Automatic discovery of discriminative parts as a quadratic assignment problem

Ronan Sicre[1], Julien Rabin[2], Yannis Avrithis[1], Teddy Furon[1], Frederic Jurie[2]

[1] *INRIA - Campus Beaulieu, Rennes, France*
[2] *GREYC UMR 6072 - University of Caen, 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. The training of parts is cast 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*. State-of-the art results are obtained on these datasets.

Key words: Image classification, part-based models, parts discovery.

1. Introduction

The representation of images as set of patches has a long history in computer vision, especially for object recognition [1], image classification [2] or object detection [3]. 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 recognition 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 take a decision e.g. to assign a category to an image. As an illustration, [4] proposed to select informative patches using an entropy based criterion while the decision relies on a naive-Bayes classifier. Following [4], recent approaches separate the construction of the model (i.e. the learning of the parts) and the decision function [5] [6]. The reason behind this choice is that the number of candidate regions in the training images is very large and would lead to a highly non-convex decision function.

Optimizing both (parts and decision) is however possible for simple enough part detectors and decision functions. For instance, [7] unifies the two stages by jointly learning the image classifiers and a set of shared parts. Their claim is that the definition of the parts is directly related to the final classification function.

While this argument is true, the objective function of this joint optimization is highly non-convex with no guaranty of convergence. We believe that deciding which one of the two alternatives – the joint optimization vs separate one – is still an open problem. As an insight, the two stage part-based model of [8] performs better than the joint learning of [7]. We note that there are other differences between the two approaches, e.g. [7] models both positive and negative parts while [8] focuses only on the positive ones.

Interestingly, [8] addresses the learning of parts as an assignment problem. On one hand, regions are sampled randomly from the training images. On the other hand, the model is considered as a set of parts. The assignment is constrained by imposing that each part should be assigned to one image region in each positive image (those belonging to the category to be modeled). This results in a bipartite graph linking parts and regions.

The assignment problem of [8] poses the learning of part-based models in a very appealing way, yet their solution is based on heuristics leaving room for improvements. This paper's contribution is an extensive study of this assignment problem: We first present of a well-founded formulation of the problem and propose different solutions in a rigorous way. These different methods are evaluated and compared on two different datasets and state-of-the-art performance is obtained. By experimenting with improvements in the underlying description and encoding, we demonstrate that the benefit of our part learning methodology remains complementary to the benefit of more powerful visual representations obtained by state of the art deep learning approaches.

2. Previous work

Image classification has received a lot of attention during the last decades, with most of the related approaches focused on models based on aggregated features [9, 10] or the Spatial Pyramid Matching [11]. This was before the Convolutional Network revolution [12] still at the heart of most of the recent methods [13].

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 all the image regions are pooled without selecting sparse discriminative parts. For instance, [14] 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), during an iterative procedure which alternates between clustering and training classifiers. Similarly, [5] addresses this problem by learning parts incrementally, starting from a single part occurrence with an Exemplar SVM and collecting more occurrences from the training images.

In a different way, [6] poses the discovery of visual elements as a discriminative mode seeking problem solved with the mean-shift algorithm: it discovers visually-coherent patch clusters that are maximally discriminative. In [15], Maji *et al* investigate the problem of parts discovery when some correspondences between instances of a category are known. The work of [16] 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 on the number of regions a part is assigned to (from zero to multiple); given this fixed assignment, part detectors are optimized using stochastic gradient descent.

The recent papers related to part-based models are those of Sicre *et al* [8] and Parizi *et al* [7]. As said before, the part-based representation of [7] 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* [8] follow the two stage formulation, formulating the discovery of parts as an assignment problem. We also mention the recent and unpublished work of Mettes *et al* [17] arguing that image categories may share parts and proposing a method to model them as such.

Finally, this paper is related to the assignment problem which finds a maximum weight matching in a weighted bipartite graph. A survey on this topic is the work of Burkard *et al* [18].

3. Discovering and Learning Parts

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

This section presents different ways to formalize this task giving birth to interesting optimization alternatives in Sect. 4. We first present the parts learning problem as defined in [8]. We show that it boils down to a concave minimization under non convex constraints, which is recast as a quadratic assignment problem.

