Unsupervised and Transfer Learning under Uncertainty: from Object Detections to Scene Categorization

Grégoire Mesnil\textsuperscript{1,2}, Salah Rifai\textsuperscript{1}, Antoine Bordes\textsuperscript{3}, Xavier Glorot\textsuperscript{1}, Yoshua Bengio\textsuperscript{1} and Pascal Vincent\textsuperscript{1}

\textsuperscript{1}LISA, Université de Montréal, Québec, Canada
\textsuperscript{2}LITIS, Université de Rouen, France
\textsuperscript{3}CNRS - Heudiasyc UMR 7253, Université de Technologie de Compiègne, France

gregoire.mesnil@umontreal.ca

Keywords: Unsupervised Learning, Transfer Learning, Deep Learning, Scene Categorization, Object Detection

Abstract: Classifying scenes (e.g. into “street”, “home” or “leisure”) is an important but complicated task nowadays, because images come with variability, ambiguity, and a wide range of illumination or scale conditions. Standard approaches build an intermediate representation of the global image and learn classifiers on it. Recently, it has been proposed to depict an image as an aggregation of its contained objects: the representation on which classifiers are trained is composed of many heterogeneous feature vectors derived from various object detectors. In this paper, we propose to study different approaches to efficiently combine the data extracted by these detectors. We use the features provided by Object-Bank (Li-Jia Li and Fei-Fei, 2010a) (177 different object detectors producing 252 attributes each), and show on several benchmarks for scene categorization that careful combinations, taking into account the structure of the data, allows to greatly improve over original results (from +5\% to +11\%) while drastically reducing the dimensionality of the representation by 97\% (from 44,604 to 1,000).

1 Introduction

Automatic scene categorization is crucial for many applications such as content-based image indexing (Smeulders et al., 2000) or image understanding. This is defined as the task of assigning images to predefined categories (“office”, “sailing”, “mountain”, etc.). Classifying scene is complicated because of the large variability of quality, subject and conditions of natural images which lead to many ambiguities w.r.t. the corresponding scene label.

Standard methods build an intermediate representation before classifying scenes by considering the image as a whole (Torralba, 2003; Vogel and Schiele, 2004; Fei-Fei and Perona, 2005; Oliva and Torralba, 2006). In particular, many such approaches rely on power spectral information, such as magnitude of spatial frequencies (Oliva and Torralba, 2006) or local texture descriptors (Fei-Fei and Perona, 2005). They have shown to perform well in cases where there are large numbers of scene categories.

Another line of work conveys promising potential in scene categorization. First applied to object recognition (Farhadi et al., 2009), attribute-based methods have now proved to be effective for dealing with complex scenes. These models define high-level representations by combining semantic lower-level elements, e.g., detection of object parts. A precursor of this tendency for scenes was an adaptation of pLSA (Hofmann, 2001) to deal with “visual words” proposed by (Bosch et al., 2006). An extension of this idea consists in modeling an image based on its content i.e. its objects (Espinace et al., 2010; Li-Jia Li and Fei-Fei, 2010a). Hence, the Object-Bank (OB) project (Li-Jia Li and Fei-Fei, 2010b) aims at building high-dimensional over-complete representations of scenes (of dimension 44,604) by combining the outputs of many object detectors (177) taken at various poses, scales and positions in the original image (leading to 252 attributes per detector). Experimental results indicate that this approach is effective since simple classifiers such as Support Vector Machines trained on their representations achieve state-of-the-art performance. However, this approach suffers from two flaws: (1) curse of dimensionality (very large number of features) and (2) individual object detectors have a poor precision (30\% at most). To solve (1), the
original paper proposes to use structured norms and group sparsity to make best use of the large input. Our work studies new ways to combine the very rich information provided by these multiple detectors, dealing with the uncertainty of the detections. A method designed to select and combine the most informative attributes would be able to carefully manage redundancy, noise and structure in the data, leading to better scene categorization performance.

