Matrix Completion for Weakly-supervised Multi-label Image Classification

Ricardo Cabral, Fernando De la Torre, João P. Costeira, Alexandre Bernardino

***Abstract*—In the last few years, image classification has become an incredibly active research topic, with widespread applications. Most methods for visual recognition are fully supervised in nature, as they make use of bounding boxes or pixelwise segmentations to locate objects of interest. However, this type of manual labeling is typically time consuming and error prone. This paper proposes a weakly-supervised system for multi-label image classification. In this setting, training images are annotated with a set of keywords describing their contents, but the visual concepts are not explicitly segmented in the images.**

We formulate the weakly-supervised image classification as a matrix completion problem. Compared to previous work, our proposed framework has three advantages: (1) Unlike existing solutions based on multiple-instance learning methods, our model is convex. We propose two alternative algorithms for matrix completion based on a rank minimization criterion specifically tailored to visual data, and prove their convergence. (2) Unlike existing discriminative methods, our algorithm is robust to labeling errors, background noise and partial occlusions. (3) Our method can potentially be used for semantic segmentation. Experimental validation on several datasets shows that our method outperforms state-of-the-art classification algorithms, while effectively capturing each class appearance.

***Index Terms*—Weakly-supervised learning, multi-label image classification, segmentation, rank minimization, nuclear norm.**

I. Introduction

With the ever-growing amount of digital image data in multimedia databases, there is a great need for algorithms that can provide effective semantic indexing. Attributing keywords to digital images, however, is the quintessential example of a challenging classification problem. Several aspects contribute to the difficulty of this problem, including the large variability in appearance, illumination and pose any object class can exhibit. To add to this complexity, natural images often depict a multitude of objects rather than a single one. In this multi-label setting, the interaction between objects needs to be modeled so classifiers are able to discern between co-occurring concepts. To address this issue, standard discriminative approaches such as Support Vector Machines (SVMs) or Linear Discriminant Analysis have been extended to the multi-label case [1]. A major limitation of these approaches, however, is that the location for objects of interest has to be known in the training images, usually in the form of bounding boxes or a full-blown pixelwise segmentation. While efforts have been made to provide datasets with this information [2], [3], manual labeling is still labor intensive, subjective and error prone. Moreover, it has been shown that manual segmentations are not necessarily the optimal spatial enclosure for object classifiers [4]. To cope with an increasing number of concepts and larger scale datasets, there has been an increased interest in transitioning away from these fully supervised approaches.

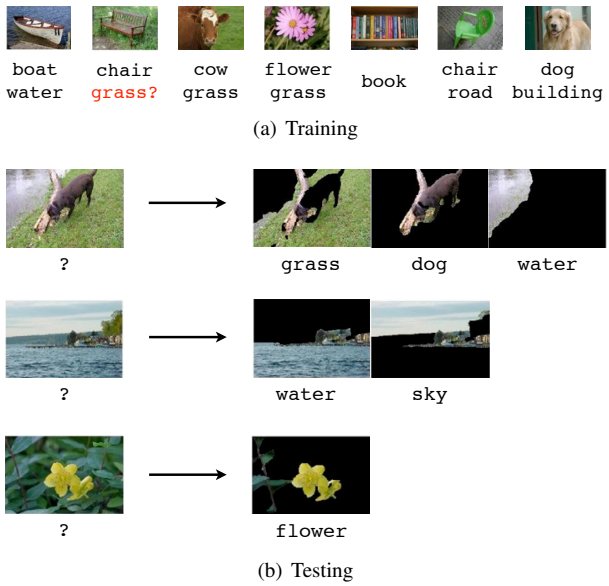


Fig. 1. This work proposes a weakly-supervised method for multi-label image classification. The training set images (a) are labeled with the objects that are present but their location in the image is unknown. Given unseen test images (b), our method is able to classify which classes are present in the image and segment the image into regions that correspond to the classes.

Weakly-supervised algorithms [4]–[7] relieve the labeling burden by learning using labels without localization information. Figure 1 illustrates this setting and the problem we address in this paper: given a weakly-labeled training set (Figure 1(a)), we segment and label new test images (Figure 1(b)). Several Multiple Instance Learning (MIL) methods [4], [8]–[13] have been proposed in the literature for solving this type of weakly supervised learning problem. However, existing MIL methods have three major drawbacks: (1) The MIL problem is usually cast as a NP-hard binary quadratic problem [4], [8]–[12]. Most existing algorithms to solve MIL lead to non-convex models and consequently are highly sensible to initialization. Moreover, the extension of MIL to the multi-label case is not trivial. Current multi-label MIL approaches [9]–[11] heavily rely on an explicit enumeration of instances, which are then solved by single class MIL or Multi-label learning. (2) They lack robustness to outliers. Recall that most

Ricardo Cabral is with the ECE Department, Carnegie Mellon University, USA and with the ISR, Instituto Superior Técnico, Portugal. E-mail: rscabral@cmu.edu. Fernando De la Torre is with the Robotics Institute, Carnegie Mellon University, USA. E-mail: ftorre@cs.cmu.edu João Paulo Costeira and Alexandre Bernardino are with the ISR, Instituto Superior Técnico, Portugal. E-mail: {jpc,alex}@isr.ist.utl.pt

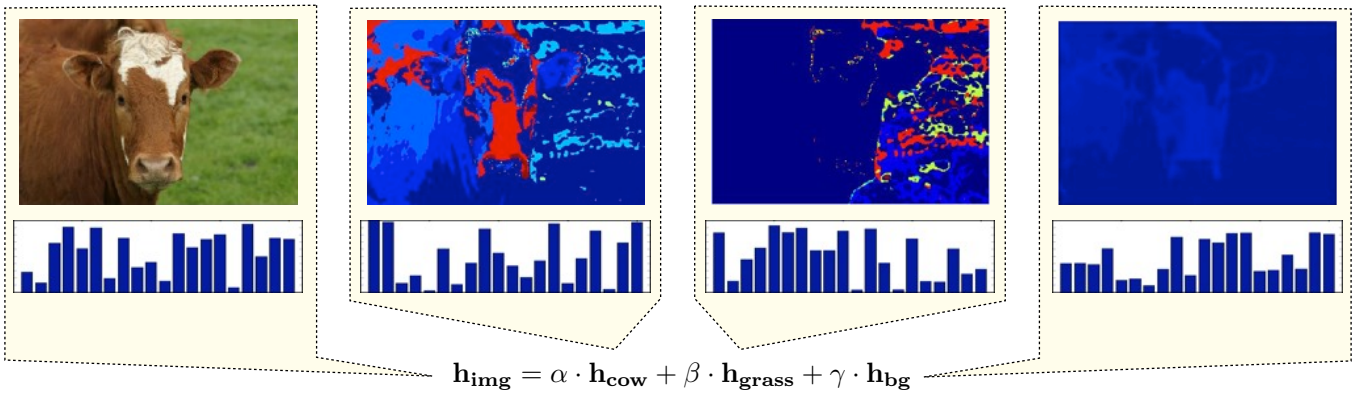


Fig. 2. The left image represents the original training image that has been labeled with the words grass and cow. Our algorithm decomposes the histogram of this image as a linear combination of two class histogram basis (cow, grass) plus another histogram $\mathbf{h_{bg}}$ modeling errors and the background. Class localization can be visualized on the image by interpreting each histogram as a probability distribution of which words belong to the class. Best seen in color.

discriminative approaches project data directly onto linear or non-linear spaces [4], [8], [12]. Thus, a single outlier can bias the solution, severely degrading classification performance. (3) It is unclear how existing MIL approaches can be extended to use partial information, such as incomplete label assignments or missing feature descriptions. For instance, in Figure 1(a) one training image has no label for the category grass.

We observe that the specific problem of image classification and localization problem has more structure than the one exploited by MIL problems. MIL approaches consider images as bags with many instances denoting possible regions of interest. A major contribution of our work is to instead make use of the additive property of histogram representations such as Bag of Words (BOW) [14]: the histogram of an entire image is a weighted sum of the histogram information of all of its subparts. Using this property, we bypass the combinatorial nature of finding desired regions in every positive image. Instead, the main aim of our algorithm is to factorize the histogram of an image as a weighted sum of class histograms (as many as objects are present) plus an error to model the background. Figure 2 shows an illustration of the histogram factorization for one training image. By using this property, image classification can be posed as a rank minimization problem, since class histograms are shared across images, and the number of class histograms is small compared to the number of images. This paper proposes to cast the weakly supervised multi-label image classification problem under a matrix completion framework. Contrary to typical MIL approaches, our proposed approach is fueled by recent advances in rank minimization [15], [16] and therefore is convex. Figure 3 illustrates the main idea of the paper. Each column of $\mathbf{Z}^{obs}$ has a concatenation of the labels (1 if the class is present and zero otherwise) and the histogram $\mathbf{h}_i^{tr}$ for one training image (Figure 3 (a)). In the test set (Figure 3 (b)), labels are unknown and denoted as question marks $(?)$. Our method fills the unknown entries and corrects known features and labels such that $\mathbf{Z}$ has the smallest rank possible. It can also infer the feature descriptor of a particular class (Figure 3 (c)). This is achieved by looking for the unknown histograms whose label vector denotes the presence of only this class. In doing so, we obviate the need for training with precise localization or expensive combinatorial MIL models, as required by previous methods. To summarize, the main contributions of this work are threefold:

1) We show the advantages of matrix completion over classic discriminative approaches for image classification. By performing classification under this inherently multi-label paradigm, we can easily cope with missing information as well as outliers in both the feature and the label space. We present comparisons on several datasets that show how these properties lead to a classification improvement over state-of-the-art methods;
2) We exploit the additive nature of histogram features. Since histograms of images are sums of their subparts, a rank minimization criteria allows for learning latent individual representations for all classes in the dataset. Thus, we can recover the localization information without the need for fully supervised training or MIL. We show empirical validation that our approach is able to associate the semantic concepts with regions in images;
3) We propose two new convex rank minimization algorithms, MC-Pos and MC-Simplex, motivated by the multi-label image classification problem and the additive histogram property. We prove that these enjoy the same convergence properties of Fixed Point Continuation methods for rank minimization without constraints.

A preliminary version of this work was published in [17].

II. Previous Work

This section reviews related image classification work and provides a brief survey on the use of the nuclear norm as a surrogate for rank minimization problems in computer vision.