3.1. Notation

$X^\top$ and $\mathrm{tr}(X)$ are the transpose and trace of matrix X; $\mathrm{vec}(X)$ is the column vector containing all elements of 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. The Euclidean norm of vector x is $\|x\| = \sqrt{\langle x, x \rangle}$. Vector $x_{i\bullet}^\top$ ($x_{\bullet j}$) denotes the i-th row (resp. j-th column) of matrix X. The $n \times n$ identity matrix is denoted as I_n, while vector $\mathbf{1}_n$ (matrix $\mathbf{1}_{m \times n}$) is an $n \times 1$ vector (resp. $m \times n$ matrix) of ones. $\mathbb{1}_{\mathcal{A}}$ is the indicator function of set $\mathcal{A}$ and $\mathrm{Proj}_{\mathcal{A}}$ is the Euclidean projector onto $\mathcal{A}$.

Following [8], we denote by $\mathcal{I}^+$ with $n^+ = |\mathcal{I}^+|$ the set of images of the category 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 $n = |\mathcal{I}|$ 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, and in particular it is the output of a convolutional or fully connected layer. More details are given in Section 5.2. 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 category is modeled by a set of parts $\mathcal{P}$ with $|\mathcal{P}| = P$. We introduce the $P \times R^+$ matching matrix M associating image 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 represents part p, and 0 otherwise. By M_I we denote the $P \times |\mathcal{R}|$ submatrix of M that contains columns $r \in \mathcal{R}_I$ corresponding to image I.

We keep the requirements of [8]: (i) the P parts are different from one another, (ii) each part is present in every positive image, (iii) parts should occur more frequently in positive images than in negative ones. The first two requirements define constraints on the admissible set $\mathcal{A}$ of M:

$$\mathcal{A} \triangleq \left\{ M \in \{0,1\}^{P \times R^+} : M^\top \mathbf{1}_P \leq \mathbf{1}_{R^+} \text{ and } M_I \mathbf{1}_{|\mathcal{R}|} = \mathbf{1}_P \text{ for } I \in \mathcal{I} \right\}. \quad (1)$$

This implies that each sub-matrix M_I is a *partial assignment* matrix. Observe that the set $\mathcal{A}$ is not convex. The third assumption is enforced by Linear Discriminant Analysis (LDA): given matching matrix M, the model $w_p(M)$ of part p is defined as

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

where $\mu = \frac{1}{n} X \mathbf{1}_R$ and $\Sigma = \frac{1}{n}(X - \mu \mathbf{1}_R^\top)(X - \mu \mathbf{1}_R^\top)^\top$ 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 the inner product $\langle w_p(M), x_r \rangle$.

For a given category, we are looking for an optimal matching matrix

$$M^\star \in \arg\max_{M \in \mathcal{A}} J(M) \tag{3}$$

$$J(M) \triangleq \sum_{p \in \mathcal{P}} \sum_{r \in \mathcal{R}^+} m_{pr} \langle w_p(M), x_r \rangle = \left\langle M, W(M)^\top X^+ \right\rangle, \tag{4}$$

where $W(M)$ is the $d \times P$ matrix whose columns are $w_p(M)$ for all parts $p \in \mathcal{P}$.

3.3. Recasting as a quadratic assignment problem

The previous formulation limits optimization to alternatively resorting to (2) and (4), as done in [8]. Here, we express J as a function of M without W, recasting (3) as a quadratic assignment problem and opening the way to a number of alternative optimization algorithms. We define similarity matrix $C(M) \triangleq W(M)^\top X^+$. Its entries represent the similarities between parts and regions. 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

$$C(M) = \left(\frac{1}{n^+} M {X^+}^\top - \mathbf{1}_P \mu^\top\right) \Sigma^{-1} X^+ = MA - B, \tag{5}$$

where $n^+ \times n^+$ matrix $A = \frac{1}{n^+} {X^+}^\top \Sigma^{-1} X^+$ is symmetric and positive definite and $P \times n^+$ matrix $B = \mathbf{1}_P \mu^\top \Sigma^{-1} X^+$ has identical rows (rank 1). Now, observe that problem (3) is equivalent to

$$\text{find} \quad M^\star \in \arg\min_{M \in \mathcal{A}} \ J_0(M) \tag{6}$$

$$J_0(M) \triangleq \langle M, B - MA \rangle = \operatorname{vec}(M)^\top Q \operatorname{vec}(M) + \operatorname{vec}(B)^\top \operatorname{vec}(M) \tag{7}$$

for a $PR^+ \times PR^+$ matrix Q that is a function of A only. This shows that our task is closely related to the *quadratic assignment problem* [18], a NP-hard combinatorial problem. Moreover, in our setting, the objective function to be minimized is strictly concave.