Hence, in the following, we propose a sequential 2-steps strategy for combining the feature representations provided by the OB object detectors on which the linear SVM classifier is destined to be trained for categorizing scenes. The first step adapts Principal Components Analysis (PCA) to this particular setting: we show that it is crucial to take into account the structure of the data in order for PCA to perform well. The second one is based on Deep Learning. Deep Learning has emerged recently (see (Bengio, 2009) for a review) and is based on neural network algorithms able to discover data representations in an unsupervised fashion (Hinton et al., 2006; Bengio et al., 2007; Ranzato et al., 2007; Kavukcuoglu et al., 2009; Jarrett et al., 2009). We propose to use this ability to combine multiple detector features. Hence, we present a model trained using Contractive Auto-Encoders (Rifai et al., 2011b; Rifai et al., 2011a), which have already proved to be efficient on many image tasks and has contributed to winning a transfer learning challenge (Mesnil et al., 2012).

We validate the quality of our models in an extensive set of experiments in which several setups of the sequential feature extraction process are evaluated on benchmarks for scene classification (Lazebnik et al., 2006; Li and Fei-Fei, 2007; Quattoni and Torralba, 2009; Xiao et al., 2010). We show that our best results substantially outperform the original methods developed on top of OB features, while producing representations of much lower dimension. The performance gap is usually large, indicating that advanced combinations are highly beneficial. We show that our method based on dimensionality reduction followed by deep learning offers a flexibility which makes it able to benefit from semi-supervised and transfer learning.

2 Scene Categorization with Object-Bank

Let us begin by introducing the approach of the OB project (Li-Jia Li and Fei-Fei, 2010a). First, the 177 most useful (or frequent) objects were selected from popular image datasets such as LabelMe (Russell et al., 2008), ImageNet (Deng et al., 2009) and Flickr. For each of these 177 objects, a specific detector, existing in the literature (Felzenszwalb et al., 2008; Hoiem et al., 2005), was trained. Every detector is composed of 2 root filters depending on the pose, each one coming with its own deformable pattern of parts, e.g., there is one root filter for the front-view of a bike and one for the side-view. These 354 = 177 × 2 part-based filters (each composed by a root and its parts) are used to produce features of natural images. For a given image, a filter is convolved at 6 different scales. At each scale, the max-response among 21 = 1 + 4 + 16 positions (whole image, quadrants, quadrants within each quadrant) is kept, producing a response map of dimension 126 = 6 × 21. All 2 × 177 maps are finally concatenated to produce an over-complete representation \( x \in \mathbb{R}^{44,604} \) of the original image.

In the original OB paper (Li-Jia Li and Fei-Fei, 2010a), classifiers for scene categorization are learned directly on these feature vectors of dimension 44,604. More precisely, \( C \) classifiers (Linear SVM or Logistic Regression) are trained in a 1-versus-all setting in order to predict the correct scene category \( y_{\text{category}}(x) \) among \( C \) different categories. Various strategies using structured sparsity with combinations of \( \ell_1/\ell_2 \) norms have been proposed to handle the very large input.

3 Unsupervised Feature Learning

The approach of OB for the task of scene categorization, based on specific object detectors, is appealing since it works well in practice. This suggests that a scene is better recognized by first identifying basic objects and then exploiting the underlying semantics in the dependencies between the corresponding detectors.

However, it appears that none of the individual object detectors reaches a recognition precision of more than 30%. Hence, one may question whether the ideal view that inspired this approach (and expressed above) is indeed the reason of OB’s success. Alternatively, one may hypothesize that the 44,604 OB features are more useful for scene categorization because they represent high level statistical properties of images than because they precisely report the absence/presence of objects — see Figure 1. OB tried structured sparsity to handle this feature selection but there may be other ways — simpler or not.

This paper investigates several ways of learning higher-level features on top of the high dimensional representation provided by OB, expecting that capturing further structure may improve categorization.
performance. Our approach employs unsupervised feature learning/extraction algorithms, i.e. generic feature extraction methods which were not developed specifically for images. We will consider both standard Principal Component Analysis and Contractive Auto-Encoders (Rifai et al., 2011b; Rifai et al., 2011a). The latter is a recent machine learning method which has proved to be a robust feature extraction tool.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933) is the most prevalent technique for linear dimensionality reduction. A PCA with \( k \) components finds the \( k \) orthonormal directions of projection in input space that retain most of the variance of the training data. These correspond to the eigenvectors associated with the leading eigenvalues of the training data’s covariance matrix. Principal components are ordered, so that the first corresponds to the direction along which the data varies the most (largest eigenvalue), etc. . .

