Automatic discovery of discriminative parts as a quadratic assignment problem

Ronan Sicre
IRISA, France

Julien Rabin
Normandie Univ., France

Yannis Avrithis
Inria, France

Teddy Furon
Inria, France

Frédéric Jurie
Normandie Univ., France

Ewa Kijak
Univ. Rennes 1 / IRISA, France

Abstract

Part-based image classification consists in representing categories by small sets of discriminative parts upon which a representation of the images is built. This paper addresses the question of how to automatically learn such parts from a set of labeled training images. We propose to cast the training of parts as a quadratic assignment problem in which optimal correspondences between image regions and parts are automatically learned. The paper analyses different assignment strategies and thoroughly evaluates them on two public datasets: Willow actions *and* MIT 67 scenes.

1. Introduction

The representation of images as set of patches has a long history in computer vision, especially for object recognition [2], image classification [9, 36] or object detection [18]. Its biggest advantages are the robustness to spatial transformations (rotation, scale changes, etc.) and the ability to focus on the important information of the image while discarding clutter and background.

Part-based classification raises the questions of i) how to automatically identify what are the parts to be included in the model and ii) how to use them to classify a query image. The work of [38] selects informative patches using an entropy based criterion while the decision relies on a Bayes classifier. Following [38], recent approaches separate the construction of the model (i.e. the learning of the parts) and the decision function [14, 8].

Jointly optimizing modeling and classification is however possible for simple enough part detectors and decision functions [24]. This work aims at defining the parts directly related to the final classification function.

While this argument is understandable, the objective function of this joint optimization is highly non-convex with no guaranty of convergence. Deciding which alternative is better – the joint or separate design – is still an open problem. As an insight, the two stage part-based model of [30] performs better than the joint learning of [24]. We note other differences: [24] models both positive and negative parts while [30] focuses only on the positive ones.

Interestingly, [30, 29] addresses the learning of parts as an assignment problem. Regions are sampled randomly from the training images, and a class is modeled as a set of parts. The assignment region-part is constrained: each part is assigned to one region in each positive image (belonging to the class to be modeled). This yields a bipartite graph. Solving the learning of part-based models via an assignment problem is appealing, yet solution [30] is based on heuristics leaving room for improvements.

Our contribution is an extensive study of this assignment problem: We present a well-founded formulation of the problem and propose different solutions in a rigorous way. We revisit the model of [30] and introduce an alternative constraint of *one-to-many assignment*, where a part may be assigned to more than one regions in each image. We cast part learning as a *quadratic assignment problem*, and study a number of convex relaxations and optimization algorithms.These methods are evaluated and compared on two different public datasets and we demonstrate that our methodology remains complementary to the powerful visual representations obtained by state of the art deep learning approaches.

The paper is organized as follows: Section 2 gives the related works, Section 3 presents our new formulation. Then, Section 4 discusses convex relaxations, while Section 5 introduces several optimization algorithms. Finally, Section 6 is devoted to the experimental validation.

2. Previous work

Image classification has received a lot of attention during the last decades. The literature used to focus on models based on aggregated features [6, 25] or the Spatial Pyramid Matching [16]. This was before the Convolutional Network revolution [15] at the heart of the recent methods [31].

Several authors have investigated part-based models in which some parts of the image are combined in order to

determine if a given object is depicted. This is in contrast to aggregation approaches where the image regions are pooled without selecting the discriminative parts. For instance, [32] discovers sets of regions used as mid-level visual representation; the regions are selected for being representative (occurring frequently enough) and discriminative (different enough from others). This iterative procedure alternates between clustering and training classifiers. Similarly, [14] learns parts incrementally, starting from a single part occurrence with an Exemplar SVM and collecting more and more occurrences from the training images. Also, [23] propose to learn discriminative parts with LSSVM.

In a different way, [8] poses the discovery of visual elements as a discriminative mode seeking problem solved with the mean-shift algorithm. This method discovers visually-coherent patch clusters that are maximally discriminative. The work of [21] investigates the problem of parts discovery when some correspondences between instances of a category are known. The work of [34] bears several similarities to our work in the encoding and classification pipeline. However, parts are assigned to regions using spatial max pooling without any constraint.

The part-based representation of [24] relies on the joint learning of informative parts (using heuristics that promote distinctiveness and diversity) and linear classifiers trained on vectors of part responses. On the other hand, Sicre *et al* [30] follow the two stage design, formulating the discovery of parts as an assignment problem. Recently, Mettes *et al* [22] argue that image categories may share parts and they propose a method taking into account this redundancy.

Finally, this paper uses algorithms finding the assignment maximizing the total weight in a bipartite graph. A survey on this topic is the work of Burkard *et al* [3].

3. Discovering and Learning Parts

Our approach comprises three steps: (i) distinctive parts are discovered and learned, (ii) a global image signature is computed based on the presence of these parts, and (iii) the signature is classified by a linear SVM. This paper focuses on the first step. For each class, we learn a set of P distinctive parts which are representative and discriminative.

This section formalizes the parts learning problem in different ways giving birth to interesting optimization alternatives in Sect. 5. We show that it boils down to a concave minimization under non convex constraints, which can be cast as a quadratic assignment problem.

3.1. Notation

Column vector $\mathrm{vec}(X)$ contains all elements of matrix X in column-wise order. Given matrices X, Y of the same size, $\langle X, Y\rangle = \sum_{i,j} X_{ij}Y_{ij}$ is their (Frobenius) inner product, $\|X\|$ and $\|X\|_F = \sqrt{\langle X, X\rangle}$ are the spectral and Frobenius norms. Vector $x_{i\bullet}^\top$ ($x_{\bullet j}$) denotes the i-th row (resp. j-th column) of matrix X. Vector $\mathbf{1}_n$ (matrix $\mathbf{1}_{m\times n}$) is an $n \times 1$ vector (resp. $m \times n$ matrix) of ones. The dot product between vectors x and y is also denoted by $\langle x, y\rangle$. $\mathbb{1}_{\mathcal{A}}$ is the indicator function of set $\mathcal{A}$ and $\mathrm{Proj}_{\mathcal{A}}$ is the Euclidean projector onto $\mathcal{A}$.