A. Image Classification

Since the seminal work of Barnard and Forsyth [18], many researchers have addressed the problem of associating words to images. Image semantic understanding is now typically formulated as a multi-label problem. In this setting, each image may be simultaneously assigned to more than one class. This is an important difference from multi-class classification, where

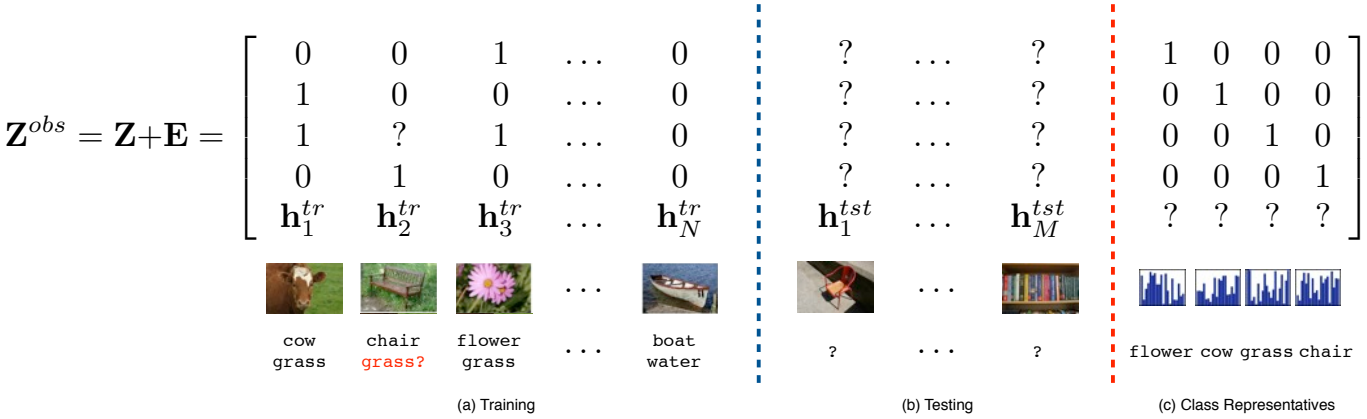


Fig. 3. Our weakly supervised classification algorithm works by completing a matrix $\mathbf{Z}^{obs}$ as shown above, where the question marks denote unknown entries. We complete this matrix such that it can be factorized into a low rank matrix $\mathbf{Z}$ and an error matrix $\mathbf{E}$. This ensures that background distributions and feature/label outliers are captured in $\mathbf{E}$, since they increase the rank of $\mathbf{Z}$. In the training submatrix (a), the i^{th} column concatenates the image histogram $\mathbf{h_i^{tr}}$ with its respective $\{0, 1\}$ label assignments. Note that a partially labeled example such as the second training image (a) is trivially handled by our framework. In the test submatrix (b), the j^{th} column is a concatenation of histogram $\mathbf{h_j^{tst}}$ with unknown assignments. In this transductive setting, the statistics of the test set are also used in the learning. By completing (c), we obtain a representative histogram for each class, in spite of their co-occurrence in the images.

classes are assumed to be independent and mutually exclusive. While multi-label can trivially be handled in multi-class approaches by dropping the mutual exclusivity constraint, Desai *et al*. [19] have shown the need to model object interactions. Therefore, many multi-class techniques such as SVM and LDA have been modified to make use of label correlations to improve multi-label classification performance [1]. In these approaches, localization is achieved by detection, using *e.g*., a sliding window. This is, however, at the expense of a fully supervised training set where localization is known a priori.

Several researchers have addressed the problem of classifying an image and providing precise class localization. Deselaers *et al*. [20] use a CRF to learn new class appearances from previously known ones obtained with supervised training. Blaschko *et al*. [21] learn a supervised structured output regression where the outputs are coordinates of a bounding box enclosing the object. Jamieson *et al*. [7] associate configurations of SIFT features to captions. Tighe and Lazebnik [22] propose lazy learning for large scale image parsing.

Alternatively to these approaches, Multiple Instance Learning (MIL) has surfaced as a reliable framework for performing learning in the presence of unknown latent factors. First proposed in [23], this class of learning problems extends the typical classification setting to the case where labels are no longer applied individually, but to multi-sets or "bags": a bag is labeled positive if at least one of its instances is positive and negative if none of its constituents are. In computer vision, this framework has been used for weakly supervised learning tasks such as learning deformable part models [12] and to explicitly model the relations between labels and specific regions of the image, as initially proposed by Maron and Lozano-Perez [24].

This method allows for the localization and classification tasks to benefit from each other, thus reducing noise in the corresponding feature space and making the learned semantic models more accurate [4], [8]–[11], [25], [26]. Although promising, the MIL framework is combinatorial, so several approaches have been proposed to avoid local minima and deal with the prohibitive number of possible subregions in an image. Zha *et al*. [10] make use of hidden CRFs while Vijayanarasimhan *et al*. [11] recur to multi-set kernels to emphasize instances differently. Yang *et al*. [8] exploit asymmetric loss functions to balance false positives and negatives. These methods, however, require an explicit enumeration of instances in the image. This is usually obtained by breaking images in a small fixed number of segments or applied in settings where detectors perform well, such as the problem of associating faces to captioned names [27]. On the other hand, to avoid explicitly enumerating the instances, Nguyen *et al*. [4] coupled constraint generation algorithms with a branch and bound method for fast localization. This is also seen in the negative data-mining process of [12]. Yakhnenko *et al*. [26] propose a MIL algorithm of linear complexity in the number of instances by using a non-convex Noisy-Or model. Multi-task learning has also been proposed as a way to regularize the MIL problem to avoid local minima due to many available degrees of freedom. In this setting, the MIL optimization is jointly learned with a fully supervised task [25].

To the best of the authors' knowledge, the only work modeling MIL as a convex problem is by Li and Sminchisescu [13]. They replace the classifier loss and the non-convex constraints on the positive bags by convex alternatives (f-divergence family loss and class likelihood ratios for each instance). They show promising results over standard MIL formulations as the ratio of positive instances in positive bags increase. Unfortunately, this is not the setting in image classification, as the percentage of possible negative bounding boxes in an image largely exceeds that of the positives.

B. Nuclear norm as a surrogate for rank minimization

Rank minimization has recently received much attention due to its success in collaborative filtering problems such as the Netflix challenge. Rank minimization techniques have also been applied to minimum order control [28], [29], to find least complex solutions achieving a performance measure.

A breakthrough by Candés and Recht [15] states the minimization of the rank function, under broad conditions, can

be achieved with the nuclear norm (sum of singular values). This result is a clear parallel with the results in [30] for the ℓ_1 and ℓ_0 norms. Since the natural reformulation of the nuclear norm gives rise to a Semidefinite Program, off-the-shelf optimizers can only handle problems of limited size. Thus, several methods have been devised for its efficient [15], [16], [31]–[34] and incremental [35], [36] optimization.

In the context of computer vision, the nuclear norm has been applied to several problems: Structure from motion [36], [37], robust PCA [38], subspace clustering [39], segmentation [40], tag refinement [41], camera calibration [42].

The nuclear norm regularizer has been applied to classification tasks in *e.g.*, [26], [41], [43]–[47]. Most of these approaches exploit the nuclear norm to enforce correlations between classifiers [47] or to allow for dimensionality reduction [43] in discriminative settings. Harchaoui *et al.* [47] decompose the nuclear norm into a surrogate infinite-dimensional optimization, allowing the feasibility of coordinate descent in large scale settings with smooth losses. Instead, we propose a generative approach based on [46] that is able to decouple appearance descriptions of co-occurring classes, allows for a recovery of segments and thus localization in the images.

This work can also relate to Latent Semantic Analysis, as the low rank justifications provided in Sec. III are similar in nature to the ones provided for subspaces obtained from document-term matrices. Bosch *et al.* [48] provide preliminary results that visual words associated with high probability to a given category can provide cues for localization. Our method, however, is not subject to local minima and estimates subspace ranks automatically.

III. Matrix Completion for Multi-Label Classification of Visual Data

This section describes the main contributions of this paper: We start by presenting the use of matrix completion for general classification tasks. Then, we describe its use for weakly supervised multi-label image classification and localization.

A. Classification as matrix completion

In a supervised setting, a classifier learns a mapping (see footnote[1] for notation) $\mathcal{W} : \mathcal{X} \rightarrow \mathcal{Y}$ between the space of features $\mathcal{X}$ and the space of labels $\mathcal{Y}$. This learning is done from a dataset of N training tuples $(\mathbf{x}_i^{tr}, \mathbf{y}_i^{tr}) \in \mathbb{R}^D \times \mathbb{R}^K$, where D is the feature dimension and K the number of classes. In particular, linear classifiers minimize a loss $l(\cdot)$ between the output space and the projection of the input space, as

$$\underset{\mathbf{W},\mathbf{b}}{\text{minimize}} \sum_{i=1}^{N} l\left(\mathbf{y}_i^{tr}, [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{x}_i^{tr} \\ 1 \end{bmatrix}\right), \tag{1}$$

where parameters $\mathbf{W} \in \mathbb{R}^{D\times K}$ and $\mathbf{b} \in \mathbb{R}^{K\times 1}$ describe the class decision boundaries. After the training stage, these parameters are used to estimate unknown labels for test samples $\mathbf{y}_j^{tst}$ from their feature descriptors $\mathbf{x}_j^{tst}$. This is typically done independently for each test entry, as

$$\mathbf{y}_j^{tst} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{x}_j^{tst} \\ 1 \end{bmatrix}. \tag{2}$$

In this paper, we formulate the problem of jointly classifying M test samples as one of matrix completion. For this purpose, let us define the feature matrices $\mathbf{X}^{tr} \in \mathbb{R}^{D\times N}$ and $\mathbf{X}^{tst} \in \mathbb{R}^{D\times M}$. These matrices respectively collect in each column feature vectors for N training and M test samples. Without loss of generality, the linear model assumed in (2) can be written in matrix form. Specifically, it states that $\mathbf{Y}^{tr} \in \mathbb{R}^{K\times N}$, the matrix concatenating the labels for all training images, can be obtained by the linear combination

$$\mathbf{Y}^{tr} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tr} - \mathbf{E}_X^{tr} \\ \mathbf{1}^\top \end{bmatrix} - \mathbf{E}_Y^{tr}, \tag{3}$$

where $\mathbf{E_Y}^{tr}$ and $\mathbf{E_X}^{tr}$ denote errors in the labels and features, respectively. The test labels $\mathbf{Y}^{tst} \in \mathbb{R}^{K\times M}$ are obtained as

$$\mathbf{Y}^{tst} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tst} - \mathbf{E}_X^{tst} \\ \mathbf{1}^\top \end{bmatrix}, \tag{4}$$

with no error in the labels since they are unknown. Concatenating labels and features in (3) and (4) in one matrix yields

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X^{tr} & \mathbf{E}_X^{tst} \\ \mathbf{0}^\top \end{bmatrix} = \mathbf{Z}^{obs} - \mathbf{E}, \tag{5}$$

where $\mathbf{Z}^{obs} \in \mathbb{R}^{(K+D+1)\times(M+N)}$ holds all observed entries (with $\mathbf{Y}^{tst}$ unknown) and $\mathbf{E}$ is a matrix of errors.