Since we will consider an auto-encoder variant (presented next), we should mention here a well-known result: a linear auto-encoder with \( k \) hidden units, trained to minimize squared reconstruction error, will learn projection directions that span the same subspace as a \( k \) component PCA (Baldi and Hornik, 1989). However the regularized non-linear auto-encoder variant that we consider below is capable of extracting qualitatively different, and usually more useful, nonlinear features.

3.2 Contractive Auto-Encoders

Contractive Auto-Encoders (CAEs) (Rifai et al., 2011b; Rifai et al., 2011a) are among the latest developments in a line of machine learning research on nonlinear feature learning methods, that started with the success of Restricted Boltzmann Machines (Hinton et al., 2006) for pre-training deep networks, and was followed by other variants of auto-encoders such as sparse (Ranzato et al., 2007; Kavukcuoglu et al., 2009; Goodfellow et al., 2009) and denoising auto-encoders (Vincent et al., 2008). It was selected here mainly due to its practical ease of use and recent empirical successes.

Unlike PCA that decomposes the input space into leading global directions of variations, the CAE learns features that capture local directions of variation (in some regions of input space). This is achieved by penalizing the norm of the Jacobian of a latent representation \( h(x) \) with respect to its input \( x \) at training samples. Rifai et al. (Rifai et al., 2011b) show that the resulting features provide a local coordinate system for a low dimensional manifold of the input space. This corresponds to an atlas of charts, each corresponding to a different region in input space, associated with a different set of active latent features. One can think about this as being similar to a mixture of PCAs, each computed on a different set of training samples that were grouped together using a similarity criterion (and corresponding to a different input re-
tion), but without using an independent parametra-
ization for each component of the mixture, i.e., allow-
ing to generalize across the charts, and away from the
training examples.

In the following, we summarize the formulation
of the CAE as a regularized extension of a
basic Auto-Encoder (AE). In our experiments, the
parametrization of this AE consists in a non-linear
encoder or latent representation $h$ of $m$ hidden
units with a linear decoder or reconstruction $g$ towards
an input space of dimension $d$.

Formally, the latent variables are parametrized by:

$$h(x) = s(Wx + b_h),$$  \hspace{1cm} (1)

where $s$ is the element-wise logistic sigmoid $s(z) = \frac{1}{1 + e^{-z}}, W \in \mathbb{M}_{m \times d}(\mathbb{R})$ and $b_h \in \mathbb{R}^m$ are the parameters
to be learned during training. Conversely, the units of
the decoder are linear projections of $h(x)$ back into
the input space:

$$g(h(x)) = W^T h(x).$$  \hspace{1cm} (2)

Using mean squared error as the reconstruction objec-
tive and the L2-norm of the Jacobian of $h$ with respect
to $x$ as regularization, training is carried out by min-
imizing the following criterion by stochastic gradient
descent:

$$J_{CAE}(\Theta) = \sum_{x \in \mathcal{D}} \|x - g(h(x))\|^2 + \lambda \sum_{i=1}^m \sum_{j=1}^d \left| \frac{\partial h_i}{\partial x_j}(x) \right|^2,$$  \hspace{1cm} (3)

where $\Theta = \{W, b_h\}$, $\mathcal{D} = \{x^{(i)}\}_{i=1,...,n}$ corre-
sponds to a set of $n$ training samples $x \in \mathbb{R}^d$ and $\lambda$
is a hyper-parameter controlling the level of contraction
of $h$. A notable difference between CAEs and
PCA is that features extracted by CAEs are non-linear
w.r.t. the inputs, so that multiple layers of CAEs can
be usefully composed (stacked), whereas stacking lin-
ear PCAs is pointless.

4 Extracting Better Features with
Advanced Combination Strategies

In this work, we study two different sub-structures
of OB. We consider the pose response defined by the
output of only one part-based filter at all positions and
scales, and the object response which is the concate-
nation of all pose responses associated to an object.
Combination strategies are depicted in Figure 2.

4.1 Simplistic Strategies: Mean and
Max Pooling

The idea of pooling responses at different locations
or poses has been successfully used in Convolutional
Neural Networks such as LeNet-5 (LeCun et al.,
1999) and other visual processing (Serre et al., 2005)
architectures inspired by the visual cortex.