Following [30], we denote by $\mathcal{I}^+$ the set of n^+ images of the class to be modeled, i.e. positive images, while $\mathcal{I}^-$ represents the negative images. The training set is $\mathcal{I} = \mathcal{I}^+ \cup \mathcal{I}^-$ and contains n images. A set of regions $\mathcal{R}_I$ is extracted from each image $I \in \mathcal{I}$. The number of regions per image is fixed and denoted $|\mathcal{R}|$. The total number of regions is thus $R = n|\mathcal{R}|$. $\mathcal{R}^+$ is the set of regions from positive images whose size is $R^+ = n^+|\mathcal{R}|$.

Each region $r \in \mathcal{R}_I$ is represented by a descriptor $x_r \in \mathbb{R}^d$. In this work, this descriptor is obtained by a CNN (see Sect. 6.2). By X (X^+) we denote the $d \times R$ (resp. $d \times R^+$) matrix whose columns are the descriptors of the complete training set (resp. positive images only).

3.2. Problem setting

A class is modeled by a set of parts $\mathcal{P} \subset \mathbb{R}^d$ with $|\mathcal{P}| = P$. The $P \times R^+$ matching matrix M associates regions of positive images to parts. Element m_{pr} of M corresponds to region r and part p. Ideally, $m_{pr} = 1$ if region r is deemed to represent part p, and 0 otherwise. For a given image I, we denote by M_I the $P \times |\mathcal{R}|$ submatrix of M that contains columns $r \in \mathcal{R}_I$.

We remind the requirements of [30]: (i) the P parts are different from one another, (ii) each part is present in every positive image, (iii) parts occur more frequently in positive images than in negative ones. The first two requirements define the following subset of $\mathbb{R}^{P\times R^+}$:

$$\mathcal{M}_1 \triangleq \left\{ M^\top \mathbf{1}_P \leq \mathbf{1}_{R^+} \text{ and } M_I \mathbf{1}_{|\mathcal{R}|} = \mathbf{1}_P, \forall I \in \mathcal{I}^+ \right\}. \tag{1}$$

We note that the constraint forcing columns to sum to maximum one encourage regions to be assigned to at most one part. Since we wish M to represent a one-to-one assignment of regions to parts, the admissible set of M is $\mathcal{A}_1 \triangleq \{0,1\}^{P\times R^+} \cap \mathcal{M}_1$. Note that set $\mathcal{A}_1$ is not convex.

The third requirement is enforced by Linear Discriminant Analysis (LDA): given M, the model $w_p(M) \in \mathbb{R}^d$ of part p is defined as

$$\begin{aligned} w_p(M) &\triangleq \Sigma^{-1}\left(\frac{\sum_{r\in\mathcal{R}^+} m_{pr}x_r}{\sum_{r\in\mathcal{R}^+} m_{pr}} - \mu\right) \\ &= \Sigma^{-1}\left(\frac{1}{n^+}X^+ m_{p\bullet} - \mu\right), \end{aligned} \tag{2}$$

where $\mu \triangleq \frac{1}{n}X\mathbf{1}_R$ and $\Sigma \triangleq \frac{1}{n}(X - \mu\mathbf{1}_R^\top)(X - \mu\mathbf{1}_R^\top)^\top$ are the empirical mean and covariance matrix of region descriptors over all training images. The similarity between region r and a part p is then computed as $\langle w_p(M), x_r\rangle$.

For a given class, we are looking for the optimal matching matrix $M^\star \in \arg\max_{M\in\mathcal{A}_1} J(M)$ with

$$J(M) \triangleq \sum_{p\in\mathcal{P}, r\in\mathcal{R}^+} m_{pr}\langle w_p(M), x_r\rangle. \tag{3}$$

For a given class, we define $W(M)$ as the $d \times P$ matrix whose columns are $w_p(M)$ for all parts $p \in \mathcal{P}$, and the similarity matrix $C(M) \triangleq W(M)^\top X^+$. This matrix stores the $P \times R^+$ similarities between parts and regions. In the end, we can compactly rewrite the objective function as $J(M) = \langle M, C(M)\rangle$.

In this work, we further deviate from the original requirement (ii) of [30] by observing that each part may not only be present once in every positive image, but with more than one instances allowed. It is expected, for instance, to find more than one chair in an office scene. The case of overlapping regions with similar descriptors is also common. With this modification, the first two requirements define the subset of $\mathbb{R}^{P\times R^+}$

$$\begin{aligned}\mathcal{M}_\kappa \triangleq \quad & \left\{M^\top \mathbf{1}_P \leq \mathbf{1}_{R^+}\right\} \cap \\ & \left\{\mathbf{1}_P \leq M_I \mathbf{1}_{|\mathcal{R}|} \leq \kappa \mathbf{1}_P, \forall I \in \mathcal{I}^+\right\}.\end{aligned} \tag{4}$$

Observe that $\mathcal{M}_1$ defined in (1) is a special case for $\kappa = 1$, representing a *one-to-one* assignment. On the other hand, for $\kappa > 1$, $\mathcal{M}_\kappa$ represents a *one-to-many* assignment between parts and regions. This enables assigning a part to up to κ regions, provided that $\kappa \leq K \triangleq \frac{|\mathcal{R}|}{P}$. We assume $|\mathcal{R}|$ is a multiple of P. The admissible set of M becomes $\mathcal{A}_\kappa \triangleq \{0,1\}^{P\times R^+} \cap \mathcal{M}_\kappa$.