Note that according to (3) and (4), the matrix $\mathbf{Z}$ defined in (5) is rank deficient. That is, the rows comprising the labels are linearly dependent on the feature rows. In the absence of error ($\mathbf{E} = \mathbf{0}$), the input matrix $\mathbf{Z}^{obs}$ is also low rank, as

$$\text{rank}(\mathbf{Z}) = \text{rank}(\mathbf{Z}^{obs}) = \text{rank}\left(\begin{bmatrix} \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix}\right). \tag{6}$$

In this case, we observe that (3) becomes

$$\mathbf{Y}^{tr} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tr} \\ \mathbf{1}^\top \end{bmatrix}, \tag{7}$$

and thus the $\mathbf{Y}^{tst}$ in (4) does not increase the rank of $\mathbf{Z}$, since

$$\mathbf{Y}^{tst} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix}. \tag{8}$$

Using this result, Goldberg *et al.* [46] suggest unknown test labels in $\mathbf{Y}^{tst}$ can be recovered by completing these entries such that the rank of $\mathbf{Z}$ is minimized. This can be written as

$$\begin{aligned} \underset{\mathbf{Y}^{tst}}{\text{minimize}} \quad & \text{rank}(\mathbf{Z}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix}. \end{aligned} \tag{9}$$

[1] Bold capital letters denote matrices (*e.g.*, $\mathbf{D}$), bold lower-case letters represent column vectors (*e.g.*, $\mathbf{d}$). All non-bold letters denote scalar variables. $\mathbf{d}^j$ is the j^{th} column of the matrix $\mathbf{D}$. d_{ij} denotes the scalar in the row i and column j of $\mathbf{D}$. $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ denotes the inner product between two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$. $||\mathbf{d}||_2^2 = \langle \mathbf{d}, \mathbf{d} \rangle = \sum^i d_i^2$ denotes the squared Euclidean Norm of the vector $\mathbf{d}$. $\text{tr}(\mathbf{A}) = \sum^i a_{ii}$ is the trace of the matrix $\mathbf{A}$. $||\mathbf{A}||_*$ designates the nuclear norm (sum of singular values) of $\mathbf{A}$.$\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^\top\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}^\top)$ designates the squared Frobenius Norm of $\mathbf{A}$. $\mathbf{1}_k \in \mathbb{R}^{K\times 1}$ is a vector of ones, and $\mathbf{e}_j \in \mathbb{R}^{K\times 1}$ denotes the j^{th} canonical vector, with 1 at the j^{th} position and zero otherwise. $\mathbf{0}_{K\times N} \in \mathbb{R}^{K\times N}$ is a matrix of zeros and $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ denotes the identity matrix.

In practice $\mathbf{E} \neq \mathbf{0}$, so we modify (9) to include (5). To avoid trivial solutions, we penalize errors with a loss $l(\cdot)$, as

$$\begin{aligned} \underset{\mathbf{Y}^{tst},\mathbf{E}_Y^{tr},\mathbf{E}_X}{\text{minimize}} \quad & \text{rank}(\mathbf{Z}) + \lambda l(\mathbf{E}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X \\ \mathbf{0}^\top \end{bmatrix}, \end{aligned} \tag{10}$$

where λ is a tradeoff parameter and $\mathbf{E}_X = [\mathbf{E}_X^{tr}\ \mathbf{E}_X^{tst}]$. We discuss the choices of loss functions $l(\cdot)$ in detail in Sec. III-D.

There are three fundamental advantages in casting a general classification problem as the matrix completion in (10).

First, it bypasses the estimation of the model parameters $\mathbf{W}$ and $\mathbf{b}$. This allows our formulation to estimate errors in the features $\mathbf{E}_X$. Parametric models that estimate $\mathbf{W}$ and $\mathbf{b}$ (such as linear regression or SVMs) do not model this error, and thus implicitly assume $\mathbf{E}_X = \mathbf{0}$. Note that the product $\mathbf{W}^\top \mathbf{E}_X$ in (3) will result in a non-convex problem when both $\mathbf{W}$ and $\mathbf{E}_X$ are considered as optimization variables. While (10) is also non-convex, we show in Sec. III-C that a convex relaxation exists, backed by the recent advances in rank minimization.

Second, errors in features and labels are estimated jointly, so missing data in labels and features can easily be estimated.

Third, we minimize the rank of $\mathbf{Z}$, containing training and test samples. This transductive setting allows the model to leverage the statistics of the test set.

B. Image classification as matrix completion

In spite of justifying the applicability of matrix completion as a generic classification framework, the explanation provided by Goldberg *et al.* [46] described in Sec. III-A only spans the row space of $\mathbf{Z}$. In this section, we provide an alternative explanation for the low rank of $\mathbf{Z}$ in (5), based instead on its column space. Let us assume the case when histograms are used as feature vectors. Note that several popular techniques for obtaining global representations of images in computer vision, such as Bag of Words or HOG, fall under this assumption. Let $\mathbf{h}^i$ denote such a histogram representation for image i. In this case, the feature submatrix $\mathbf{X} = [\mathbf{X}^{tr}\ \mathbf{X}^{tst}]$ in (5) contains one histogram per column, as

$$\mathbf{X} = \left[\begin{array}{ccc|ccc} \mathbf{h}_1^{tr} & \cdots & \mathbf{h}_N^{tr} & \mathbf{h}_1^{tst} & \cdots & \mathbf{h}_M^{tst} \end{array} \right]. \tag{11}$$

One property of image histograms is that they can be represented by a sum of the histograms of its segments (see Figure 2). Without loss of generality, we consider these latent histograms as $\mathbf{C}_k \in \mathbb{R}^{D \times N_k}$, the N_k canonical histogram representations for class k. Therefore, we have that the histogram of image i can be written as a sum of class representatives $\mathbf{C}_k$ weighted by coefficients $\mathbf{a}_{k,i} \in \mathbb{R}^{N_k \times 1}$, as

$$\mathbf{h}_i = \sum_k \mathbf{C}_k \mathbf{a}_{k,i} + \mathbf{E}_{X\,i}, \tag{12}$$

where $\mathbf{E}_{X\,i}$ collects errors (*e.g.*, words in the background that do not pertain to any class). If we concatenate the representatives $\mathbf{C}_k$ in the matrix

$$\mathbf{C} = \left[\begin{array}{cccc} \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_K \end{array} \right], \tag{13}$$

and collect weights $\mathbf{a}_{k,i}$ in a matrix $\mathbf{A}$ we can write (11) as

$$\mathbf{X} = \mathbf{C}\mathbf{A} + \mathbf{E}_X. \tag{14}$$

Additionally, since we postulated each $\mathbf{c}_{kj}$ as belonging to only class k, the correspondent label matrix for $\mathbf{C}$ is given by

$$\mathbf{Y}_C = \left[\begin{array}{ccc} \mathbf{e}_1 \mathbf{1}_{N_1}^\top & \cdots & \mathbf{e}_K \mathbf{1}_{N_K}^\top \end{array} \right], \tag{15}$$

where $\mathbf{e}_i$ denotes the i^{th} canonical vector. Merging (11) and (15), we obtain the data matrix $\mathbf{Z}^{obs}$ in (5) as

$$\mathbf{Z}^{obs} = \begin{bmatrix} \mathbf{Y}_C \\ \mathbf{C} \end{bmatrix} \mathbf{A} + \begin{bmatrix} \mathbf{E}_Y \\ \mathbf{E}_X \end{bmatrix} = \mathbf{Z} + \mathbf{E}, \tag{16}$$

the sum of a low rank component matrix $\mathbf{Z}$ with an error matrix $\mathbf{E}$. A close inspection of (16) allows us to state that $\mathbf{Z}^{obs}$ is low rank also due to its column space, in absence of background noise, since class histograms are shared across images and therefore $\sum_k N_k < N+M$. Additionally, it allows for the observation that the appearance of individual classes can be recovered from a multi-label dataset by estimating $\mathbf{C}$. In this paper, we assume that for localization purposes, each class can be well represented by a single histogram. In this case, (15) becomes $\mathbf{Y}_C = \mathbf{I}_K$, and therefore our approach can obtain an estimate of $\mathbf{C}$ by completing in $\mathbf{Z}^{obs}$ the features correspondent to the canonical labels (see Figure 3 (c)). By directly estimating $\mathbf{C}$, we are able to recover the appearance of each class and thus provide the localization for each concept in the images. This is done despite the weakly supervised setting and bypassing the combinatorial nature of searching for bounding boxes such as in MIL problems. Also, note that this assumption is not used in the classification, where our algorithm estimates class subspace dimensions automatically.

C. Nuclear norm as a convex surrogate of the rank function

Since the rank is a highly non-convex and non-differentiable function, it is nontrivial to minimize. Therefore, we relax (10) by using the convex envelope of the rank function, the nuclear norm. Let $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the SVD of $\mathbf{Z}$. The nuclear norm is defined as $\text{tr}(\mathbf{\Sigma})$, the sum of singular values of $\mathbf{Z}$. It has been shown that under general assumptions of low coherence of the singular vectors of $\mathbf{Z}$, minimizers obtained using the nuclear norm are equal to minimizers of rank with high probability [15]. Therefore, we rewrite (10) as

$$\begin{aligned} \underset{\mathbf{Y}^{tst},\mathbf{E}_Y{}^{tr},\mathbf{E}_X}{\text{minimize}} \quad & \|\mathbf{Z}\|_* + \lambda l(\mathbf{E}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X \\ \mathbf{0}^\top \end{bmatrix}. \end{aligned} \tag{17}$$

We provide a simple intuition as to why this norm is in fact the largest possible convex underestimator of the rank function: Since the singular values of matrices are always positive, the nuclear norm can be interpreted as an ℓ_1-norm of the singular values. Under this interpretation, one can easily identify it as the convex envelope of the rank function, since the latter is the cardinality (or ℓ_0-norm) of the singular values.

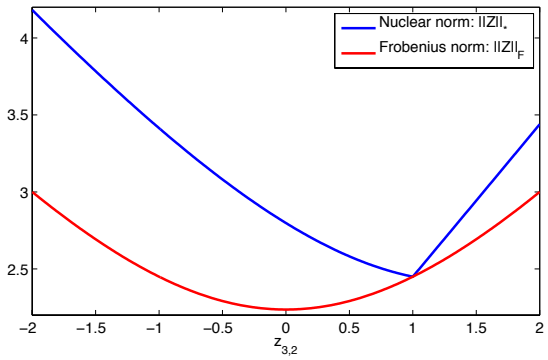


Fig. 4. Comparison of Nuclear and Frobenius norms as function of one single unknown entry $z_{2,3}$ for the matrix in (18).