Here, we pool the 252 responses of each object
detector into one component (using the mean or max
operator) leading to a representation of size $177 =
44604/252$. It corresponds to the mean/max over the
object responses at different scales and locations. One
may view the object max responses as features encod-
ing absence/presence of objects while discarding all
the information about the detector’s positions.

4.2 Combination Strategies with PCA

PCA is a standard method for extracting features from
high dimensional input, so it is a good starting point.
However, as we find in our experiments, exploiting
the particular structure of the data, e.g., according to
poses, scales, and locations, can yield to improved re-

Whole PCA. An ordinary PCA is trained on the raw
output of OB ($x \in \mathbb{R}^{44\times604}$) without looking for any
structure. Given the high-dimensionality of OB’s rep-
resentation, we used the Randomized PCA algorithm
of the scikits toolbox\footnote{Available from http://scikits.appspot.com/}.

Pose-PCA. Each of the two poses associated with
each object detector is considered independently.
This results in 354 = 2 × 177 different PCAs, which
are trained on pose outputs ($x \in \mathbb{R}^{126}$) – see Figure 2.

Object-PCA. Only each object response ($x \in \mathbb{R}^{252}$)
is considered separately, therefore 177 PCAs are
trained in total. It allows the model to capture
variations among all pose responses at various scales
and positions – see Figure 2.

Note that, in all cases, whitening the PCA (i.e. di-
viding each eigenvector’s response by the correspond-
ing squared root eigenvalue) performs very poorly.
For post-processing, the PCA outputs $\tilde{x}$ are always
normalized: $\tilde{x} \leftarrow (\tilde{x} - \mu)/\sigma$ according to mean $\mu$ and
the deviation $\sigma$ of the whole, per object or per pose
PCA outputs. Thereby, we ensure contributions from
all objects or poses to be in the same range. The num-
ber of components in all cases has been selected ac-

4.3 Improving upon PCA with CAE

Due to hardware limitations and high-dimensional
input, we could not train a CAE on the whole OB output
Figure 2: Different Combination Strategies (a) and (b) pose and object PCAs (c) high-level CAE: pose-PCA as dimensionality reduction technique in the first layer and a CAE stacked on top. We denote it high-level because it can learn context information i.e. plausible joint appearance of different objects.

(“whole CAE”). However, we address this problem with the sequential feature extraction steps below.

To overcome the tractability problem that forbids a CAE to be trained on the whole OB output, we pre-process it by using the pose-PCAs as a dimensionality reduction method. We keep only the 5 first components of each pose. Given this low-dimensional representation (of dimension 1,770), we are able to train a CAE – see Figure 2. The CAE has a global view of all object detectors and can thus learn to capture context information, defined by the joint appearance of combinations of various objects. Moreover, instead of using an SVM on top of the learned representations, we can use a Multi-Layer Perceptron whose weights would be initialized by those of this CAE. This setting is where the CAE has shown to perform best in practice (Rifai et al., 2011a).

5 Experiments

5.1 Datasets

We evaluate our approach on 3 scene datasets, cluttered indoor images (MIT Indoor Scene), natural scenes (15-Scenes), and event/activity images (UIUC-Sports). Images from a large scale scene recognition dataset (SUN-397 database) have also been used for unsupervised learning.

* MIT Indoor is composed of 67 categories and, following (Li-Jia Li and Fei-Fei, 2010a; Quattoni and Torralba, 2009), we used 80 images from each category for training and 20 for testing.

* UIUC-Sports contains 8 event classes. We randomly chose 70 / 60 images for our training / test set respectively, following the setting of (Li-Jia Li and Fei-Fei, 2010a; Li and Fei-Fei, 2007).

* SUN-397 contains a full variety of 397 well sampled scene categories (100 samples per class) composed of 108,754 images in total.

5.2 Tasks

We consider 3 different tasks to evaluate and compare the considered combination strategies. In particular, various supervision settings for learning the CAE are explored. Indeed, a great advantage of this kind of method is that it can make use of vast quantities of unlabeled examples to improve its representations. We thus illustrate this by proposing experiments in which the CAE has been trained in supervised or in semi-supervised way and also in a transfer context.

MIT Indoor (plain). Only the official training set of the MIT Indoor scene dataset (5,360 images) is used for unsupervised feature learning. Each representation is evaluated by training a linear SVM on top of the learned features.

