to 3. (c): true class 9 misclassified to 8. (d): true class 9 misclassified to 4. (e)-(f): corresponding reconstructed images by DL-FUMI

TABLE III
CLASSIFICATION ACCURACIES ON USPS DATA SET

| Alg. | Acc.(%) | Alg. | Acc.(%) |
|---|---|---|---|
| mi-SVM [3] | 81.1 | GD-MIL[3] [14] | 83.4 |
| EM-DD [6] | 76.55 | MMDL [12] | 80.8 |
| **DL-FUMI** | **86.5** | | |

## VI. CONCLUSION

In this paper, a multiple-instance dictionary learning algorithm, DL-FUMI, is proposed. DL-FUMI leverages the *shared* information between positive and negative classes to improve the discriminative ability of the estimated target atoms. Experimental results show that the estimated DL-FUMI target atoms provide a good representation of the positive class and improves target detection and classification performance over comparison methods.

## VII. DERIVATION OF DL-FUMI UPDATE EQUATIONS

This section provides a derivation of DL-FUMI update equations. When updating the dictionary $\mathbf{D}$, the sparse weights $\{\boldsymbol{\alpha}_i\}_{i=1}^N$ are held fixed. To update one of the atoms in $\mathbf{D}$, (4) is minimized with respect to the corresponding atom while keeping all other atoms constant. The resulting update equations for $\mathbf{d}_t^+$ and $\mathbf{d}_k^-$ are shown in (9) and (8).

[2] Database at: http://www-i6.informatik.rwth-aachen.de/∼keysers/usps.html

[3] Results as stated in the literature [14].

$$\begin{aligned}\mathbf{d}_k^- \quad = \quad & \left\{\sum_{i=1}^{N^+}\left[P(z_i=1)\psi\alpha_{ik}(\mathbf{x}_i-\sum_{t=1}^{T}\alpha_{it}\mathbf{d}_t^+-\sum_{l=1,l\neq k}^{M}\alpha_{il}\mathbf{d}_l^-)+P(z_i=0)\psi\alpha_{ik}(\mathbf{x}_i-\sum_{l=1,l\neq k}^{M}\alpha_{il}\mathbf{d}_l^-)\right]+\sum_{i=1}^{N^-}\left[\alpha_{ik}(\mathbf{x}_i-\sum_{l=1,l\neq k}^{M}\alpha_{il}\mathbf{d}_l^-)\right]\right.\\ & \left.-\Gamma\sum_{t=1}^{T}\cos\theta_{kt}\mathbf{d}_{told}^+\right\}\left\{\sum_{i=1}^{N+}\psi\alpha_{ik}^2+\sum_{i=1}^{N-}\alpha_{ik}^2\right\}^{-1} \qquad (8)\end{aligned}$$

$$\mathbf{d}_t^+=\frac{\sum_{i=1}^{N+}\left[P(z_i=1)\alpha_{it}(\mathbf{x}_i-\sum_{l=1,l\neq t}^{T}\alpha_{il}\mathbf{d}_l^+-\sum_{k=1}^{M}\alpha_{ik}\mathbf{d}_k^-)\right]}{\sum_{i=1}^{N+}\left[P(z_i=1)\alpha_{it}^2\right]} \qquad (9)$$

Note, $P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{(t-1)})$ is denoted as $P(z_i)$ for simplicity.

When updating the sparse weights, $\{\boldsymbol{\alpha}_i\}_{i=1}^{N}$, it should be noted that the sparse weight vector $\boldsymbol{\alpha}_i$ for instance $\mathbf{x}_i$ is not dependent on any other instances.

The gradient with respect to $\boldsymbol{\alpha}_i$ without considering the $l_1$ penalty term is:

$$\begin{aligned}\frac{\partial F^+}{\partial\boldsymbol{\alpha}_i}=&-\left[P(z_i=1)\mathbf{D}^+\quad\mathbf{D}^-\right]^T\mathbf{x}_i+\Big(P(z_i=1)\mathbf{D}^T\mathbf{D}\\ &+P(z_i=0)\left[\mathbf{0}_{d\times T}\quad\mathbf{D}^-\right]^T\left[\mathbf{0}_{d\times T}\quad\mathbf{D}^-\right]\Big)\boldsymbol{\alpha}_i. \qquad (10)\end{aligned}$$

Then $\boldsymbol{\alpha}_i$ at $l^{th}$ iteration can be updated using gradient descent,

$$\boldsymbol{\alpha}_i^l=\boldsymbol{\alpha}_i^{l-1}-\eta_i\frac{\partial F^+}{\partial\boldsymbol{\alpha}_i}, \qquad (11)$$

followed by a soft-thresholding:

$$\left\{\begin{array}{l}\boldsymbol{\alpha}_i^{l+}=S_{\lambda P(z_i=1)}\left(\boldsymbol{\alpha}_i^{l+}\right)\\ \boldsymbol{\alpha}_i^{l-}=S_{\lambda}\left(\boldsymbol{\alpha}_i^{l-}\right)\end{array}\right., \qquad (12)$$

s.t. $S_\lambda\left(\mathbf{x}[i]\right)=sign(\mathbf{x}[i])\max(|\mathbf{x}[i]|-\lambda,0)$, $i=1,...,d$.