1) Toy example: To understand why the singular value sparsity induced by the nuclear norm is important for the matrix completion in (17), consider completing a rank-1 matrix

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & ? \end{bmatrix}, \tag{18}$$

where only one entry $z_{2,3}$ is unknown. The results shown in Figure 4 plot the nuclear norm and Frobenius norm of $\mathbf{Z}$ for all possible completions in a range around the value that minimizes its rank $z_{2,3} = 1$. In this case, the sparsity induced by the nuclear norm (ℓ_1-norm on the singular values) yields the optimal solution for $\mathbf{Z}$ with singular values $\sigma = [2.4495\ 0]$, a rank-1 matrix. In opposition, the Frobenius Norm (ℓ_2-norm of singular values) will set the entries to zero, thus leading to a solution with singular values $\sigma = [2.1358\ 0.6622]$, a rank-2 matrix. This key difference can be attributed to the fact that completing a matrix under the rank or nuclear norm favors the interaction between rows and columns to find a global solution, while the Frobenius norm treats each entry in the matrix independently (recall that $\|\mathbf{Z}\|_F^2 = \sum_{ij} z_{ij}^2$).

D. Adding robustness into matrix completion

In practical applications, we have several sources of errors in the features (*e.g.*, changes in pose, illumination, background noise) and missing data in the training samples (*e.g.*, missing labels), which will translate into nonzero error matrices in the models of (5) and (16). We account for these possible violations by allowing the matrix $\mathbf{Z}$ in (17) to deviate from the original data matrix. The resulting optimization problem finds the best label assignment $\mathbf{Y}^{tst}$ and error matrices $\mathbf{E}_X = [\mathbf{E}_{X^{tr}}\ \mathbf{E}_{X^{tst}}], \mathbf{E}_{Y^{tr}}$ such that the rank of $\mathbf{Z}$ is minimized, as

$$\begin{aligned} \underset{\mathbf{Y}^{tst},\mathbf{E}_{Y^{tr}},\mathbf{E}_X}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + l_x(\mathbf{E}_X) + \lambda l_y(\mathbf{E}_Y^{tr}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top & \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X & \\ \mathbf{0}^\top & \end{bmatrix}. \end{aligned} \tag{19}$$

Here, distortions of $\mathbf{Z}$ from known labels and features are penalized according to $l_y(\cdot)$ and $l_x(\cdot)$, respectively. The parameters λ, μ are positive trade-off weights between better feature adaptation and label error correction. We rewrite (19) by defining sets Ω_X and Ω_Y of known feature and label entries and $\mathbf{Z}_Y, \mathbf{Z}_X, \mathbf{Z}_1$ as the label, feature and last rows of $\mathbf{Z}$, as

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\sum_{ij\in\Omega_X} l_x(z_{ij}, z_{ij}^{obs}) \\ & + \frac{\lambda}{|\Omega_Y|}\sum_{ij\in\Omega_Y} l_y(z_{ij}, z_{ij}^{obs}) \\ \text{subject to} \quad & \mathbf{Z}_1 = \mathbf{1}^\top \end{aligned}, \tag{20}$$

where the constraint that $\mathbf{Z}_1$ be equal to one is necessary for dealing with the bias $\mathbf{b}$ in (3). The model in (20) can be solved using Fixed Point Continuation [16], described in Sec. III-F.

In [46], $l_x(\cdot)$ was defined as the least squares error and $l_y(\cdot)$ a log loss to emphasize the error on entries switching classes as opposed to their absolute numerical difference. We note that in this model (MC-1), the log loss in $l_y(\cdot)$, albeit asymmetric, incurs in unnecessary penalization of entries belonging to the same class as the original entry.Therefore, we generalize this loss to a smooth approximation of the Hinge loss, controlled by a parameter γ. For labels $\{-1, 1\}$, we have

$$l_y(z_{ij}, z_{ij}^{obs}) = \frac{1}{\gamma}\log(1 + \exp(-\gamma z_{ij}^{obs} z_{ij})), \tag{21}$$

and for the case of labels $\{0, 1\}$, we have

$$l_y(z_{ij}, z_{ij}^{obs}) = \frac{1}{\gamma}\log\left(1 + \exp\left(-\gamma(2z_{ij}^{obs} - 1)(z_{ij} - z_{ij}^{obs})\right)\right). \tag{22}$$

Also, in the bag of words model, visual data are encoded as histograms. In this setting, (20) is inadequate as it introduces negative values to the histograms in $\mathbf{Z}_X$. Thus, we replace the least-squares penalty in $l_x(\cdot)$ by a χ^2 distance,

$$\chi^2(\mathbf{z}^j, \mathbf{z}_0^j) = \sum_{i=1}^{F} \chi_i^2(z_{ij}, z_{ij}^{obs}) = \sum_{i=1}^{F} \frac{\left(z_{ij} - z_{ij}^{obs}\right)^2}{z_{ij} + z_{ij}^{obs}}. \tag{23}$$

and constrain all feature vectors to be positive (MC-Pos model)

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\sum_{ij\in\Omega_X} \chi_i^2(z_{ij}, z_{ij}^{obs}) \\ & + \frac{\lambda}{|\Omega_Y|}\sum_{ij\in\Omega_Y} l_y(z_{ij}, z_{ij}^{obs}) \\ \text{subject to} \quad & \mathbf{Z}_X \geq \mathbf{0} \\ & \mathbf{Z}_1 = \mathbf{1}^\top, \end{aligned} \tag{24}$$

or in the Probability Simplex $\mathcal{P}$ (MC-Simplex model)

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\sum_{ij\in\Omega_X} \chi_i^2(z_{ij}, z_{ij}^{obs}) \\ & + \frac{\lambda}{|\Omega_Y|}\sum_{ij\in\Omega_Y} l_y(z_{ij}, z_{ij}^{obs}) \\ \text{subject to} \quad & \mathbf{Z}_X \in \mathcal{P} \\ & \mathbf{Z}_1 = \mathbf{1}^\top, \end{aligned} \tag{25}$$

depending on whether we wish to perform normalization on the data or not. Observe that (20) and (25) are both convex.

E. Comparison to other subspace techniques

It is important to note that many standard dimensionality reduction techniques such as PCA and LDA have been robustified by using a nuclear norm penalization typically coupled with an ℓ_1 error function [38], [49]. The differences and similarities between the method presented in Section III-D and these techniques can be analyzed if one interprets (24),(25) as forms of PCA with missing data. Our method can be seen as an extension of Robust PCA in two ways: 1) it includes labels as additional "features" in the data samples 2) it penalizes label and features errors with different losses l_x and l_y.

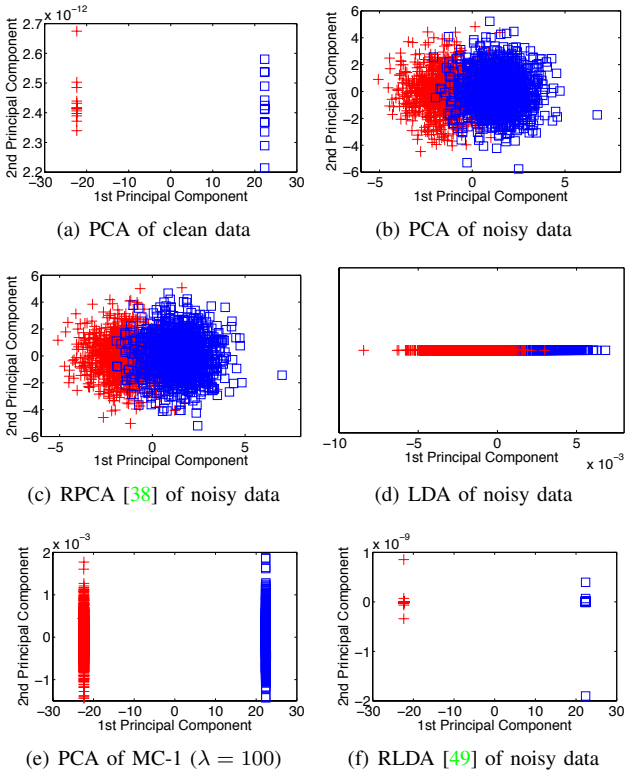
A comparison between the behavior of PCA, LDA, RPCA [38], RLDA [49] and our method in the presence of noise can be seen in Figure 5. We generated a two-class dataset of $2,000$ 500-dimensional vectors. The positive and negative classes (resp.) have $1,000$ samples of the form $-\mathbf{1}_{500}$ and $\mathbf{1}_{500}$ (resp.). We refer to this as clean data. The first two principal components of this clean data are in Figure 5(a). Then, we added to the clean data noise sampled from a Normal distribution with zero mean and standard deviation $20\mathbf{I}_{2\times 2}$. We plot the two principal components data in Figure 5(b). Note that PCA does not recover the underlying structure of the clean data due to the significant amount of noise.

In this example, because the data does not have outliers and the noise does not follow a Laplacian distribution, the ℓ_1 error function assumed by RPCA [38] is not able to clean the noisy data (Figure 5(c)). Similarly, augmenting the space by adding the labels as an additional dimension does not help since for RPCA the errors in features and labels are weighted equally. In both these cases, the output of RPCA (Figure 5(c)) is similar to the one obtained by regular PCA (Figure 5(b)). LDA (Figure 5(d)) is able to find a projection which classifies most of the points correctly. However, observe that it fails to clean the data, which results in several misclassified points on the class boundary. Our matrix completion approach, in turn, balances a trade-off between correcting the data points, correcting the labels and minimizing the rank. Therefore, it is able to correct the feature data (Figure 5(e)) by giving more weight to the information on the labels. This capability of correcting the errors in features is only matched by our work in Robust LDA [49], which achieved the result in Figure 5(f). While this method has the advantage of obtaining an explicit transformation from the feature to the label space, the matrix completion has the ability to clean the test data during training.