MIT+SUN (semi-supervised). This task, like the previous one, uses the official train/test split of the MIT Indoor scene dataset for its supervised training and evaluation of scene categorization performance. For the initial unsupervised feature extraction however, we augmented the MIT Indoor training set with the whole dataset of images from SUN-397 (108,754 images). This yields a total of 123,034 images for unsupervised feature learning and corresponds to a semi-supervised setting. Our motivation for adding scene images from SUN, besides increasing the number of training samples, is that on MIT Indoor, which contains only indoor scenes, OB detectors specialized on
Table 1: MIT Indoor. Results are reported on the official split (Quattoni and Torralba, 2009) for all combination strategies described in Section 4. Only the unsupervised feature learning strategies (PCA and CAE based) can benefit from the addition of unlabeled scenes from SUN. Object Bank + SVM refers to the original system (Li-Jia Li and Fei-Fei, 2010a) and DPM + Gist + SP (Pandey and Lazebnik, 2011) corresponds to the state-of-the-art method on MIT Indoor.

5.3 SVMs on Features Learned with each Strategy

In order to evaluate the quality of the features generated by each strategy, a linear SVM is trained on the features extracted by each combination method. We used LibLinear (Fan et al., 2008) as SVM solver and chose the best C according to 5-fold cross-validation scheme. We compare accuracies obtained by features provided by all considered combination methods against the original OB performances (Li-Jia Li and Fei-Fei, 2010a). Results obtained with SVM classifiers on all MIT-related tasks are displayed in Table 1 and those concerning UIUC and 15-scenes in Table 2.

The simplistic strategy object mean-pooling performs surprisingly well on all datasets and tasks whereas object max-pooling obtained the worst results. It suggests that taking the mean response of an object detector across various scales and positions is actually meaningful compared to considering presence/absence of objects as max-pooling does.

On MIT and MIT+SUN, object or pose PCAs reach almost the same range of performance slightly above the current state-of-the-art performances (Pandey and Lazebnik, 2011), except for whole-PCA which performs poorly: one must consider the structure of OB to combine features efficiently. In the experiments, keeping the 10 (resp. 15) first principal components gave us the best results for pose-PCA (resp. object-PCA).

Besides, Table 3 shows that both PCAs and PCA+CAE allow a huge reduction of the dimension of the OB feature representation. Results obtained for the UIUC-Sports and 15-
Scenes transfer learning tasks are displayed in Table 2. Representations learned on MIT+SUN generalize quite well and can be easily used for other datasets even if images from those datasets have not been seen at all during unsupervised learning.

## 5.4 Deep Learning with Fine Tuning

Previous work (Larochelle et al., 2009) on Deep Learning generally showed that the features learned through unsupervised learning could be improved upon by fine-tuning them through a supervised training stage. In this stage (which follows the unsupervised pre-training stage), the features and the classifier on top of them are together considered to be a supervised neural network, a Multi-Layer Perception (MLP) whose hidden layer is the output of the trained features. Hence we apply this strategy to the pose-PCA+CAE architecture, keeping the PCA transformation fixed but fine-tuning the CAE and the MLP altogether. These results are given at the bottom of tables 1 and 2. The MLP are trained with early stopping of joint appearance of several objects learned by the CAEs; these learned representations capture statistical dependencies in the co-occurrence of detections the object detectors (OB features) that have constituted the basis of very promising trend of approach for scene classification (Li-Jia Li and Fei-Fei, 2010a). These higher-level representations are mostly trained in an unsupervised way, following the trend of so-called Deep Learning (Hinton et al., 2006; Bengio, 2009; Jarrett et al., 2009), but can be fine-tuned using the supervised detection objective.

Finally, we trained a non-linear SVM (with rbf kernel) to verify whether this gap in performances was simply due to the replacement of a linear classifier (SVM) by a non-linear one (MLP) or to the detectors’ outputs combination. The poor results of the rbf-SVM (see tables 1 and 2) suggests that the careful combination strategies are essential to reach good performance.

## 6 Discussion

In this work, we add one or more levels of trained representations on top of the layer of object and part detectors (OB features) that have constituted the basis of very promising trend of approach for scene classification (Li-Jia Li and Fei-Fei, 2010a). These higher-level representations are mostly trained in an unsupervised way, following the trend of so-called Deep Learning (Hinton et al., 2006; Bengio, 2009; Jarrett et al., 2009), but can be fine-tuned using the supervised detection objective. These learned representations capture statistical dependencies in the co-occurrence of detections the object detectors from (Li-Jia Li and Fei-Fei, 2010a). In fact, one can see in Table 4 plausible contexts of joint appearance of several objects learned by the CAE. These detectors, which can be quite imperfect when seen as actual detectors, contain a lot of information when combined altogether. The extraction of those context semantics with unsupervised feature-learning algorithms has empirically shown better performances.