Following a similar proof to that in [26], when $\eta_i\in\left(0,\left(\lambda_{max}\left(P(z_i=0)\left[\mathbf{0}_{d\times T}\ \mathbf{D}^-\right]^T\left[\mathbf{0}_{d\times T}\ \mathbf{D}^-\right]+P(z_i=1)\mathbf{D}^T\mathbf{D}\right)\right)^{-1}\right)$, the update of $\boldsymbol{\alpha}_i$ using a gradient descent method with step length $\eta_i$ monotonically decreases the value of the objective function, where $\lambda_{max}(\mathbf{A})$ denotes the maximum eigenvalue of $\mathbf{A}$. For simplicity, $\eta$ was set as $\eta=\frac{1}{\lambda_{max}(\mathbf{D}^T\mathbf{D})}$ for all $\boldsymbol{\alpha}_i$, $\mathbf{x}_i\in\mathbf{B}_j^+$.

A similar update can be used for points from negative bags. The resulting update equation for negative points is:

$$\boldsymbol{\alpha}_i^l=S_\lambda\left(\boldsymbol{\alpha}_i^{l-1}+\frac{1}{\lambda_{max}\left(\mathbf{D}^{-T}\mathbf{D}^-\right)}\left(\mathbf{D}^{-T}(\mathbf{x}_i-\mathbf{D}^-\boldsymbol{\alpha}_i^{l-1})\right)\right) \qquad (13)$$

The sparse weights corresponding to target dictionary atoms are set to 0 for all points in negative bags.

## References

[1] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Found. and Trends in Comput. Graph. and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles," *Artificial Intell.*, vol. 89, no. 1-2, pp. 31–17, 1997.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Inf. Process. Syst.*, 2002, pp. 561–568.

[4] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 28, no. 12, pp. 1931–1947, 2006.

[5] O. Maron and T. Lozano-Perez, "A framework for multiple-instance learning." in *Advances in Neural Inf. Process. Syst.*, vol. 10, 1998.

[6] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Advances in Neural Inf. Process. Syst.*, vol. 2. MIT; 1998, 2002, pp. 1073–1080.

[7] C. Jiao and A. Zare, "Functions of multiple instances for learning target signatures," *IEEE Trans. on Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4670 – 4686, 2015.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[10] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, 2012.

[11] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, 2013.

[12] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," in *Int. Conf. On Mach. Learning*, 2013, pp. 846–854.

[13] A. Shrivastava, J. K. Pillai, V. M. Patel, and R. Chellappa, "Dictionary-based multiple instance learning," in *IEEE Int. Conf. on Image Process.*, 2014, pp. 160–164.

[14] A. Shrivastava, V. M. Patel, j. K. Pillai, and R. Chellappa, "Generalized dictionaries for multiple instance learning," *Int. J. of Comput. Vision*, vol. 114, no. 2, pp. 288–305, Septmber 2015.

[15] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Comput. Vision and Pattern Recognition*, 2010, pp. 3501–3508.

[16] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. of Mach. Learning Research*, vol. 11, pp. 19–60, 2010.

[18] M. A. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, 2003.

[19] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. on Pure and Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[20] A. M. Martínez, "The AR face database," *CVC Tech. Rep.*, vol. 24, 1998.

[21] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, 2001.

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[23] P. D. Gader and M. A. Khabou, "Automatic feature generation for handwritten digit recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1256–1261, 1996.

[24] Z.-H. Zhou and M.-L. Zhang, "Ensembles of multi-instance learners," in *European Conf. on Mach. Learning*. Springer, 2003, pp. 492–502.

[25] S. Kraut, L. Scharf, and L. McWhorter, "Adaptive subspace detectors," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 1–16, 2001.

[26] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

# APPENDIX

In this appendix, the relationship between two forms of the discriminative terms are examined. The two forms considered are:

$$\sum_{k=1}^{M}\sum_{t=1}^{T}\frac{\Gamma}{2}\cos\theta_{kt}\|\mathbf{d}_k^- + \mathbf{d}_{t_{\text{old}}}^+|\|_2^2 \tag{A.1}$$

$$\sum_{k=1}^{M}\sum_{t=1}^{T}\Gamma\cos\theta_{kt}\langle\mathbf{d}_k^-, \mathbf{d}_{t_{\text{old}}}^+\rangle \tag{A.2}$$

The update equations for the dictionary atoms given these two versions of the discriminative terms are identical after normalization. Thus, there is no difference in the results obtained using either term. This relationship is shown as follows.