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

3.4. Convex relaxation with entropic regularization

In the specific case of **fixed cost matrix** C, the previous problem becomes tractable.

3.4.1. Convex relaxation

Solving a linear assignment problem is numerically demanding, with a complexity about $O(N^3)$ [18]. It can be done exactly with dedicated methods, such as the Hungarian algorithm; or equivalently, with linear programming methods that assume convex relaxation of the binary constraints, i.e. considering bi-stochastic instead of permutation matrices.

3.4.2. Soft assignment

To reduce the complexity, the problem is approximated using negative-entropy regularization. Considering a $P \times P$ bi-stochastic matrix M, the soft-assignment problem is

$$\underset{\substack{M \in \mathcal{R}_+^{P\times P},\\ M\mathbf{1}_P = \mathbf{1}_P, M^\top \mathbf{1}_P = \mathbf{1}_P}}{\operatorname{argmin}} \left\{ \langle C, M\rangle - \frac{1}{\beta} E(M) = \langle C, M\rangle + \frac{1}{\beta}\langle \log(M), M\rangle \right\} \qquad (8)$$

where E is the entropy of the bistochastic matrix M, and $\beta > 0$ is the regularization parameter. As β increases, the problem converges to the hard-assignment problem. Paper [8] uses the Sinkhorn algorithm [19], which normalizes iteratively the rows and the columns of M to one, initializing from the regularized cost matrix $\exp(\beta C)$.

Observe that in our setting M is not square as we consider partial assignments between P rows and $R > P$ columns. To solve this more general problem, a simple trick is to add as many rows than required and to define a maximal cost value when affecting columns to them.

Soft assignment has gained a lot attention because it solves large scale problems [20]. However, a major limitation is the loss of sparsity of the solution. As a consequence, approximate solutions of the linear soft-assignment are not

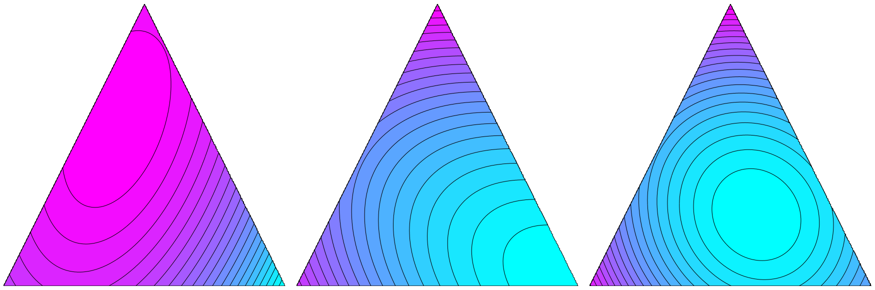

Figure 1: **Illustration of the convex relaxation of our assignment problem in 3D.** *Black lines are level-sets of the objective function J_0 in the plane of the simplex, which is a triangle in $\mathbb{R}^3$. Lower values are displayed in cyan, larger in magenta. (Left) The original problem is the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

suitable for our problem, as observed in experiments. We describe in the section 4.2 how the authors of [8] have circumvented this problem by iterating soft assignment.

3.5. Convex relaxation with quadratic regularization

We consider now the quadratic regularization of the problem (see Figure 1):

$$\begin{aligned} J_\rho(M) &\triangleq \langle M, M(\rho I_{n^+} - A) + B\rangle = \langle M, B - MA\rangle + \rho\|M\|_F^2 \\ &= J_0(M) + \rho\, P\, n^+ \text{ for } M \in \mathcal{A}. \end{aligned} \tag{9}$$

This means that provided $M \in \mathcal{A}$, $J_\rho(M)$ and $J_0(M)$ differ by a constant. Therefore, the minimizers of J_ρ on $\mathcal{A}$ are the minimizers of J_0, for any value of ρ. Indeed, if ρ is sufficiently large ($\rho > \|A\|$), J_ρ becomes convex (see Figure 1).