F. Fixed Point Continuation (FPC) for MC-Pos/MC-Simplex

Albeit convex, the nuclear norm makes (24) and (25) not smooth. Since nuclear norm problems are naturally cast as Semidefinite Programs, existing interior point methods are inapplicable due to the large dimension of $\mathbf{Z}$. Thus, several methods have been devised to efficiently optimize this problem class [15], [16], [31]–[35]. The FPC method [16], in particular, consists of a series of gradient descent updates $h(\cdot) = I(\cdot) - \tau g(\cdot)$ with step size τ and gradient $g(\cdot)$ as

$$g(z_{ij}) = \begin{cases} \frac{\lambda}{|\Omega_Y|} \frac{-z_{ij}^{obs}}{1+\exp\left(\gamma z_{ij}^{obs} z_{ij}\right)} & \text{if } z_{ij} \in \Omega_Y, \\ \frac{1}{|\Omega_X|} \frac{z_{ij}^2 + 2 z_{ij} z_{ij}^{obs} - 3 z_{ij}^{obs\,2}}{(z_{ij}+z_{ij}^{obs})^2} & \text{if } z_{ij} \in \Omega_X, \end{cases} \tag{26}$$



(a) PCA of clean data (b) PCA of noisy data

(c) RPCA [38] of noisy data (d) LDA of noisy data

(e) PCA of MC-1 ($\lambda = 100$) (f) RLDA [49] of noisy data

Fig. 5. Comparison of results obtained for two-class classification of the random dataset in III-E. The data error correction in Robust LDA (f) and Matrix Completion (e) allow for recovery of the original data, when other component analysis methods don't.

and 0 otherwise. These steps are alternated with a shrinkage operator $S_\nu(\cdot) = \max(0, \cdot - \nu)$ on the singular values of the resulting matrix, to minimize its rank. Provided $h(\cdot)$ is non-expansive, FPC converges to the optimal solution for the unconstrained problem. FPC was originally devised in [16] for unconstrained problems and extended in [46] to solve the formulation MC-1 (20) by adding a projection step. However, its convergence was only empirically verified. In Appendix V, we prove the convergence of FPC for (20), (24), (25) using the fact that projections onto convex sets are non-expansive.

Key to the feasibility of FPC is an efficient way to project $\mathbf{Z}$ onto the constraint sets in (24) and (25). While for MC-Pos (24) the non-negative orthant projection is done in closed form by setting negative components to zero, efficiently projecting onto the probability simplex in MC-Simplex (25) is not straightforward. We note, however, this is a projection onto a convex subset of an ℓ_1 ball. Therefore, we can explore the dual of the projection problem and use a sorting procedure to implement this projection in closed form, as described in [17], [50]. The algorithms are summarized in Alg. 1. We note that the computational bottleneck is the computation of the SVD of $\mathbf{Z}$. State-of-the-art methods for SVD (*e.g.*, Lanczos bidiagonalization algorithm with partial reorthogonalization) take a flop count of $O((K+D+1)(M+N)^2 + (M+N)^3)$.

Algorithm 1 FPC for MC-Pos (24) and MC-Simplex (25)

Input: Initial Matrix $\mathbf{Z}^{obs}$, known entries sets Ω_X, Ω_Y
  Initialize $\mathbf{Z}$ as the rank-1 approximation of $\mathbf{Z}^{obs}$
  for $\mu = \mu_1 > \mu_2 > \cdots > \mu_k$ **do**
    while Rel. Error $> \epsilon$ **do**
      Gradient Descent: $\mathbf{A} = \mathbf{Z} - \tau g(\mathbf{Z})$
      Shrink 1: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$
      Shrink 2: $\mathbf{Z} = \mathbf{U}S_{\tau\mu}(\mathbf{\Sigma})\mathbf{V}^\top$
      Project $\mathbf{Z}_X$: $\mathbf{Z}_X = \max(\mathbf{Z}_X, \mathbf{0})$ for (24)
      Project $\mathbf{Z}_X$ onto the probability simplex $\mathcal{P}$ for (25)
      Project $\mathbf{Z}_1$: $\mathbf{Z}_1 = \mathbf{1}^\top$
    end while
  end for
Output: Complete Matrix $\mathbf{Z}$

IV. Experiments

This section presents the performance evaluation of the proposed algorithms MC-Pos (24) and MC-Simplex (25) in several tasks. In the first experiment (Sec. IV-B), we validated the low rank assumption of Sec. III-B using two multi-label datasets, MSRC[2] and SIFTFlow [51]. In the second experiment (Sec. IV-C), we evaluated the classification performance and the localization abilities of our method on the CMU-Face dataset [4] (a single-class problem). In the third experiment, we evaluated the performance of our method for multi-label classification (Sec. IV-D) in the MSRC and PASCAL VOC2007 datasets. We also perform an experiment for localization (Sec. IV-E) in MSRC. We compared our methods with MC-1 (20), an SVM baseline, and several state-of-the-art MIL approaches [4], [9]–[11], [25], [26], [52].

A. Parameters

For MC-Pos, MC-Simplex and MC-1, the values considered for parameter tuning were $\gamma \in [1, 30], \lambda \in [10^{-4}, 10^2]$. The continuation steps require a decreasing sequence of μ, which we chose as $\mu_k = 0.25\mu_{k-1}$, stopping when $\mu = 10^{-9}$. We used $\mu_0 = 0.25\sigma_1$, where σ_1 is the largest singular value of $\mathbf{Z}^{obs}$, with unknown entries set to zero. Convergence was defined as a relative change in the objective function smaller than 10^{-2}. In a transduction setting, since the task is to classify an already known test set, one could choose the parameters which perform best on the final test set. However, to be fair to other baselines, we tuned the parameters in a cross validation setting. As such, the results reported are for the choice of parameters which, from the aforementioned ranges, yielded the best average result on all the validation sets provided by cross-validation. The results reported for the SVM baselines were obtained using libSVM, with parameter $C \in [10^{-6}, 10^6]$.

B. Low rank assumption validation

In this experiment, we empirically validated the assumption in (16) that histograms of objects of the same class share a low-dimensional subspace. We constructed a bag of words representation for the MSRC dataset, which consists of 591 real world images distributed among 21 classes, with an average of 3 classes present per image. We mimicked the setup of [10], [11] and dismissed the classes void, mountain and horse. To obtain a bag of words (BOW) descriptor, we clustered texton filter responses [53] obtained from all three CIELab color channels into a codebook by applying k-means to a random subset of $40,000$ descriptors. In this model [14], images are encoded as histograms representing the distribution of the 400 words from the codebook. Then, using the ground truth segmentation labeling, we collected feature matrices $\mathbf{X}_1$ by concatenating all the histograms of the same class. We compared these with feature matrices $\mathbf{X}_2$ of the same dimension with an equal amount of elements from all classes (including elements from the class of $\mathbf{X}_1$). In order to compare the singular value distribution of these matrices, we normalized them so columns have unit ℓ_2 norm. Then, we measured their nuclear norm ratio (NNR), defined as

$$NNR(\mathbf{X}_1, \mathbf{X}_2) = \frac{\|\mathbf{X}_1\|_*}{\|\mathbf{X}_2\|_*}. \quad (27)$$

This measure provides an empirical validation of our assumption and is linked to what our model is optimizing and is an indirect measure of the rank of a matrix, as explained in Sec. III-C. Results on Table I(a) show that for all classes in the MSRC dataset, we obtained a NNR lower than 1. An assignment of test entries to incorrect class labels yields a higher nuclear norm of $\mathbf{Z}$, thus validating our model. For visualization, we plot the singular value distribution of $\mathbf{X}_1$ and corresponding $\mathbf{X}_2$ for some classes in the dataset (Figure 6).

It might be argued that explanation of (16) only holds when the columns dominate the estimate of the rank, *i.e.*, $\text{rank}(\begin{bmatrix} \mathbf{X}^{tr} & \mathbf{X}^{tst} \end{bmatrix}) \leq N + M \leq F$. However, we also validated this hypothesis in the case when the feature dimension F is smaller than the number of images $N+M$ in the dataset. Since there are only 591 images in the MSRC dataset and some classes exhibit a small number of exemplars, we validated this assumption in the larger scale SIFTFlow dataset [51]. This dataset is a collection of $2,688$ images distributed among 33 classes. Following [51], we extracted a dense HoG feature map [54] from every image in the dataset and built a BoW codebook of 200 words. We collected the histograms for all the $25,758$ ground truth segments in the dataset according to their class label. Then, we calculated the distribution of singular

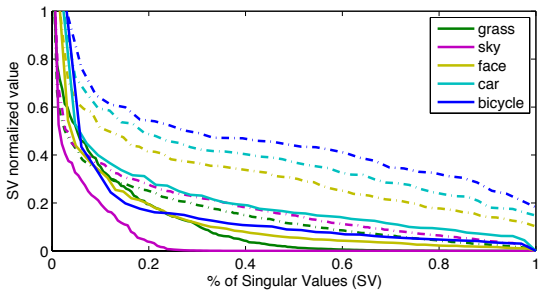


Fig. 6. Comparison of singular value distribution of matrices $\mathbf{X}_1$ with histograms of the same class (solid) versus corresponding matrices $\mathbf{X}_2$ of the same dimension with an equal amount of histograms from all classes (dashed) for different classes on the MSRC dataset.

[2] http://research.microsoft.com/en-us/projects/objectclassrecognition/

TABLE I
NUCLEAR NORM RATIOS (NNR) FOR ALL CLASSES IN THE MSRC DATASET (A) AND FOR ALL CLASSES WHICH HAVE MORE THAN 200 SEGMENTS IN THE SIFTFLOW DATASET (B).

(a) MSRC dataset.

Class	NNR
building	0.8595
grass	0.6987
tree	0.8325
cow	0.9092
sheep	0.7653
sky	0.4530
aeroplane	0.7831
water	0.8224
face	0.6622
car	0.8392
bicycle	0.6525
flower	0.8741
sign	0.9491
bird	0.8793
book	0.9217
chair	0.9397
road	0.7070
cat	0.8402
dog	0.8420
body	0.9465
boat	0.9123

(b) SIFTflow dataset.

Class	NNR
building	0.9074
tree	0.8620
car	0.8989
sky	0.8455
window	0.7513
mountain	0.8657
road	0.8568
person	0.8673
plant	0.8655
sidewalk	0.9038
rock	0.8728
door	0.6554
sea	0.6073
field	0.7719
sign	0.9098
grass	0.8181
streetlight	0.9439
river	0.8719
balcony	0.7458

values for matrices $\mathbf{X}_1$ as aforementioned, for all classes with more than 200 samples in the dataset. We compared the NNR of the matrices $\mathbf{X}_1$ with matrices $\mathbf{X}_2$ of the same dimension comprised by an equal amount of elements from all classes. Results in Table I(b) corroborate the MSRC dataset results, showing our assumption is also valid when the feature dimensions are smaller than the number of images.

C. Classification and localization on a two-class problem

In this experiment, we tested the classification performance of our method in a two-class classification problem. We used the CMU Face dataset [55], which consists of 624 images of 20 subjects. All subjects are captured with varying expression and poses, with and without sunglasses. Figure 7 shows examples of our positive (wearing sunglasses) and negative class (not wearing sunglasses). We have two goals: First, we want to build a classifier that, given a new face image, determines whether the subject is wearing sunglasses or not. Second, Nguyen *et al*. [4] argue that better results are obtained when the classifier training is restricted to the region that has the discriminative information (*e.g*., the glasses region in this case). They propose using a Multiple Instance Learning framework (MIL-SegSVM) that localizes the most discriminative region in each image while learning a classifier to discriminate between classes. We show how our method is also able to estimate the histogram of the discriminative region (*i.e*., sunglasses) and localize it in the training and test set.