3.3. Recasting as a quadratic assignment problem

Paper [30] solves the problem by alternatively resorting to (2) and (3). Here, we rewrite the objective function J as a function of M only by injecting an explicit expression of $W(M)$. This gives birth to a quadratic assignment problem, which allows a number of alternative algorithms as detailed in the next section. According to LDA (2), $W(M) = \Sigma^{-1}\left(\frac{1}{n^+}X^+M^\top - \mu\mathbf{1}_P^\top\right)$, which in turn gives

$$C(M) = MA - B, \tag{5}$$

where $R^+ \times R^+$ matrix $A = \frac{1}{n^+}X^{+\top}\Sigma^{-1}X^+$ is symmetric and positive definite and $P \times R^+$ matrix $B = \mathbf{1}_P\mu^\top\Sigma^{-1}X^+$ has identical rows (rank 1). Our problem becomes equivalent to finding $M^\star \in \arg\min_{M\in\mathcal{A}_\kappa} J_0(M)$

$$\begin{aligned}J_0(M) &\triangleq \langle M, B - MA\rangle \\ &= \operatorname{vec}(M)^\top Q \operatorname{vec}(M) + \operatorname{vec}(B)^\top \operatorname{vec}(M),\end{aligned} \tag{6}$$

for a $PR^+ \times PR^+$ matrix Q that is only a function of A. This shows that our task is closely related to the *quadratic assignment problem* [3], which is NP-hard. But, the objective function $J_0(M)$ is strictly concave.

This new formalism enables to leverage a classical procedure in optimization: the convex relaxation.

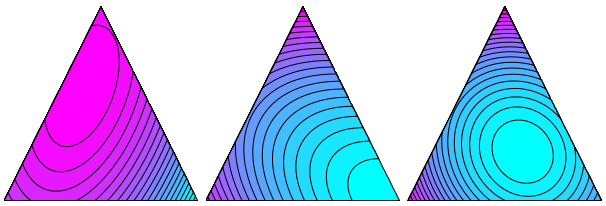

Figure 1. **Illustration of the convex relaxation of our assignment problem in 3D.** *Black lines are level-sets of the objective function J_0 in the plane of the simplex, which is a triangle in $\mathbb{R}^3$. Lower values are displayed in cyan, larger in magenta. (Left) The original problem is the minimization of a concave quadratic function that lies on the vertices of the simplex. (Middle) A small quadratic regularization of the objective function together with the relaxation of the binary constraint preserves the solution. (Right) A too large regularization yet shifts the minimum inside the simplex, thus giving less sparse solutions.*

4. Convex relaxation

It is common to relax the constraint of M being binary: $M^\star = \arg\min_{M\in\mathcal{S}_\kappa} J_0(M)$ with the admissible set of M relaxed to $\mathcal{S}_\kappa \triangleq [0,1]^{P\times R^+} \cap \mathcal{M}_\kappa$, with $\kappa = 1$ in the one-to-one case and $1 < \kappa \leq K$ in the one-to-many case. Domain $\mathcal{S}_\kappa$ is the convex hull of $\mathcal{A}_\kappa$ and we refer to $\mathcal{S}_\kappa$ to as a *κ-simplex*. Unless otherwise stated, we assume this convex relaxation below.

A convex relaxation of the objective function is also common. We examine two regularization methods in the following.

4.1. Entropic regularization

Entropy regularization can be used to approximate the binary constraint by considering the objective function

$$J'_\beta(M) \triangleq \langle M, B - MA\rangle - \frac{1}{\beta}H(M) \tag{7}$$

where $H(M) = -\langle \log(M), M\rangle$ is the entropy of matrix M, and $\beta > 0$ is the regularization parameter.

In the simpler case where $C(M)$ is a *fixed cost matrix*, the minimization over $\mathcal{S}_1$ becomes tractable and is referred to as *soft assignment*. This problem has gained a lot attention because it can be solved efficiently at large scales [33]. However, a major limitation is the loss of sparsity of the solution. We describe in the section 5.2 how the authors of [30] have circumvented this problem by *iterative soft assignment* (ISA), also without assuming C fixed.

Dedicated methods are also common in the case of fixed C. The Hungarian algorithm examined in section 5.1 gives an exact solution to the *linear (hard) assignment problem*, without the relaxation of the binary constraint.

4.2. Quadratic regularization

We consider now the quadratic regularization of the problem, see Figure 1 for an illustration:

$$J_\rho(M) \triangleq \langle M, B - MA\rangle + \rho\|M\|_F^2 \quad (8)$$
$$= J_0(M) + \rho\, P\, n^+, \quad (9)$$

where (9) holds provided $M \in \mathcal{A}_1$. In this case, $J_\rho(M)$ and $J_0(M)$ differ by a constant. Therefore, the minimizers of J_ρ on $\mathcal{A}_\kappa$ are the minimizers of J_0, for any value of ρ. Indeed, if ρ is sufficiently large ($\rho > \|A\|$), J_ρ becomes convex (see Fig. 1). In general however, $J_\rho(M) \neq J_0(M) + \rho P n^+$ when $M \in \mathcal{S}_\kappa \setminus \mathcal{A}_\kappa$. We may find different solutions as illustrated in Fig. 1.

Over-relaxing the problem for the sake of convexity is not interesting as it promotes parts described by many regions instead of a few ones. Indeed, when $\rho > \|A\|$, the minimum of J_ρ is achieved for the rank-1 matrix $\frac{1}{2}B(A - \rho I_{R^+})^{-1}$, which may lie inside $\mathcal{S}$. Conversely, when ρ is negative, the regularization term acts as a force towards the set $\mathcal{A}_\kappa$, driving the solution to a binary matrix which may be an interesting way to avoid the aforementioned problem of over relaxing the constraints.

5. Optimization

The previous section formalizes the part learning task as an optimization problem. This section now presents two alternatives to numerically solve them: (i) hard assignment optimization directly finding $M^\star \in \mathcal{A}_\kappa$, (ii) soft assignment optimization (Sect. 4.1 and 4.2) finding $M^\star \in \mathcal{S}_\kappa$. This latter strategy is not solving the initial problem. However, as already observed in [30] and [19] for classification, soft-assignment may provide good performance. This observation deserves an experimental investigation in our context.