We also relax the constraints:

$$M^\star = \arg\min_{M\in\mathcal{S}} \; J_\rho(M) \tag{10a}$$

$$\mathcal{S} \triangleq \left\{ M \in [0,1]^{P\times R^+} : \; M^\top \mathbf{1}_P \le \mathbf{1}_{R^+}, \text{ and } M_I \mathbf{1}_{|\mathcal{R}|} = \mathbf{1}_P \text{ for } I \in \mathcal{I}^+ \right\}. \tag{10b}$$

In words, domain $\mathcal{S}$ is the convex hull of the set $\mathcal{A}$ and we will refer it to as a simplex. Yet, in general, $J_\rho(M) \neq J_0(M) + \rho P n^+$ when $M \in \mathcal{S} \setminus \mathcal{A}$. We may find different solutions than problem (6), as illustrated in Figure 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 $\frac{1}{2}BA_\rho^{-1}$ which may lie inside $\mathcal{S}$. This shows that ρ implicitly controls the number of regions used to describe parts.

4. Optimization

The previous section formalizes the part learning task into optimization problems. This section now presents various methods to numerically solve them. We see two kinds of techniques: i) hard assignment optimization directly finding $M^\star \in \mathcal{A}$, ii) soft assignment optimization (Sect. 3.4 and 3.5) that finds $M^\star \in \mathcal{S}$.

This latter strategy is not solving the initial problem. However, as already observed in [8] and [21] for classification, soft-assignment affecting several regions to describe a part, may provide better results. This lesson learned from previous works deserves an experimental investigation in our context.

4.1. Hard assignment methods

4.1.1. Hungarian Algorithm

As mentioned in Sect. 3.4, problem (3), when the cost matrix $C(M)$ is fixed, is a variant of the *linear assignment problem* for which several dedicated methods give an exact solution. Solving this approximated problem can be seen as computing the orthogonal projection of the matrix $C(M_0)$ (M_0 being an initial guess, see Section 5.2.1) onto the set $\mathcal{A}$

$$M^\star_{hun} = \text{Proj}_{\mathcal{A}}\left(C(M_0)\right) = \underset{M \in \mathcal{A}}{\text{argmax}}\ \langle M, C(M_0) \rangle. \qquad (11)$$

In our setting, we use the fast Hungarian algorithm variant of [22]. The experimental section shows that this gives surprisingly good results in comparison to more sophisticated methods.

4.1.2. IPFP

The Integer Projected Fixed Point (IPFP) method [23] can be seen as the iteration of the previous method, alternating between similarity matrix $C(M)$ update and projections onto the constraints set $\mathcal{A}$. More precisely, a first order Taylor approximation of the objective function is maximized (e.g. using the Hungarian algorithm) and combined with a linesearch (see Algorithm 1). This approach guaranties the convergence to a local minimizer of $J(M)$ on the set $\mathcal{A}$.

Algorithm 1 IPFP algorithm for problem (6)

```
Init: $M_0$, set: $k \leftarrow 0$, $M_{-1} \leftarrow \mathbf{0}$
while $M_{k+1} \neq M_k$ do
    $k \leftarrow k + 1$
    $G_k \leftarrow 2M_k A - B$    (gradient $\nabla J(M_k)$)
    $P_{k+1} \leftarrow \mathrm{Proj}_{\mathcal{A}}(G_k)$    (projection using partial Hungarian algorithm [22])

    $\Delta_{k+1} \leftarrow P_{k+1} - M_k$
    $c_k \leftarrow \langle G_k, \Delta_{k+1} \rangle$
    $d_k \leftarrow \langle \Delta_{k+1} A, \Delta_{k+1} \rangle$
    $t_k = \min(-\frac{c_k}{2d_k}, 1)$ 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 without improving much results. This is explained by the very specific structure of our problem, where the quadratic matrix Q of (7) is very sparse and negative definite.

4.2. Iterative Soft-assignment (ISA)

The strategy of [8] referred here 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. 3.4) and learns the parts by yielding $W(M)$ for a given M thanks to LDA. The assignation resorted to a soft-assign algorithm (see [24] for instance) which is also an iterative algorithm solving a sequence of entropic-regularized problems (Section 3.4) that converges to the target one. The general scheme of the algorithm is drawn in Algorithm 2.

Algorithm 2 Iterated-SoftAssign algorithm

```
Init: $M = M_0$
while $M \notin \mathcal{A}$ 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 linear Soft-Assignment problem (8)
    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). Nevertheless, as reported in [8], 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 $\mathcal{A}$ nor $\mathcal{S}$).