To allow for direct comparison, we used the setup and features of [4]: Our training set is built using images of the first 8 subjects (126 images with sunglasses and 128 without), leaving the remainder for testing (370, equally split among the positive and negative classes). We represented each image with the BoW model by extracting $10,000$ SIFT features [56] at random scales and positions and quantizing them onto a

(a) (b)

Fig. 7. Example images of the CMU-Face dataset. (a) shows the positive class (wearing sunglasses) and (b) shows the negative class (no sunglasses).

$1,000$ visual codebook, obtained by performing hierarchical k-means clustering on $100,000$ features randomly selected from the training set. For the first part of the experiment, we compared the results of our classifier to what is obtained using several methods: (1) SVM-Img: a Support Vector Machine (SVM) trained using the entire image, (2) SVM-Region: an SVM trained using a manually labeled discriminative region (in this case, the region of the glasses), (3) MIL-SegSVM: a MIL SVM method proposed by [4]. For MC-1, MC-Pos and MC-Simplex, we proceeded as follows: we built $\mathbf{Z}$ with the label vector and the BoW histograms of each entire image and left the test set labels $\mathbf{Y}^{tst}$ as unknown entries. For the MC-Simplex case, we further preprocessed $\mathbf{Z}$ by dividing each histogram in $\mathbf{Z}_X$ by its sum. This was done to avoid the Simplex projection step in Alg. 1 picking a single bin and zeroing out the others, due to scale disparities in the bin counts.

The performance, measured using the area under ROC curve (AUROC), is shown in Table II. These results indicate both the fully supervised (SVM-FS) and the MIL approach (MIL-SegSVM) are more robust to the noise introduced by non-discriminative parts of the images, when compared to training without localization (SVM-Img). However, this is done at either the cost of cumbersome labeling efforts or by iteratively approximating the solution of the MIL problem, an integer quadratic problem. The matrix completion approaches (MC-1, MC-Pos, MC-Simplex), in turn, are able to surpass these classification scores by solving a convex minimization.

Beyond improving the classification performance, our algorithm is able to localize the discriminative region of interest (the sunglasses region, in this dataset). Recall that the error $\mathbf{E}_\mathbf{X}$ removes the portion of the histogram introduced by the non-discriminative regions of the image. To illustrate this property, after we run the matrix completion classification, we obtain the most discriminative bounding box for all images in the dataset. For each image i in the dataset, we searched for the bounding box that best matches the features of the i-th column of the completed matrix $\mathbf{z}_\mathbf{X}{}^i = \mathbf{h}^i - \mathbf{e}_\mathbf{X}{}^i$ (recall Figure 3). We use a sliding window detector varying scale and position using the size criteria in [4] and measure similarity using the χ^2 distance. The results are shown in Figure 8 for MC-Pos (similar results were obtained with MC-Simplex). Similarly to MIL-SegSVM, which used a linear SVM score for the subwindow search, our methods correctly localized the eyes region, that discriminates between the classes. Note that MC-1 does not allow to pursue localization of the class representative since it may introduce negative numbers in the histograms.

Fig. 8. A sliding window search shows that histograms corrected by MC-Pos (24) are most similar to the discriminative region of the eyes in the images.

TABLE II
AUROC RESULT COMPARISON FOR THE CMU FACE DATASET.

Method	AUROC
SVM-Img [4]	0.90
SVM-FS [4]	0.94
MIL-SegSVM [4]	0.96
MC-1 [46]	0.96
MC-Pos	**0.97**
MC-Simplex	0.96

D. Classification in multi-label datasets

In this experiment, we ran our method on two multi-label datasets: MSRC and PASCAL VOC 2007. The MSRC dataset consists of 591 photos distributed among 21 classes, with an average of 3 classes present per image. We mimicked the setup of [10], [11] and used as features histograms of textons [53]. Then, we obtained a 400 word codebook by applying k-means clustering to a random subset of $40,000$ descriptors.

In this task, all training images are labeled with one or several classes, and the goal is to label the test images. Observe that the test image can have several labels (*i.e.*, multi-label classification). We proceeded as in the experiment described in Sec. IV-C. We compared MC-Pos and MC-Simplex with MC-1 and several state-of-the-art multi-label MIL approaches: Multiple Set Kernel MIL (MSK-MIL) by Vijayanarasimhan and Grauman [11], Multi-label Multiple Instance Learning (ML-MIL) by Zha *et al*. [10], Discriminative Multiple Instance Multiple Label model by Yakhnenko and Honavar [26]. We also compared to a one-vs-all linear SVM.

The obtained average AUROC classification scores on the test set using 5-fold cross validation are shown in Table III(a). Results show that our methods outperformed MC-1, thus showing the improvement introduced by the additional constraints and improved loss functions. Moreover, they outperformed results given by state-of-the-art MIL techniques, including the non-linear classifier MSK-MIL. This can be explained by the fact that MIL methods work by selecting regions from images to be the positive examples for a class while learning that class boundary. Since possible regions are enumerated by a segmentation algorithm, it is not guaranteed that they match exactly the ground truth segmentation. The feature error correction in MC-Pos and MC-Simplex does not require this segmentation step and thus allows for superior results in this weakly supervised multi-label scenario.

We also tested our method in the PASCAL VOC 2007 dataset. This dataset consists of 9963 images labeled with at least one of 20 classes, split into `trainval` and `test` sets. We used the same features as the winning approach (INRIA_Genetic) [2]. This method achieved a mean average precision (mAP) of 0.594. Given that it is a fusion method, we followed its simplest feature setting (reported to yield mAP of 0.477) and represented each image by extracting dense SIFT features [56] and quantizing them onto a $4,000$ dimension codebook, built by k-means clustering on $100,000$ features randomly selected from the training set. We used this codebook to code images as histograms, which we ℓ_1 normalized. Results in Table III(b) show increased performance for the same features, demonstrating the power of transduction having access to the test set statistics. To confirm this, we further explored a transduction setting, by tuning parameters on the test set. Here, whilst SVM gets an mAP of 0.4966, MC-1/Pos/Simplex yield 0.952, 1.000 and 0.992 respectively. This is only achievable by an SVM if training is done with the test set. Our approach, however, only requires its features.

TABLE III
5-FOLD CROSS VALIDATION AVERAGE AUROC COMPARISON FOR THE IMAGE AND LABELING TASKS ON MSRC (A) AND CLASSIFICATION TASK RESULTS (B) IN THE VOC 2007 DATASET.

(a) MSRC Image labeling

Method	Image
MSK-MIL [11]	0.90
ML-MIL [10]	0.90
DMIML-ℓ_2 [26]	0.91
MC-1 [46]	0.91
MC-Pos	**0.95**
MC-Simplex	0.92
Linear SVM	0.89

(b) PASCAL VOC 2007

Method	mAP
INRIA_Genetic	0.594
MC-1 [46]	0.426
MC-Pos	0.617
MC-Simplex	**0.713**
Linear SVM	0.416

E. Localization in a multi-label dataset

While our method is able to competitively classify pre-segmented images, when compared to the state-of-the-art, in this section, we propose an alternative exploratory paradigm for the association of labels to regions in the image. The purpose of the method presented herein is not to provide competitive state-of-the-art results for semantic segmentation, but merely to build a working prototype that builds on the histogram representatives naturally obtained by our method, and discuss its advantages and current limitations. Recall that in the single-class example of Sec. IV-C, we used each corrected histogram in the training and test set to localize the bounding box containing the most discriminative region. In the multi-label case, however, several classes coexist in one image. Since corrected histograms contain a mixture of classes, they can't be used for class localization in the images.

One possible approach to solve this problem is to pre-segment the test images and use the learned class models to classify each region individually. However, this approach has several drawbacks: 1) having to select a fixed number of segments, 2) the segments are obtained through only texture and color cues, so they might not match the ground truth regions of the classes, and 3) contextual information between segments is lost, which results in poorer classification performance when compared to the classifiers learned on the entire image.

We propose an alternative method that does not suffer from these drawbacks, by explicitly recovering representative histograms for each class. We proceeded as in IV-D, but padded the matrix $\mathbf{Z}$ with 21 extra columns where the labels

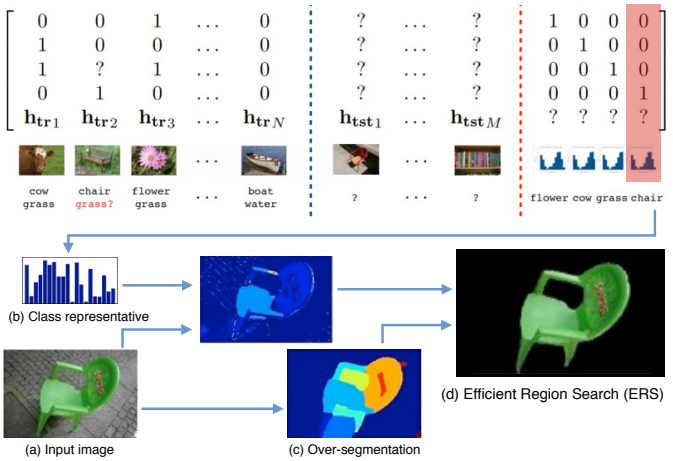


Fig. 9. Illustration of our proposed method for class localization with Matrix Completion.

are the identity and the features are unknown, to recover one representative histogram per class (see Figure 3(c)). Observe that we do not require segmentation for this classification. For each class in an image (Figure 9 (a)), we plot a heatmap of which words belong to the class using its respective histogram (Figure 9(b)). Then, we oversegmented each image using the hierarchical segmentation of Arbelaez *et al*. [58] (Figure 9(c)). We used code provided by the authors and set the parameter boundary segmentation scale to $k = 0.1$. Last, in order to get the localization for a class in an image, we used the class histograms and the obtained segments for that image as the input to the Efficient Region Search (ERS) method of Vijayanarasimhan and Grauman [59]. ERS selects a group of connected segments (Figure 9(d) that maximizes a detection score as measured by an SVM classifier. Since the output of our algorithm is a probability map, we emulated the SVM weight vector by using the class representative subtracted by its mean. We show qualitative results of this approach on Figure 10 for independent recovery of classes in the same image. The failures of our approach can be generally attributed to one of two cases: class confusion in both the classification and the fact that ERS is applied individually to each class (Figure 12(a)); the fact that the solution obtained by ERS is by design a single contiguous region (Figure 12(b)).