5.1. Hungarian methods

Hungarian Algorithm (Hun): When the cost matrix $C(M)$ is fixed and consists of n^+ square blocks (i.e. $P = |\mathcal{R}|$), the minimization of $J_0(M)$ onto $\mathcal{A}_1$ is a linear program which solves a bipartite graph matching. Several dedicated methods give an exact solution, including the well-known Hungarian algorithm with $O(P^3)$ complexity [3]. Starting from an initial guess M_0 (see Sect. 6.2), this solution can be seen as computing the orthogonal projection of matrix $C(M_0)$ onto $\mathcal{A}_1$

$$M^\star_{hun} \triangleq \text{Proj}_{\mathcal{A}_1}(C(M_0)) = \underset{M\in\mathcal{A}_1}{\text{argmax}}\ \langle M, C(M_0)\rangle. \quad (10)$$

In our setting, M is not square as we consider partial assignments between P rows and $|\mathcal{R}| > P$ columns per image. In the case of one-to-one assignment $\kappa = 1$, a simple trick is to add an extra row which sums to $|\mathcal{R}| - P$ and to define a maximal cost value when affecting columns to it [1].

To achieve one-to-many assignment for $1 < \kappa \leq K$, we still add rows but define cost as follows. Since the number of columns $|\mathcal{R}|$ per image is a multiple of the number of rows P, we define an $|\mathcal{R}| \times R^+$ cost matrix consisting of κ blocks equal to $C(M_0)$ stacked vertically, while the extra row that sums to $(K - \kappa)P$ has a constant maximal cost value. Observe that this does not solve the problem onto $\mathcal{A}_\kappa$. Rather, constraint $\mathbf{1}_P \leq M_I \mathbf{1}_{|\mathcal{R}|} \leq k\mathbf{1}_P$ in (1) is replaced by $M_I \mathbf{1}_{|\mathcal{R}|} = \kappa \mathbf{1}_P$. Hence its solution is sub-optimal. We refer to this approach as Hun$^\kappa$.

We use the fast Hungarian algorithm variant of [1]. The experimental section shows that this method gives surprisingly good results in comparison to more sophisticated methods.

Integer Projected Fixed Point (IPFP): The IPFP method [17] can be seen as the iteration of the previous method, alternating between updates of the similarity matrix $C(M)$ and projections onto the constraints set $\mathcal{A}_1$. More precisely, a first order Taylor approximation of the objective function is maximized (e.g. by the Hungarian algorithm) and combined with a linesearch (see Algorithm 1). This approach guarantees the convergence to a local minimizer of $J(M)$ on the set $\mathcal{A}_1$.

Algorithm 1 IPFP algorithm

Init: M_0, **set:** $k \leftarrow 0$, $M_{-1} \leftarrow \emptyset$
while $M_{k+1} \neq M_k$ **do**
  $k \leftarrow k + 1$
  $G_k \leftarrow 2M_k A - B$ (gradient $\nabla J(M_k)$)
  $P_{k+1} \leftarrow \text{Proj}_{\mathcal{A}_1}(G_k)$
  (projection using partial Hungarian algorithm [1])
  $\Delta_{k+1} \leftarrow P_{k+1} - M_k$
  $c_k \leftarrow \langle G_k, \Delta_{k+1}\rangle$
  $d_k \leftarrow \langle \Delta_{k+1} A, \Delta_{k+1}\rangle$
  $t_k = \min(-\frac{c_k}{2d_k}, 1)$ if $d_k < 0$ and $t_k = 1$ otherwise
  $M_{k+1} \leftarrow t_k P_{k+1} + (1 - t_k)M_k$ (linesearch)
end while
Output: P_k

We observed that IPFP converges very fast nevertheless results are not improving. This is explained by the specific structure of our problem where the quadratic matrix Q of (6) is sparse and negative definite.

5.2. Iterative Soft-assignment (ISA)

The strategy of [30] referred to as *Iterative Soft-Assign* (ISA) solves a sequence of approximated linear assignment problems. It is based on the rationale: if we better detect regions matching a part, we will better learn that part; if we better learn a part, we will better detect region matching that part. Hence, the approach iteratively assigns regions to parts

by yielding a M for a given $C(M)$ (Sect. 4.1) and learns the parts by yielding $W(M)$ for a given M thanks to LDA (2). The assignment resorted to a soft-assign algorithm, see [28] for instance, which is also an iterative algorithm solving a sequence of entropic-regularized problems (Sect. 4.1) that converges to the target one. The general scheme of the algorithm is drawn in Algorithm 2.

Algorithm 2 ISA algorithm

Init: $M = M_0$
while $M \notin \mathcal{A}_1$ **do**
  $\beta \leftarrow \beta \times \beta_r$ (decreases regularization)
  while M has not converged **do**
    update $C(M)$ using definition (5)
    update M by solving Soft-Assignment problem (7)
  end while
end while

The approach suffers from two major drawbacks: it is computationally demanding due to the three intricate optimization loops, and it is numerically very difficult to converge to an hard-assignment matrix (due to the entropy regularization). Yet, as reported in [30], the latter limitation turns out to be an advantage for this classification task. Indeed, the authors found out that early stopping the algorithm actually improves the performance. However, the obtained matrix M does not satisfy the constraints (neither $\mathcal{A}^\kappa$ nor $\mathcal{S}_\kappa$).