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

To address the relaxed problem (10b), we split the constraints on the matching matrix M for rows and columns: for each row $m_{p\bullet}$ and each column $m_{\bullet r}$ of M

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

Problem (10b) is then equivalent to the following

$$\underset{M=M_1=M_2\,\in\,\mathbb{R}^{P\times R^+}}{\operatorname{argmin}} \quad J_\rho(M)+G_1(M_1)+G_2(M_2) \tag{12}$$

where G_1 and G_2 respectively encode constraints on parts and regions:

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

The General Forward Backward (GFB) algorithm [25] 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. The positive parameter L controls the gradient descent step. Experimentally, we set $L = \frac{1}{10}\|A\|$ and estimate $\|A\|$ using power-iteration. The projector onto $\mathbb{P}$ is computed in linear time [26]. The projection onto $\mathbb{P}_\leq$ is trivial. Note that other splitting schemes are possible and have been tested (for instance, using non-negativity constraint on a third variable), but this combination was particularly efficient (faster convergence). The main advantage of this algorithm is that it can be massively parallelized.

Algorithm 3 GFB algorithm for problem (12)

```
$M \leftarrow M_0$      (initialization)
while not converge do
    $\nabla J_\rho(M) = 2MA_\rho + B$      (gradient)
    update $M_1$: $m^1_{p\bullet} \leftarrow m^1_{p\bullet} - m_{p\bullet} + \mathrm{Proj}_{\mathbb{P}}\left(2m_{p\bullet} - m^1_{p\bullet} - \frac{1}{L}\nabla J_\rho(M)_{p\bullet}\right)$ $\forall\, p \in \mathcal{P}$
    update $M_2$: $m^2_{\bullet r} \leftarrow m^2_{\bullet r} - m_{\bullet r} + \mathrm{Proj}_{\mathbb{P}_\leq}\left(2m_{\bullet r} - m^2_{\bullet r} - \frac{1}{L}\nabla J_\rho(M)_{\bullet r}\right)$ $\forall\, r \in \mathcal{R}^+$
    update $M \leftarrow \frac{1}{2}(M_1 + M_2)$
end while
```

5. Experiments

5.1. Datasets

5.1.1. The Willow actions dataset [27]

is a dataset for action classification, which contains 911 images split into 7 classes of common human actions, namely *interacting with a computer, pho-*

tographing, 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.

5.1.2. The MIT 67 scenes dataset [28]

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.

5.2. Improved description and classification pipeline

We follow the general learning and classification pipeline of [8], however we also introduce significant improvements. Such improvements makes sense in order to compete with recent works. In summary, 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 is used to classify test images. For each stage, we briefly describe the choices made in [8] and discuss our improvements.

5.2.1. Initialization

The initialization step is achieved as in [8]. All training positive regions are clustered and for each cluster an LDA classifier is computed over all regions of the cluster. Maximum responses to the classifiers are then selected per image and averaged over positive and negative sets to obtain two scores. 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. This scheme is followed for all optimization algorithms, even if a part model matrix is not explicitly formed during iterations.

5.2.2. Extraction of image regions

Two strategies are investigated:

- *Random regions* ('R'). As in [8], $|\mathcal{R}|$ 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 [17], up to $|\mathcal{R}|$ regions are obtained based on selective search [29]. If less than P regions are found, randomly sampled regions are added to complete the set.

5.2.3. Region descriptors

Again two strategies are investigated, based on fully connected CNN or convolutional layers:

- *Fully connected* ('FC'). As in [8], we use the output of the 7th layer of the CNN proposed by [30] on the rescaled region, resulting in a 4,096-dimensional vector. For the Willow dataset, we use the standard Caffe CNN architecture [30] trained on ImageNet. For MIT67, we use the hybrid network [31] trained on ImageNet and on the Places dataset. The descriptors are square-rooted and ℓ_2-normalized.

- *Convolutional* ('C'). As an improvement, we use the last convolutional layer, after ReLU and max pooling, of the very deep VGG-VD19 CNN [13] 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 and feed it to the network; rather, the entire image is fed to the network only once, as in [32, 33]. Further, following [34], 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 has been shown to be beneficial [35].

5.2.4. Encoding

Given an image, either training or testing, region descriptors are tested against the learned part model to generate a global image descriptor, which is then used by a SVM classifier. We use several alternative strategies:

- *Bag-of-Parts* ('BoP') and *Spatial Bag-of-Parts* ('SBoP'). According to BoP [8], for each part classifier, the maximum and average score is computed over all regions; the scores for all parts are then concatenated. Here we introduce SBoP, which adds weak spatial information to BoP by using Spatial Pyramids as [6]. In this case, maximum scores are computed over the four cells of a 2×2 grid over the image and appended to the original BoP.