The expectation of the objective function with the discriminative term in (A.1) is shown in (A.3):

$$\begin{aligned}E[F] = \sum_{z_i\in\{0,1\}} P(z_i)\Bigg[\frac{1}{2}\sum_{i=1}^{N} w_i\left\|\mathbf{x}_i - z_i\sum_{t=1}^{T}\alpha_{it}\mathbf{d}_t^+ - \sum_{k=1}^{M}\alpha_{ik}\mathbf{d}_k^-\right\|_2^2\\ +\lambda\sum_{i=1}^{N} w_i\left\|\begin{bmatrix} z_i\boldsymbol{\alpha}_i^+\\ \boldsymbol{\alpha}_i^-\end{bmatrix}\right\|_1\Bigg] + \sum_{k=1}^{M}\sum_{t=1}^{T}\frac{\Gamma}{2}\cos\theta_{kt}\|\mathbf{d}_k^- + \mathbf{d}_{t_{\text{old}}}^+|\|_2^2\end{aligned} \tag{A.3}$$

To update one of the atoms in $\mathbf{D}$, (A.3) is minimized with respect to the corresponding atom while keeping all other atoms constant, *i.e.*, $\mathbf{d}_t^+ \leftarrow \text{argmin}_{\mathbf{d}_t^+\in\mathbb{R}^d, \|\mathbf{d}_t^+\|_2=1} E[F]$ and $\mathbf{d}_k^- \leftarrow \text{argmax}_{\mathbf{d}_k^-\in\mathbb{R}^d, \|\mathbf{d}_k^-\|_2=1} E[F]$. The resulting update equations of $\mathbf{d}_t^-$ is exactly the same as (9) since $\mathbf{d}_t^-$ is viewed as a constant in the robust terms. To determine the update equation of $\mathbf{d}_k^-$ with the discriminative term shown in (A.1), take the partial derivative of (A.3) with respect to $\mathbf{d}_k^-$:

$$\begin{aligned}\frac{\partial E[F]}{\partial\mathbf{d}_k^-} = \sum_{i=1}^{N^+}\Bigg[-P(z_i=0)\psi\alpha_{ik}\left(\mathbf{x}_i - \sum_{k=1}^{M}\alpha_{ik}\mathbf{d}_k^-\right)\\ -P(z_i=1)\psi\alpha_{ik}\left(\mathbf{x}_i - \sum_{t=1}^{T}\alpha_{it}\mathbf{d}_t^+ - \sum_{k=1}^{M}\alpha_{ik}\mathbf{d}_k^-\right)\Bigg]\\ +\sum_{i=1}^{N^-}\left[-\alpha_{ik}\left(\mathbf{x}_i - \sum_{k=1}^{M}\alpha_{ik}\mathbf{d}_k^-\right)\right] + \Gamma\sum_{t=1}^{T}\cos\theta_{kt}\left(\mathbf{d}_k^- + \mathbf{d}_{t_{\text{old}}}^+\right)\end{aligned} \tag{A.4}$$

After setting (A.4) to zero and solving for $\mathbf{d}_k^-$, (A.5) is obtained as update equation for $\mathbf{d}_k^-$.

If we compare (A.5) with (8), it can be clearly seen that the only difference between the two update equations is a constant normalization scaling term found in the denominator (*i.e.*, $\{\cdot\}^{-1}$). This means that updated $\mathbf{d}_k^-$ by the two update equations, (A.5) and (8), has the same shape but different scale. However, since we add a norm 1 constraint on each dictionary atom estimated, $\mathbf{d}_k^-$ will be normalized in Step 6 Alg. **??**, $\mathbf{d}_k^- \leftarrow \frac{1}{\|\mathbf{d}_k^-\|_2}\mathbf{d}_k^-, k = 1,\cdots,M$. This ensures that the update of $\mathbf{d}_k^-$ is identical with either discriminative term.

$$\begin{aligned}\mathbf{d}_k^- \quad = \quad & \Bigg\{\sum_{i=1}^{N^+}\Bigg[P(z_i=1)\psi\alpha_{ik}(\mathbf{x}_i - \sum_{t=1}^{T}\alpha_{it}\mathbf{d}_t^+ - \sum_{l=1,l\neq k}^{M}\alpha_{il}\mathbf{d}_l^-) + P(z_i=0)\psi\alpha_{ik}(\mathbf{x}_i - \sum_{l=1,l\neq k}^{M}\alpha_{il}\mathbf{d}_l^-)\Bigg] + \sum_{i=1}^{N^-}\Bigg[\alpha_{ik}(\mathbf{x}_i - \sum_{l=1,l\neq k}^{M}\alpha_{il}\mathbf{d}_l^-)\Bigg]\\ & -\Gamma\sum_{t=1}^{T}\cos\theta_{kt}\mathbf{d}_{t_{old}}^+\Bigg\}\left\{\sum_{i=1}^{N+}\psi\alpha_{ik}^2 + \sum_{i=1}^{N-}\alpha_{ik}^2 + \Gamma\sum_{t=1}^{T}\cos\theta_{kt}\right\}^{-1}\end{aligned} \tag{A.5}$$