5.3. Quadratic soft assignment with Generalized Forward Backward (GFB)

To address the relaxed and regularized problem which minimize J_ρ over the set $\mathcal{S}_\kappa$, we split the constraints on the matching matrix M for rows and columns. Assuming the matrix $M = (m_{pr})_{p\in P, r\in R^+}$ has non-negative values, for each row $m_{p\bullet}$ and each column $m_{\bullet r}$ we have

- $m_{p\bullet} \in \mathbb{P}_\kappa \triangleq \{x \in \mathbb{R}^{|\mathcal{R}|} : \langle x, \mathbf{1}_{|\mathcal{R}|}\rangle \in [1, \kappa]\}$ is a vector summing between 1 and κ;
- $m_{\bullet r} \in \mathbb{P}_{\leq 1} \triangleq \{x \in \mathbb{R}_+^P : \langle x, \mathbf{1}_P\rangle \leq 1\}$ is a vector that sums at most to 1;

The optimization problem can then be written as

$$\underset{M=M_1=M_2=M_3 \in \mathbb{R}^{P\times R^+}}{\operatorname{argmin}} J_\rho(M) + \sum_{i=1}^{3} G_i(M_i) \qquad (11)$$

where functions G_i, $i = 1..3$ respectively encode constraints on parts, regions and non-negativity:

$$\begin{cases} G_1(M) & = \sum_{p\in\mathcal{P}} \mathbb{1}_{\{m_{p\bullet}\in\mathbb{P}_\kappa\}} \\ G_2(M) & = \sum_{I\in\mathcal{I}^+, r\in\mathcal{R}(I)} \mathbb{1}_{\{m_{\bullet r}\in\mathbb{P}_{\leq 1}\}} \\ G_3(M) & = \sum_{p\in\mathcal{P}, r\in R^+} \mathbb{1}_{\{m_{p,r}\geq 0\}} \end{cases}.$$

The Generalized Forward Backward (GFB) algorithm [27], described in Alg. 3, alternates between explicit gradient descent on the primal problem and implicit gradient ascent on the dual problem. It offers theoretical convergence guaranties in the convex case and can benefit from parallelization.

Algorithm 3 GFB$^\kappa_\rho$ algorithm for problem (11)

$M, M^1, M^2, M^3 \leftarrow M_0$ (initialization)
while not converge **do**
  $G \leftarrow \nabla J_\rho(M) = 2MA_\rho + B$ (gradient)
  update M^1: $\forall p \in \mathcal{P}$
$m^1_{p\bullet} \leftarrow m^1_{p\bullet} + \tau\left(\text{Proj}_{\mathbb{P}_\kappa}\left(2m_{p\bullet} - m^1_{p\bullet} - \frac{1}{L}G_{p\bullet}\right) - m_{p\bullet}\right)$
  update M^2: $\forall r \in \mathcal{R}^+$
$m^2_{\bullet r} \leftarrow m^2_{\bullet r} + \tau\left(\text{Proj}_{\mathbb{P}_{\leq 1}}\left(2m_{\bullet r} - m^2_{\bullet r} - \frac{1}{L}G_{\bullet r}\right) - m_{\bullet r}\right)$
  update M^3: $\forall p \in \mathcal{P},\ r \in \mathcal{R}^+$
$m^3_{pr} \leftarrow m^3_{pr} + \tau\left(\text{Proj}_{\mathbb{R}^+}\left(2m_{pr} - m^3_{pr} - \frac{1}{L}G_{pr}\right) - m_{pr}\right)$
  update $M \leftarrow \frac{1}{3}(M^1 + M^2 + M^3)$
end while

The positive parameters τ and L controls the gradient descent step. Experimentally, we set $\tau = \frac{1}{4}$ for $\kappa = 1$, $\tau = \frac{1}{2}$ for $\kappa = 10$, and $L = \frac{1}{10}\|A_\rho\|$, estimating $\|A_\rho\|$ using power-iteration. Note that other splitting schemes are possible and have been tested but this combination was particularly efficient (faster convergence) due to the simplicity of the projectors onto $\mathbb{P}_{\leq 1}$ and $\mathbb{P}_\kappa$ that can be computed in linear time [5] using the projection onto κ-simplex.

6. Experiments

6.1. Datasets

The Willow actions dataset [7] is a dataset for action classification, which contains 911 images split into 7 classes of common human actions, namely *interacting with a computer, photographing, playing music, riding cycle, riding horse, running, walking*. There are at least 108 images per actions, with around 60 images used as training and the rest as testing images. The dataset also offers bounding boxes, but we do not use them as we want to detect the relevant parts of images automatically.

The MIT 67 scenes dataset [26] is an indoor scene classification dataset, composed of 67 categories. These include stores (e.g. bakery, toy store), home (e.g. kitchen, bedroom), public spaces (e.g. library, subway), leisure (e.g. restaurant, concert hall), and work (e.g. hospital, TV studio). Scenes may be characterized by their global layout (corridor), or by the objects they contain (bookshop). Each category has around 80 images for training and 20 for testing.

6.2. Description and classification pipeline

We essentially follow the learning and classification setup of [30]. During part learning, $|\mathcal{R}| = 1,000$ regions are extracted from each training image and used to learn the parts. During encoding, $|\mathcal{R}|$ regions are extracted from both training and test images, and all images are encoded based on the learned parts. Finally, a linear SVM classifies the test images. For each stage, we briefly describe the choices made in [30] and discuss our improvements.

Extraction of image regions Two strategies are investigated:

- **Random regions** ('R'). As in [30], $|\mathcal{R}|$ regions are randomly sampled over the entire image. The position and scale of these regions are chosen uniformly at random, but regions are constrained to be square and have a size of at least 5% of the image size.
- **Region proposals** ('P'). Following [22], up to $|\mathcal{R}|$ regions are obtained based on selective search [39]. If less than $|\mathcal{R}|$ regions are found, random regions complete the set.