- *CNN-on-Parts* ('CoP') and *PCA on CNN-on-Parts* ('PCoP'). According to CoP [8], the CNN descriptors corresponding to the maximum scoring region per part are concatenated to form the image descriptor. Here we also investigate PCoP, whereby centering and PCA is applied to CoP as in [36].

5.2.5. Parameters of the learning algorithms

For the Iterative Soft-Assign (ISA) method, we use the same parameters as [8]. Concerning the GFB method, 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. For the GFB method with $\rho \neq 0$, reffed to as GFB$_\rho$, we choose $\rho = 10^{-3}\|A\|$ after experimental evaluation on the Willow dataset. We denote by just GFB the case where $\rho = 0$.

5.3. Results

In the following, we are showing results for (i) fully connected layer descriptor on random regions (R+FC), which follows [8], and (ii) convolutional layer descriptor on region proposals (P+C), which 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 the ISA algorithm. 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 our description methods on the full image without any part learning. Comparing to subsequent results with part learning reveals that part-based methods always provide better description of the content of an image.

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

We now focus on the part learning methods, which are evaluated in the context of action and scene classification in still images. Figure 5.3 shows some qualitative results of learned parts on MIT67. Then, Table 2 shows the performance of ISA, IPFP, Hungarian, GFB, and GFB_ρ on Willow and MIT67 datasets. After some evaluation on both MIT67 and Willow, IPFP was not evaluated in further experiments since it performs on par with the Hungarian or worst, as previously explained in 4.1.2. On the Willow dataset, we observe that $\mathrm{GFB}_\rho >$ GFB $>$ Hungarian and IPFP $>$ ISA. However, on MIT67 the results are different and we have ISA $>$ Hungarian and GFB $> \mathrm{GFB}_\rho$. When using the improved P+C descriptor, we observe a similar trend for the BoP. Nevertheless, note that all methods perform similarly when using the CoP encoding.

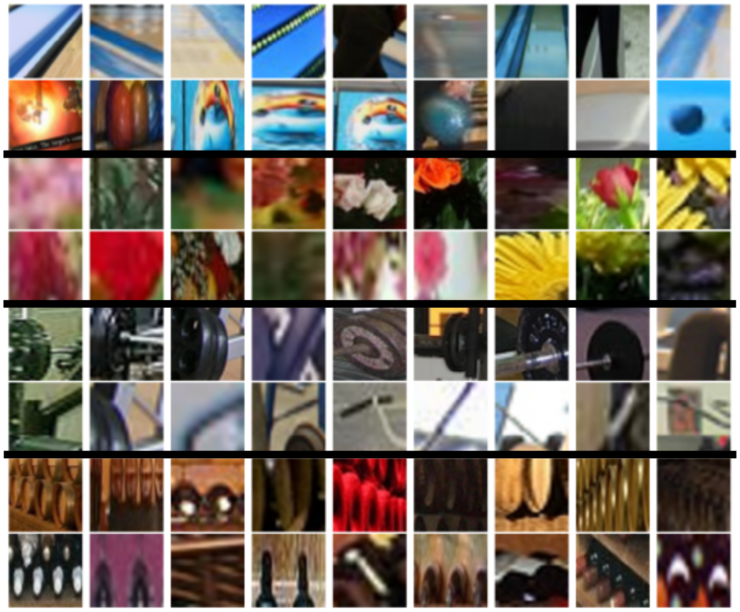

Figure 2: Top scoring parts for various images of bowling, florists, gym and wine cellar

Based on this experimentation, we can draw two main conclusions. First, methods based on soft assignment (ISA, GFB) clearly outperform methods based on hard assignment. This result is also confirmed, in almost all cases, by the results of Table 3, where one iteration of the Hungarian algorithm is performed on the assignment matrix obtained after ISA (i.e. ISA+H).

Second, while the GFB offers some significant practical advantage, when combined with quadratic regularization it is out-performed by Iterative Soft-Assign (except on the Willow dataset with BoP and CoP, first and third line in Table 2). Our explanation is that it demonstrates that quadratic regularization is less appropriate than entropic regularization for this problem. Indeed, as illustrated in Section 3.5, over-relaxing the objective function J_ρ tends to yield a matrix with very similar rows, meaning that parts are described by the same regions, which is highly undesirable. While this problem also occurs when solv-

Table 2: Performance of various learning methods on Willow and MIT67.