V. Proof of convergence of MC-1/Pos/Simplex

This appendix proves the convergence of FPC in Alg. 1 by the fact that projections onto Convex sets are non-expansive; thus, the composition of gradient, shrinkage and projection steps is also non-expansive. Since the problem is convex, a unique fixed point exists in its optimal solution.

Lemma 1: The gradient operator $h(\cdot)$ for (21), (22), (23) is non-expansive for step sizes $\tau \in [0, \min(\frac{4|\Omega_Y|}{\lambda\gamma}, \tau_X|\Omega_X|)]$.

Proof: These values are obtained from (26) by noting the gradient of the Log loss function is Lipschitz continuous with $L = 0.25$ and choosing τ_X such that the χ^2 error, for the Non-Negative Orthant, is Lipschitz continuous with $L = 1$. ■

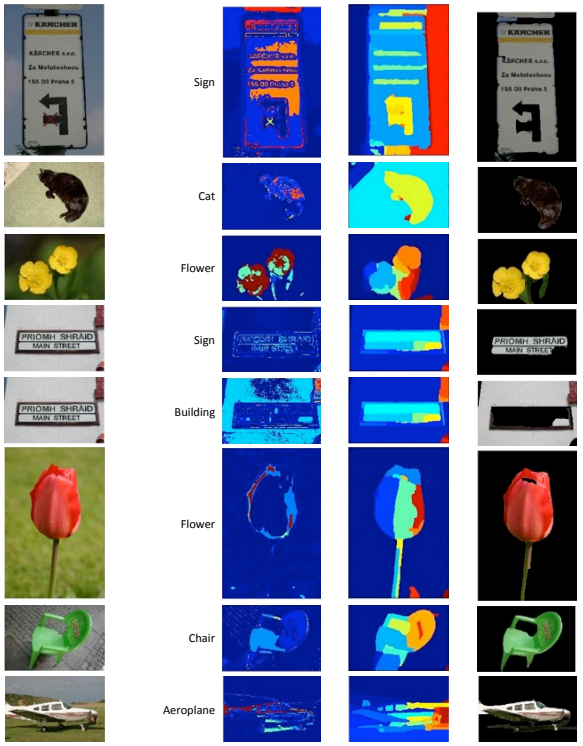


Fig. 10. Histograms corrected by our method in the MSRC dataset preserve semantic meaning. The input image is shown in (a). The heatmap generated by the class representative histogram is shown in (b). ERS [59] uses the heatmap in (b) and the over segmentation in (c) to produce the segmentation in (d).

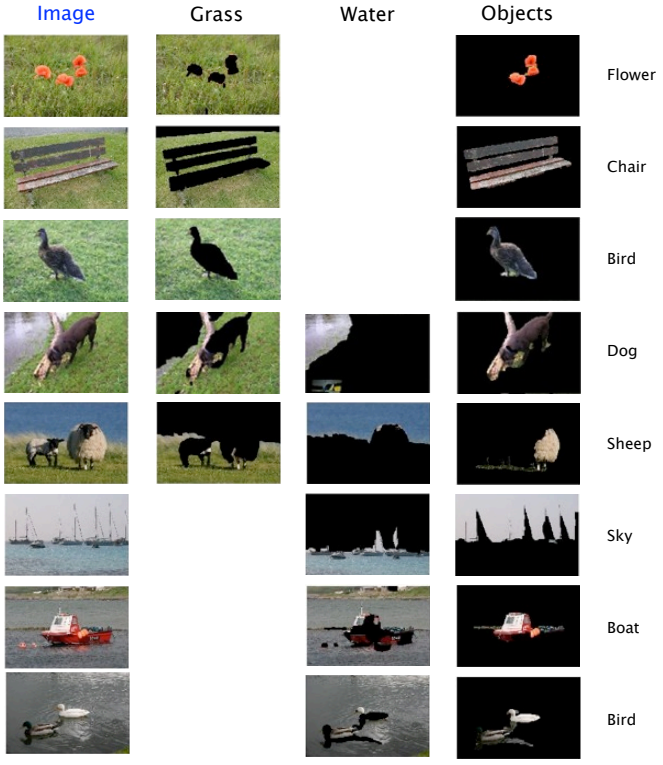
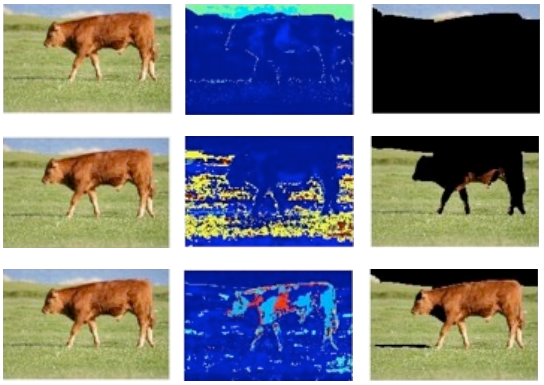
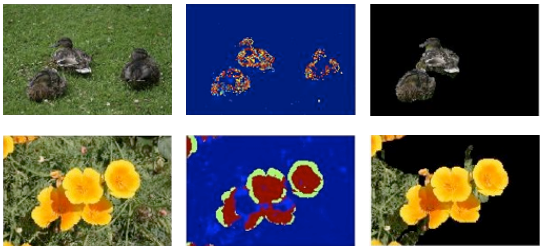


Fig. 11. Multi-label segmentation results on the MSRC dataset.



(a) Class confusion and ERS is not multi-label. Top: Sky, Middle: Grass, Bottom: Cow



(b) ERS result is a contiguous region

Fig. 12. Multi-label segmentation failure cases. Left: Original Image. Middle: Heatmap generated by the class representative histogram. Right: Segmentation obtained by ERS with class representatives.

Lemma 2: Let $p_{\mathcal{C}}(\cdot)$ be a projection operator onto any given convex set $\mathcal{C}$. It follows that $p_{\mathcal{C}}(\cdot)$ is non-expansive and $\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\| = \|\mathbf{Z} - \mathbf{Z}^*\|$ iff $p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) = \mathbf{Z} - \mathbf{Z}^*$.

Proof: For non-expansiveness, [60, Prop. 3.1.3] states that

$$\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2 \leq \langle p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*), \mathbf{Z} - \mathbf{Z}^* \rangle. \quad (28)$$

Applying the Cauchy-Schwarz inequality to (28) yields

$$\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F \leq \|\mathbf{Z} - \mathbf{Z}^*\|_F. \quad (29)$$

For the equivalence part, let us write

$$\begin{aligned} \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) - (\mathbf{Z} - \mathbf{Z}^*)\|_F^2 = \\ \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2 + \|\mathbf{Z} - \mathbf{Z}^*\|_F^2 \\ -2\langle p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*), \mathbf{Z} - \mathbf{Z}^* \rangle, \end{aligned} \quad (30)$$

where the inner product can be bounded by (28), yielding

$$\begin{aligned} \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) - (\mathbf{Z} - \mathbf{Z}^*)\|_F^2 \leq \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2 \\ +\|\mathbf{Z} - \mathbf{Z}^*\|_F^2 - 2\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2. \end{aligned} \quad (31)$$

Since our hypothesis $\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\| = \|\mathbf{Z} - \mathbf{Z}^*\|$, (31) is

$$\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) - (\mathbf{Z} - \mathbf{Z}^*)\|_F^2 \leq 0, \quad (32)$$

from which we conclude an equality is in place. ■

Theorem 3: Let $\mathbf{Z}^*$ be an optimal solution to (24) or (25). Then $\mathbf{Z}$ is also an optimal solution if

$$\|p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| = \|\mathbf{Z} - \mathbf{Z}^*\|. \quad (33)$$

Proof: By non-expansiveness of operators $p_{\mathcal{C}}(\cdot)$, $S_\nu(\cdot)$ and $h(\cdot)$ (Lemma 2 and [16, Lemmas 1,2]), we can write

$$\begin{aligned} \|\mathbf{Z} - \mathbf{Z}^*\| = \|p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| \leq \\ \leq \|S_\nu(h(\mathbf{Z})) - S_\nu(h(\mathbf{Z}^*))\| \leq \\ \leq \|h(\mathbf{Z}) - h(\mathbf{Z}^*))\| \leq \|\mathbf{Z} - \mathbf{Z}^*\|, \end{aligned} \quad (34)$$

so we conclude the inequalities are equalities. Using the second part of the Lemmas, we get

$$\begin{aligned} p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) = \\ = S_\nu(h(\mathbf{Z}^*)) - S_\nu(h(\mathbf{Z})) = h(\mathbf{Z}^*) - h(\mathbf{Z}) = \mathbf{Z} - \mathbf{Z}^*. \end{aligned}$$

Since $\mathbf{Z}^*$ is optimal, by the projected subgradient method and [16, Corollary 1], we have that

$$p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*))) = \mathbf{Z}^* \implies p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) = \mathbf{Z}, \quad (35)$$

from which we conclude $\mathbf{Z}$ is an optimal solution to (20). ■

Theorem 4: A sequence $\{\mathbf{Z}^k\}$ generated by Alg. 1 converges to $\mathbf{Z}^*$, an optimal solution of (24) ((25), resp.).

Proof: We can use the same rationale as in [16, Theorem 4], once we note the non-expansiveness of $p_{\mathcal{C}}(\cdot)$, $S_\nu(\cdot)$ and $h(\cdot)$ ensures the composite operator $p_{\mathcal{C}}(S_\nu(h(\cdot)))$ is also non-expansive. Therefore, the sequence $\{\mathbf{Z}^k\}$ lies in a compact set and must have a limit point, which we define as $\hat{\mathbf{Z}} = \lim_{k\to\infty} \mathbf{Z}^k$. Also, for any solution $\mathbf{Z}^* \in \mathcal{Z}^*$, we have

$$\begin{aligned} \|\mathbf{Z}^{k+1} - \mathbf{Z}^*\| = \|p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^k))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| \leq \\ \leq \|S_\nu(h(\mathbf{Z}^k)) - S_\nu(h(\mathbf{Z}^*))\| \leq \\ \leq \|h(\mathbf{Z}^k) - h(\mathbf{Z}^*))\| \leq \|\mathbf{Z}^k - \mathbf{Z}^*\|, \end{aligned} \quad (36)$$

so we conclude the sequence $\{\|\mathbf{Z}^k - \mathbf{Z}^*\|\}$ is monotonically non-increasing and culminates in any limit point $\hat{\mathbf{Z}}$, *i.e.*,

$$\lim_{k\to\infty} \|\mathbf{Z}^k - \mathbf{Z}^*\| = \|\hat{\mathbf{Z}} - \mathbf{Z}^*\|. \quad (37)$$

On the other hand, by the continuity of $p_{\mathcal{C}}(S_\nu(h(\cdot)))$, we have that the image of $\hat{\mathbf{Z}}$ is

$$p_{\mathcal{C}}(S_\nu(h(\hat{\mathbf{Z}}))) = \lim_{k\to\infty} p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^k))) = \lim_{k\to\infty} \mathbf{Z}^k = \hat{\mathbf{Z}} \quad (38)$$

is also a limit point of $\{\mathbf{Z}^k\}$, yielding

$$\|p_{\mathcal{C}}(S_\nu(h(\hat{\mathbf{Z}}))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| = \|\hat{\mathbf{Z}} - \mathbf{Z}^*\|, \quad (39)$$

from which we can recall Theorem 3.