Region descriptors Again two strategies are investigated:

- **Fully connected** ('FC'). As in [30], we use the output of the 7th layer of the CNN of [13] on the rescaled regions, resulting in 4,096-dimensional vectors. For the Willow dataset, we use the standard Caffe CNN architecture [13] trained on ImageNet. For MIT67, we use the hybrid network [43] trained on ImageNet and on the Places dataset. The descriptors are square-rooted and ℓ_2-normalized. We note that each region was cropped and fed to the network in [30].
- **Convolutional** ('C'). As an improvement, we use the last convolutional layer, after ReLU and max pooling, of the very deep VGG-VD19 CNN [31] trained on ImageNet. To obtain a region descriptor, we employ average pooling over the region followed by ℓ_2-normalization, resulting in a 512-dimensional vector. Contrary to 'FC', we do not need to rescale every region. The entire image is fed to the network only once, as in [11, 10]. Further, following [37], pooling is carried out by an integral histogram. These two options enable orders of magnitude faster description extraction compared to 'FC'. To ensure the feature map is large enough to sample $|\mathcal{R}|$ regions despite loss of resolution (by a factor of 32 in the case of VD-19), images are initially resized such that their maximum dimension is 768 pixels. this was shown to be beneficial in [42].

Initialization The initialization step follows [30]. All training positive regions are clustered and for each cluster an LDA classifier is computed over all regions of the cluster. The maximum responses to the classifiers are then selected per image. Two scores are then computed: the average of the maximum responses over positive and negative sets. The ratio of these scores is used to select the top P clusters to build the initial part classifiers. Finally, an initial matching matrix M is built by softmax on classifier responses.

Encoding Given an image, bellonging to training or testing set, each learned part classifier is applied to every region descriptor to generate a global image descriptor. We use several alternatives:

1. **Bag-of-Parts** ('BoP') [30]: For each part, the maximum and average classifier scores are computed over all regions. These data are then concatenated for all parts.
2. **Spatial Bag-of-Parts** ('SBoP'): In this paper, we also introduce SBoP, which adds weak spatial information to BoP by using Spatial Pyramids as [8]: Maximum scores are computed over the four cells of a 2×2 grid over the image and appended to the original BoP.
3. **CNN-on-Parts** ('CoP') [30]: The CNN descriptors corresponding to the maximum scoring region per part are concatenated to form the global image descriptor.
4. **PCA on CNN-on-Parts** ('PCoP'): This paper also investigates PCoP, where centering and PCA are applied to CoP.

The global image descriptors will be the input of the final SVM classifier.

Parameters of the learning algorithms For the Iterative Soft-Assign (ISA) method, we use the same parameters as [30]. Concerning the $\text{GFB}^{\kappa}_{\rho}$ method solving (11), we tested different configurations: w/ or w/o regularization (as controlled by ρ) and one-to-one or one-to-many assignment (as controlled by κ).

When considering 1-to-1 matching with GFB^{1}_{ρ}, we perform 2k iterations of the projection, except for the MIT67 dataset with convolutional descriptor, where iterations are limited to 1k. In all experiments performance remains stable after 1k iterations. We denote by GFB^{1}_{0} the case without regularization ($\rho = 0$). For the GFB^{1}_{ρ}, we choose $\rho = 10^{-3}\|A\|$ after experimental evaluation on the Willow dataset.

When considering the multiple assignment model $\text{GFB}^{\kappa}_{\rho}$ with $\kappa = K = 10$, we used $\rho = -\|A\|$. The motivation

Table 1. Baseline performance, without part learning.

Method	Measure	Willow		MIT67	
		FC	C	FC	C
Full-image	Acc	–	–	70.8	73.3
	mAP	76.3	88.5	72.6	75.7

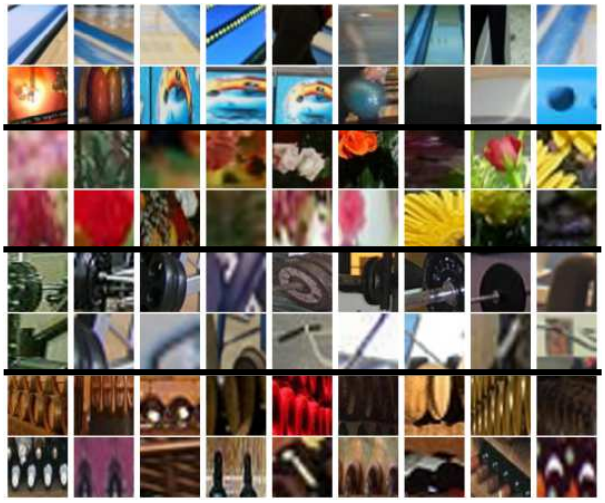

Figure 2. Top scoring parts for various images of bowling, florists, gym and wine cellar.

behind this choice is that, because the simplex $\mathcal{S}_\kappa$ is larger as κ increases, the optimal matrix is more likely to lie *inside* the simplex, with values between 0 and 1 (soft assignment). This effect can be compensated by using a negative value for ρ, which yields a hard assignment solution in practice in our experiments.

6.3. Results

In the following, we are showing results for (i) fully connected layer descriptor on random regions (R+FC), which follows [30], and (ii) convolutional layer descriptor on region proposals (P+C) that often yields the best performance. We evaluate different learning algorithms on BoP and CoP encoding, and then investigate the new encoding strategies SBoP and PCoP as well as combinations for ISA, Hun, and GFB^κ_ρ algorithms. Methods are evaluated in the context of action and scene classification in still images. On Willow we always measure mean Average Precision (mAP) while on MIT67 we calculate both mAP and classification accuracy (Acc).

We start by providing, in Table 1, a baseline corresponding to the description methods 'FC' and 'C' applied on the full image without any part learning. The comparison to subsequent results with part learning reveals that part-based methods always provides improvement.

We now focus on the part learning methods. Figure 2 shows some qualitative results of learned parts on MIT67.

Table 2. Performance of ISA and all methods satisfying $\mathcal{M}_1$ on Willow and MIT67.