Method	Measure	ISA	IPFP	Hun	GFB	GFB$_\rho$
Willow R+FC BoP	mAP	76.6	79.0	78.9	79.7	**80.6**
Willow P+C BoP		**89.2**	86.3	88.3	88.2	87.5
Willow P+C CoP		91.6	91.3	91.1	**91.8**	**91.8**
MIT67 R+FC BoP	Acc	**76.6**	–	75.4	75.7	74.7
	mAP	**78.8**	–	78.0	77.6	76.3
MIT67 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
MIT67 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 of Hard vs Soft 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

ing a soft-assignment with very large regularization, it does not happen when using ISA.

Another possible explanation of this difference in performance may lie in the fact that the Iterative Soft-Assign is stopped before convergence and does not satisfy the constraints $\mathbb{P}$ imposed on rows, whereas those constraints are satisfied when using the GFB algorithm. We conjecture that the constraint, i.e. “a part must occur in every positive image” in the original problem definition [8], is too strong and may need to be relaxed.Actually, as highlighted in the introduction (Sect. 1), the limitation of the separate optimization problem in comparison with the joint optimization is that a better optimization of the intermediate goal does not necessarily produce better final performance.

Focusing on the ISA method, the improved region description and encoding are evaluated, see Table 4. Using region proposals along with convolutional layer descriptions shows a significant performance gain, especially on the Willow dataset. We can see a consistent improvement for the SBoP and PCoP encoding as well and note that PCA yields more improvement on the descriptors based on fully connected layer than on the ones based on convolutional layers. These improvements set a new state of the art on both datasets, obtaining 91.9% mAP on Willow and 81.4% mAP on MIT67. Table 5 compares our best performance on MIT67 to a number of previous methods. Furthermore, we outperform the previous state of the art on Willow [17] with 81.7% mAP.

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. BoP+CoP and SBoP+PCoP refer to concatenated image descriptors.

Method	Measure	BoP	SBoP	CoP	PCoP	BoP+CoP	SBoP+PCoP
Willow R+FC	mAP	76.6	78.7	81.6	82.4	81.9	82.6
Willow P+C		89.2	90.1	91.6	91.7	91.8	**91.9**
MIT67 R+FC	Acc	76.6	76.1	76.8	77.1	78.1	78.3
	mAP	78.8	79.0	77.8	79.5	80.1	80.7
MIT67 P+C	Acc	75.1	76.1	80.0	80.5	81.1	**81.4**
	mAP	76.7	76.7	80.2	81.0	81.0	**81.2**

6. 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 to address

Table 5: Performance in terms of accuracy of existing part-based and non part-based methods on the MIT67 Scenes dataset.

Methods	Part-based	**MIT67**
Zhou *et al* [31]	No	70.8
Zuo *et al* [37]	Yes	76.2
Parizi *et al* [7]	Yes	77.1
Mettes *et al* [17]	Yes	77.4
Sicre *et al* [8]	Yes	78.1
Zheng *et al* [35]	No	78.4
Cimpoi *et al* [38]	No	81.0
Ours	Yes	**81.4**

this task have been investigated and compared to the previous method of [8] which achieves state of the art results.

We additionally proposed improvements on several stages of the classification pipeline, namely region extraction, region description and image encoding, using a recent very deep CNN architecture. This achieves a new state-of-the art performance on two different datasets. Furthermore, the new region description method is orders of magnitude faster, as this process was previously the bottleneck in [8].

Our experiments show that, in the context of part-based image classification, soft assignment outperforms hard assignment. Moreover, entropic regularization is more appropriate than quadratic regularization, while the best overall performance is obtained when one constraint is not fully satisfied. While it is a common constraint to consider that a part must occur in every positive image, this interesting finding shows that this constraint may need to be relaxed.

Our reformulation and investigation of different optimization methods allow the exploration of the limits of the original problem, such as defined in [8]. We believe this knowledge will help the community in the search for more appropriate models, potentially end-to-end trainable, using better network architectures.

Acknowledgment

Part of this work was achieved in the context of the IDFRAud project ANR-14-CE28-0012.

References

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

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

[3] J. J. Lim, C. L. Zitnick, 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, IEEE, 2013, pp. 3158–3165.

[4] S. Ullman, E. Sali, M. Vidal-Naquet, A Fragment-Based Approach to Object Representation and Classification, in: Visual Form 2001, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 85–100.