■

VI. Conclusions and Future Work

Weakly supervised learning algorithms allow for learning image classification models without recurring to labeling efforts based on bounding boxes or full-blown pixelwise segmentations. It has been shown that these labeling efforts are not only expensive, but subjective and error prone. Thus, the importance of reducing manual labeling from region labeling to image labeling is critical for the applicability of image classification methods, especially in large scale datasets with many classes. Limitations of existing MIL approaches include

their non-convexity and reliance on an explicit enumeration of possible regions given by a segmentation algorithm.

A key idea of our method is that histograms of full images contain the information for parts contained therein, so the weakly supervised learning problem can be formulated as a blind source separation problem, solved using a matrix completion framework. Three are the main benefits of our approach: First, unlike existing MIL approaches to weakly-supervised learning, we presented two new convex methods for performing multi-label classification of histogram data, with proven convergence properties. Second, unlike the majority of existing classifiers, we showed that matrix completion allows for handling of missing data, labeling errors, background noise and partial occlusions. Third, we were able to find class histogram representations and provide localization in the images. This is done despite of the weakly-supervised training set, where class locations are unknown.

Experiments show that our convex methods perform comparably or better than state-of-the-art MIL methods in several image datasets. Our feature error correction provides superior results for weakly supervised multi-label classification, when compared to explicitly enumerating possible regions in the image using segmentation algorithms or bounding box localization. When annotating individual regions, our method was only surpassed by a non-linear MIL method. Error correction also allows to perform localization of the discriminative regions of the image in a single class problem. Class representative histograms allow for class localization in multi-label problems.

We note that our approach is not a full replacement for MIL, since in other settings features may not respect the low rank assumptions in Section III. Despite not requiring segmentation for classification, our approach has the limitation of only capturing one representative histogram per class (Figure 3 (a)). Future work should address the extension of this framework to allow for the use of representative subspaces. As an extension of a component analysis technique, this work should be kernelized, to couple the feature error correction and the use of non-linear techniques into a single technique.

ACKNOWLEDGMENT

Support for this research was provided by the FCT (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal program under grant FCT/CMU/P11. Partially funded by FCT projects Printart PTDC/EEA-CRO/098822/2008 and PEst-OE/EEI/LA0009/2013 and project Poeticon++ from the European FP7 program (grant agreement no. 288382). Fernando De la Torre is partially supported by Grant CPS-0931999 and NSF IIS-1116583. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *European Conference on Computer Vision*, 2010.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[3] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[4] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in *International Conference on Computer Vision*, 2009.

[5] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Computer Vision and Pattern Recognition*, 2006.

[6] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Computer Vision and Pattern Recognition*, 2003.

[7] M. Jamieson, A. Fazly, S. Dickinson, S. Stevenson, and S. Wachsmuth, "Learning Structured Appearance Models from Captioned Images of Cluttered Scenes," in *International Conference on Computer Vision*, 2007.

[8] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Computer Vision and Pattern Recognition*, 2006.

[9] Z.-h. Zhou and M. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Neural Information Processing Systems*, 2006.

[10] Z.-j. Zha, X.-s. Hua, T. Mei, J. Wang, and G.-j. Q. Zengfu, "Joint multi-label multi-instance learning for image classification," in *Computer Vision and Pattern Recognition*, 2008.

[11] S. Vijayanarasimhan and K. Grauman, "What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations," in *Computer Vision and Pattern Recognition*, 2009.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[13] F. Li and C. Sminchisescu, "Convex Multiple Instance Learning by Estimating Likelihood Ratio," in *Neural Information Processing Systems*, 2010.

[14] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision and Pattern Recognition*, 2003.

[15] E. Candes and B. Recht, "Exact low-rank matrix completion via convex optimization," in *Allerton Conference on Communication, Control, and Computing*, 2008.

[16] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[17] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *Neural Information Processing Systems*, 2011.

[18] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *International Conference on Computer Vision*, 2001.

[19] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *International Conference on Computer Vision*, 2009.

[20] S. Petridis, W. Liu, and J. Pessiot, "Localizing Objects while Learning Their Appearance," in *European Conference on Computer Vision*, 2010.

[21] M. Blaschko and C. Lampert, "Learning to localize objects with structured output regression," in *European Conference on Computer Vision*, 2008.

[22] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *European Conference on Computer Vision*, 2010.

[23] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31 – 71, 1997.

[24] O. Maron and A. Ratan, "Multiple-instance learning for natural scene classification," in *International Conference on Machine Learning*, 1998.

[25] A. Vezhnevets and J. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *Computer Vision and Pattern Recognition*, 2010.

[26] O. Yakhnenko and V. Honavar, "Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies," in *British Machine Vision Conference*, 2011.

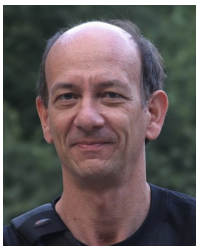
[27] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, "Who's in the Picture?" in *Neural Information Processing Systems*, 2004.

[28] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *American Conference of Control*, 2001.

[29] F. Xiong, O. I. Camps, and M. Sznaier, "Dynamic context for tracking behind occlusions," in *European Conference on Computer Vision*, 2012.
[30] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.
[31] Z. Lin and M. Chen, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," *Technical Report UILU-ENG-09-2215*, 2009.
[32] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization*, vol. 20(4), pp. 1956–1982, 2008.
[33] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems," *Pacific Journal of Optimization*, vol. 6, pp. 615–640, 2010.
[34] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Information Theory*, vol. 56, pp. 2980–2998, 2010.
[35] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Allerton Conference on Communication, Control, and Computing*, 2010.
[36] R. S. Cabral, J. P. Costeira, F. De la Torre, and A. Bernardino, "Fast incremental method for matrix completion: an application to trajectory correction," in *International Conference on Image Processing*, 2011.
[37] Y. Dai, H. Li, and M. He, "Element-wise factorization for n-view projective reconstruction," in *European Conference on Computer Vision*, 2010.
[38] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *Neural Information Processing Systems*, 2009.
[39] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Computer Vision and Pattern Recognition*, 2011.
[40] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *International Conference on Computer Vision*, 2011.
[41] G. Zhu and S. Yan, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *International Conference on Multimedia*, 2010.
[42] Z. Zhang, Y. Matsushita, and Y. Ma, "Camera calibration with lens distortion from low-rank textures," in *Computer Vision and Pattern Recognition*, 2011.
[43] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Bilinear classifiers for visual recognition," in *Neural Information Processing Systems*, 2009.
[44] N. Loeff and A. Farhadi, "Scene discovery by matrix factorization," in *European Conference on Computer Vision*, 2008.
[45] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face Liveness Detection from A Single Image with Sparse Low Rank Bilinear Discriminative Model," in *European Conference on Computer Vision*, 2010.
[46] A. B. Goldberg, X. Zhu, B. Recht, J. ming Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *Neural Information Processing Systems*, 2010.
[47] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick, "Large-scale image classification with trace-norm regularization," in *Computer Vision and Pattern Recognition*, 2012.
[48] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *European Conference on Computer Vision*, 2006.
[49] D. Huang, R. Cabral, and F. De la Torre, "Robust regression," in *European Conference on Computer Vision*, 2012.
[50] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions," in *International Conference on Machine Learning*, 2008.
[51] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2368–2382, 2011.
[52] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Computer Vision and Pattern Recognition*, 2010.
[53] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *European Conference on Computer Vision*, 2006.
[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005.
[55] T. Mitchell, *Machine Learning*. McGraw-Hill Education, 1997.
[56] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[57] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000.
[58] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. on Pattern Anaalysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, May 2011.
[59] S. Vijayanarasimhan and K. Grauman, "Efficient region search for object detection," in *Computer Vision and Pattern Recognition*, 2011.
[60] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer-Verlag, 2001.

Ricardo Cabral a PhD student from Carnegie Mellon and IST-Lisbon. He received his Masters in Electrical and Computer Eng. at IST-Lisbon and a research grant from the Portuguese Science Foundation in 2009, for work in correspondence methods for structure from motion. He received an outstanding academic achievement award in 2008 from IST-Lisbon, where he worked in several projects, including video handling for the 2012 London Olympics. His research focuses on low rank models for computer vision and machine learning.

Fernando De la Torre is an Associate Research Professor in the Robotics Institute at Carnegie Mellon University. He received his B.Sc. degree in Telecommunications, as well as his M.Sc. and Ph. D degrees in Electronic Engineering from La Salle School of Engineering at Ramon Llull University, Barcelona, Spain in 1994, 1996, and 2002, respectively. His research interests are in the fields of computer vision and Machine Learning. Currently, he is directing the Component Analysis Laboratory (http://ca.cs.cmu.edu) and the Human Sensing Laboratory (http://humansensing.cs.cmu.edu) at Carnegie Mellon University. He has over 130 publications in referred journals and conferences. He has organized and co-organized several workshops and has given tutorials at international conferences on the use and extensions of Component Analysis.



João Paulo Costeira is an Associate Professor of Electrical and Computer Eng. at Instituto Superior Técnico (IST), Portugal. He obtained his PhD from IST in 1995. From 1992-1995 he was a member of Canegie Mellon's Vision and Autonomous Systems Center. He is a researcher at Instituto de Sistemas e Robotica (ISR) since 1994 and Scientific coordinator of the thematic area " Signal Processing for Communications Networks and Multimedia" of the ISR-Associated Lab. Currently he is co-director of the CMU|Portugal Dual PhD Program in ECE, an initiative funded by the Portuguese Fundação de Ciência e Tecnologia.

Alexandre Bernardino received the PhD degree in Electrical and Computer Engineering in 2004 from Instituto Superior Técnico (IST). He is an Assistant Professor at IST and Researcher at the Institute for Systems and Robotics (ISR-Lisboa) in the Computer Vision Laboratory (VisLab). He participates in several national and international research projects in the fields of robotics, cognitive systems, computer vision and surveillance. He published several articles in international journals and conferences, and his main research interests focus on the application of computer vision, cognitive science and control theory to advanced robotic and surveillance systems.