Method	Meas.	ISA	IPFP	Hun	GFB^1_0	GFB^1_ρ
Willow						
R+FC BoP	mAP	76.6	79.0	78.9	79.7	**80.6**
P+C BoP	mAP	**89.2**	86.3	88.3	88.2	87.5
P+C CoP	mAP	91.6	91.3	91.1	**91.8**	**91.8**
MIT 67						
R+FC BoP	Acc	**76.6**	–	75.4	75.7	74.7
	mAP	**78.8**	–	78.0	77.6	76.3
P+C BoP	Acc	**75.1**	70.7	72.8	70.9	70.9
	mAP	**76.7**	72.6	75.1	73.5	73.1
P+C CoP	Acc	**80.0**	79.2	79.8	79.2	79.3
	mAP	**80.2**	79.7	79.9	79.5	79.7

Table 3. Performance ISA, when forced to satisfy the constraints $\mathcal{M}_1$ with hard assignment. ISA+H refers to performing one iteration of the Hungarian algorithm on the solution obtained by ISA.

Method	Measure	ISA	ISA+H
Willow R+FC BoP	mAP	76.6	**76.9**
Willow P+C BoP	mAP	**89.2**	88.1
Willow P+C CoP	mAP	**91.6**	89.6
MIT67 R+FC BoP	mAP	**78.8**	77.9

Then, Table 2 shows the performance of ISA against several methods satisfying the constraint $\mathcal{M}_1$, see Eq (1), i.e. a part is composed of a single region in every positive image. These methods include IPFP, Hungarian, GFB^1_0, and GFB^1_ρ.

On the Willow dataset, for the R+FC descriptor with BoP encoding, we observe that $\text{GFB}^1_\rho > \text{GFB}^1_0 >$ Hungarian and IPFP $>$ ISA. However, on MIT67 the results are different and we have ISA $>$ Hungarian and $\text{GFB}^1_0 > \text{GFB}^1_\rho$. Similar trends are observed when using the improved P+C descriptor with the BoP encoding. Nevertheless, note that all methods perform similarly when using the CoP encoding. After these evaluations, IPFP was not evaluated in further experiments since it performs on par with the Hungarian or worst, as explained in Section 5.1.

These results show that overall ISA outperforms other optimization methods, which satisfy $\mathcal{M}_1$. The explanation of this difference in performance lies in the fact that ISA is stopped before convergence and does not satisfy $\mathcal{M}_1$, as explained in Section 5.2. This result is further confirmed by running an iteration of the Hungarian algorithm on the output of ISA, see Table 3. Such experiment forces the parts resulting from ISA to satisfy $\mathcal{M}_1$ and we observe an overall drop of performance.

Table 2 also shows that region proposals combined with convolutional layer descriptions shows a significant performance gain, especially on the Willow dataset. Therefore, the improved region descriptions and encoding are evaluated using ISA, see Table 4. We can see a consistent improvement for the SBoP and PCoP encoding. Also PCA yields more improvement for descriptors based on fully

Table 4. Results on Willow and MIT67 datasets for the ISA method, with improved region descriptions P+C and improved encoding methods SBoP and PCoP.

Method	Meas.	BoP	SBoP	CoP	PCoP
Willow R+FC	mAP	76.6	78.7	81.6	**82.4**
Willow P+C		89.2	90.1	91.6	**91.7**
MIT67 R+FC	Acc	76.6	76.1	76.8	**77.1**
	mAP	78.8	79.0	77.8	**79.5**
MIT67 P+C	Acc	75.1	76.1	80.0	**80.5**
	mAP	76.7	76.7	80.2	**81.0**

Table 5. Performance of ISA and the proposed methods satisfying the constraint $\mathcal{M}_\kappa$. κ is set to 10 for Hun$^\kappa$ and GFB$^\kappa_\rho$.

Method	Meas.	ISA	Hun$^\kappa$	GFB$^\kappa_\rho$
Willow				
P+C BoP	mAP	89.2	**89.6**	**89.6**
P+C CoP		**91.6**	91.4	91.3
P+C SBOP+PCoP		91.9	**92.1**	**92.1**
MIT 67				
P+C BoP	Acc	75.1	**77.7**	77.3
	mAP	76.7	79.2	**79.4**
P+C CoP	Acc	80.0	80.4	**80.5**
	mAP	80.2	**80.6**	80.5
P+C SBOP+PCoP	Acc	81.4	**81.5**	**81.5**
	mAP	81.2	81.7	**81.9**

connected layers than on convolutional ones.

We further study the problem of part learning following the constraint $\mathcal{M}_\kappa$, see Eq. (4). Therefore, a part can be composed of several regions of the same image. We remind that each of these regions can be assigned to at most one part. We compare ISA to GFB$^\kappa_\rho$ and Hun$^\kappa$, which satisfy $\mathcal{M}_\kappa$. In our setup, we have $1 \leq \kappa \leq K = 10$. Concerning Hun$^\kappa$, κ is set to K. Results given on Table 5 show that GFB$^\kappa_\rho$ and Hun$^\kappa$ offer better results than ISA, especially with the BoP encoding. Moreover, there is a large improvement over the same methods satisfying $\mathcal{M}_1$ and we note that GFB adapts to the various types of constraints.

Since the Hungarian algorithm forces parts to be composed of a fixed number of regions κ, we evaluated the impact of this parameter on the classification performance on MIT 67 dataset, see Table 6. Interestingly, high values of κ offer the best performance. Therefore, a mixture of parts combining all regions of images allows a better description. An explanation for such results can be that parts will be more diverse and therefore more distinct one to another.

Finally, our methods offer good performance competing with the state of the art on both datasets: 92.1% mAP on Willow and 81.5% accuracy on MIT67, see Table 5 and 7.

7. Conclusion

To conclude, we have investigated in this work the problem of discovering parts for part-based image classification. We have shown that this problem can be recast as a quadratic assignment problem with concave objective function to be minimized with non-convex constraints. While being known to be a very difficult problem, several techniques have been proposed in the literature, either trying to find "hard assignment" in a greedy fashion, or based on optimization of the relaxed problem, resulting in "soft assignment". Several methods have been investigated to address this task and compared to the previous method of [30]. Of the proposed algorithms, GFB is the most adaptable and the Hungarian is the fastest. Both algorithms offer improved performance on two public datasets.