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

[6] C. Doersch, A. Gupta, A. A. Efros, Mid-level visual element discovery as discriminative mode seeking, in: Advances in Neural Information Processing Systems, 2013, pp. 494–502.

[7] S. N. Parizi, A. Vedaldi, A. Zisserman, P. Felzenszwalb, Automatic discovery and optimization of parts for image classification, in: International Conference on Learning Representations, 2015.

[8] R. Sicre, F. Jurie, Discriminative part model for visual recognition, Computer Vision and Image Understanding 141 (2015) 28 – 37. `doi:http://dx.doi.org/10.1016/j.cviu.2015.08.002`.
URL `http://www.sciencedirect.com/science/article/pii/S1077314215001642`

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

[10] F. Perronnin, J. Sanchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: European Conference on Computer Vision, 2010.

[11] S. Lazebnik, C. Schmid, 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.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.

[14] S. Singh, A. Gupta, A. A. Efros, Unsupervised discovery of mid-level discriminative patches, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 73–86.

[15] S. Maji, G. Shakhnarovich, Part discovery from partial correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 931–938.

[16] J. Sun, J. Ponce, Learning discriminative part detectors for image classification and cosegmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2013.

[17] P. Mettes, J. C. van Gemert, C. G. M. Snoek, No spare parts: Sharing part detectors for image categorization, CoRR abs/1510.04908.
URL http://arxiv.org/abs/1510.04908

[18] R. Burkard, M. Dell'Amico, S. Martello, Assignment Problems, Society for Industrial and Applied Mathematics, 2012. arXiv:http://epubs.siam.org/doi/pdf/10.1137/1.9781611972238, doi:10.1137/1.9781611972238.
URL http://epubs.siam.org/doi/abs/10.1137/1.9781611972238

[19] R. Sinkhorn, P. Knopp, Concerning nonnegative matrices and doubly stochastic matrices., Pacific J. Math. 21 (2) (1967) 343–348.
URL http://projecteuclid.org/euclid.pjm/1102992505

[20] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, L. Guibas, Convolutional wasserstein distances: Efficient optimal transportation on geometric domains, ACM Trans. Graph. 34 (4) (2015) 66:1–66:11. doi:10.1145/2766963.
URL http://doi.acm.org/10.1145/2766963

[21] L. Liu, L. Wang, X. Liu, In defense of soft-assignment coding, in: 2011 International Conference on Computer Vision, 2011, pp. 2486–2493. doi:10.1109/ICCV.2011.6126534.

[22] S. Bougleux, L. Brun, Linear sum assignment with edition, CoRR abs/1603.04380.
URL http://arxiv.org/abs/1603.04380

[23] M. Leordeanu, M. Hebert , R. Sukthankar, An integer projected fixed point method for graph matching and map inference, in: Proceedings Neural Information Processing Systems, Springer, 2009.

[24] A. Rangarajan, A. Yuille, E. Mjolsness, Convergence properties of the softassign quadratic assignment algorithm, Neural Comput. 11 (6) (1999) 1455–

1474. doi:10.1162/089976699300016313.
URL http://dx.doi.org/10.1162/089976699300016313

[25] H. Raguet, J. Fadili, G. Peyré, A generalized forward-backward splitting, SIIMS 6 (3) (2013) 1199–1226. doi:10.1137/120872802.

[26] L. Condat, Fast Projection onto the Simplex and the l1 Ball, to appear in Mathematical Programming Series A (Aug. 2015). doi:10.1007/s10107-015-0946-6.
URL https://hal.archives-ouvertes.fr/hal-01056171

[27] V. Delaitre, I. Laptev, 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, Vol. 2, 2010.

[28] A. Quattoni, A. Torralba., Recognizing indoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.

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

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

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning Deep Features for Scene Recognition using Places Database., in: Advances in Neural Information Processing Systems, 2014.

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

[33] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.

[34] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, 2016.
URL `http://arxiv.org/abs/1511.05879`

[35] L. Zheng, Y. Zhao, S. Wang, J. Wang, Q. Tian, Good practice in cnn feature transfer, arXiv preprint arXiv:1604.00133.

[36] R. Sicre, H. Jégou, Memory vectors for particular object retrieval with multiple queries, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 479–482.

[37] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, X. Jiang, Learning discriminative and shareable features for scene classification, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 552–568.

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