Our reformulation and investigation of different optimization methods explore the limits of the original problem defined in [30]. We introduce a number of new algorithms, which are designed to satisfy the constraint of the problem definition. We show that the explicit relaxation of the constraint on the assignment of regions to parts leads to a better part model. Such adaptation was not possible in the work of [30]. We believe this knowledge will help the community in the search for more appropriate models, potentially end-to-end trainable, using better network architectures.

We additionally proposed improvements on several stages of the classification pipeline, namely region extraction, region description and image encoding, using a very deep CNN architecture. Furthermore, the new region description method is orders of magnitude faster, as this process was previously the bottleneck in [30].

Table 6. Evaluation of the influence the parameter κ on the performance of the Hungarian algorithm on MIT67.

Method	Meas.	Hun$^\kappa$				
	κ	1	2	5	8	10
P+C BoP	Acc	72.8	74.9	74.9	76.9	77.7
	mAP	75.1	76.2	78.0	79.0	79.2

Table 7. Performance in terms of accuracy of existing part-based and non part-based methods on the MIT67 dataset.

Methods	Part-based	**MIT67**
Zhou *et al* [43]	No	70.8
Peng *et al* [35]	Yes	74.9
Wang *et al* [40]	Yes	75.3
Mahmood *et al* [20]	No	75.6
Zuo *et al* [44]	Yes	76.2
Parizi *et al* [24]	Yes	77.1
Mettes *et al* [22]	Yes	77.4
Sicre *et al* [30]	Yes	78.1
Zheng *et al* [42]	No	78.4
Wu *et al* [41]	Yes	78.9
Cimpoi *et al* [4]	No	81.0
Herranz *et al* [12]	No	**86.0**
Ours	Yes	**81.5**

References

[1] S. Bougleux and L. Brun. Linear sum assignment with edition. *CoRR*, abs/1603.04380, 2016. 4

[2] Y.-L. L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1

[3] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, 2012. 2, 3, 4

[4] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision*, pages 1–30, 2015. 8

[5] L. Condat. Fast Projection onto the Simplex and the l1 Ball. to appear in Mathematical Programming Series A, Aug. 2015. 5

[6] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Intl. Workshop on Stat. Learning in Comp. Vision*, 2004. 1

[7] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference*, volume 2, 2010. 5

[8] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *Advances in Neural Information Processing Systems*, pages 494–502, 2013. 1, 2, 6

[9] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes Paris look like Paris? *ACM Trans. Graph.*, 31(4), 2012. 1

[10] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 6

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, 2014. 6

[12] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: objects, scales and dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016. 8

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014. 6

[14] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 1

[17] M. Leordeanu, M. Hebert , and R. Sukthankar. An integer projected fixed point method for graph matching and map inference. In *Proceedings Neural Information Processing Systems*. Springer, December 2009. 4

[18] J. J. Lim, C. L. Zitnick, and P. Dollár. Sketch tokens: A learned mid-level representation for contour and object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3158–3165. IEEE, 2013. 1

[19] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *2011 International Conference on Computer Vision*, pages 2486–2493, Nov 2011. 4

[20] A. Mahmood, M. Bennamoun, S. An, and F. A. Sohel. Resfeats: Residual network based features for image classification. *CoRR*, abs/1611.06656, 2016. 8

[21] S. Maji and G. Shakhnarovich. Part discovery from partial correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 931–938, 2013. 2

[22] P. Mettes, J. C. van Gemert, and C. G. M. Snoek. No spare parts: Sharing part detectors for image categorization. *CoRR*, abs/1510.04908, 2015. 2, 6, 8

[23] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable models for scene recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2775–2782. IEEE, 2012. 2

[24] S. N. Parizi, A. Vedaldi, A. Zisserman, and P. Felzenszwalb. Automatic discovery and optimization of parts for image classification. In *International Conference on Learning Representations*, 5 2015. 1, 2, 8

[25] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, 2010. 1

[26] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[27] H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIIMS*, 6(3):1199–1226, 2013. 5

[28] A. Rangarajan, A. Yuille, and E. Mjolsness. Convergence properties of the softassign quadratic assignment algorithm. *Neural Comput.*, 11(6):1455–1474, Aug. 1999. 5

[29] R. Sicre, Y. Avrithis, E. Kijak, and F. Jurie. Unsupervised part learning for visual recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[30] R. Sicre and F. Jurie. Discriminative part model for visual recognition. *Computer Vision and Image Understanding*, 141:28 – 37, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 6

[32] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proceedings of the European Conference on Computer Vision*, pages 73–86. Springer, 2012. 2

[33] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein

distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, July 2015. 3

[34] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 2

[35] P. Tang, J. Zhang, X. Wang, B. Feng, F. Roli, and W. Liu. Learning extremely shared middle-level image representation for scene classification. *Knowledge and Information Systems*, pages 1–22, 2016. 8

[36] H. E. Tasli, R. Sicre, and T. Gevers. Superpixel based mid-level image description for image recognition. *Journal of Visual Communication and Image Representation*, 33:301–308, 2015. 1

[37] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. 2016. 6

[38] S. Ullman, E. Sali, and M. Vidal-Naquet. A Fragment-Based Approach to Object Representation and Classification. In *Visual Form 2001*, pages 85–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. 1

[39] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *IEEE International Conference on Computer Vision*, 2011. 6

[40] X. Wang, L. Lu, H.-c. Shin, L. Kim, M. Bagheri, I. Nogues, J. Yao, and R. M. Summers. Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. *arXiv preprint arXiv:1701.06599*, 2017. 8

[41] R. Wu, B. Wang, W. Wang, and Y. Yu. Harvesting discriminative meta objects with deep cnn features for scene classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1287–1295, 2015. 8

[42] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian. Good practice in cnn feature transfer. *arXiv preprint arXiv:1604.00133*, 2016. 6, 8

[43] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems*, 2014. 6, 8

[44] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang. Learning discriminative and shareable features for scene classification. In *Computer Vision–ECCV 2014*, pages 552–568. Springer, 2014. 8