In particular, we find that Contractive Auto-Encoder (Rifai et al., 2011b; Rifai et al., 2011a) can substantially improve performance on top of pose PCAs as a way to extract non-linear dependencies between these lower-level OB detectors (especially when fine-tuned). They also improve greatly upon the use of the detectors as inputs to an SVM or a lo-

<table>
<thead>
<tr>
<th></th>
<th>UIUC-Sports</th>
<th>15-SCENES</th>
</tr>
</thead>
<tbody>
<tr>
<td>object-MAX + SVM</td>
<td>67.23 ± 1.29%</td>
<td>71.08 ± 0.57%</td>
</tr>
<tr>
<td>object-MEAN + SVM</td>
<td>81.88 ± 1.16%</td>
<td>83.17 ± 0.53%</td>
</tr>
<tr>
<td>object-PCA + SVM</td>
<td>83.90 ± 1.67%</td>
<td>85.58 ± 0.48%</td>
</tr>
<tr>
<td>pose-PCA + SVM</td>
<td>83.81 ± 2.22%</td>
<td>85.69 ± 0.39%</td>
</tr>
<tr>
<td>pose-PCA + MLP</td>
<td>84.29 ± 2.23%</td>
<td>84.93 ± 0.39%</td>
</tr>
<tr>
<td>pose-PCA + CAE (MLP)</td>
<td><strong>85.13 ± 1.07%</strong></td>
<td><strong>86.44 ± 0.21%</strong></td>
</tr>
<tr>
<td>Object Bank + SVM</td>
<td>78.90%</td>
<td>80.98%</td>
</tr>
<tr>
<td>Object Bank + rbf-SVM</td>
<td>78.56 ± 1.50%</td>
<td>83.71 ± 0.64%</td>
</tr>
<tr>
<td>Improvement w.r.t. OB</td>
<td>+6.23%</td>
<td>+5.46%</td>
</tr>
</tbody>
</table>

Table 2: **UIUC Sports and 15-Scenes** Results are reported for 10 random splits (available at www.anonymous.org) and compared to the original OB results (Li-Jia Li and Fei-Fei, 2010a) - Object Bank + SVM - on one single split.
Table 3: **Dimensionality Reduction.** Dimension of representations obtained on MIT Indoor. The pose-PCA+CAE produces a compact and powerful combination.

<table>
<thead>
<tr>
<th>Object-Bank</th>
<th>Pooling</th>
<th>whole-PCA</th>
<th>object-PCA</th>
<th>pose-PCA</th>
<th>pose-PCA+CAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>44,604</td>
<td>177</td>
<td>1,300</td>
<td>2,655</td>
<td>1,770</td>
<td>1,000</td>
</tr>
</tbody>
</table>

Table 4: **Context semantics.** Names of the detectors corresponding to the highest weights of 8 hidden units of the CAE. These hidden units will fire when those objects will be detected altogether.

| sailboat, rock, tree, coral, blind | roller coaster, building, rail, keyboard, bridge | soil, seashore, rock, mountain, duck | attire, horse, bride, groom, bouquet | desktop computer, printer, wireless, computer screen |

This trained post-processing allows us to reach the state-of-the-art on MIT Indoor and UIUC (85.13% against 85.30% obtained by LScSPM (Gao et al., 2010)) while being competitive on 15-scenes (86.44% also versus 89.70% LScSPM). On these last two datasets, we reach the best performance for methods only relying on object/part detectors. Compared to other kinds of methods, we are limited by the accuracy of those detectors (only trained on HOG features), whereas competitive methods can make use of other descriptors such as SIFT (Gao et al., 2010), known to achieve excellent performance in image recognition.

Besides its good accuracies, it is worth noting that the feature representation obtained by the pose-PCA+CAE is also very compact, allowing a 97% reduction compared to the original data (see Table 3). Handling a dense input of dimension 44,604 is not a common thing. By providing this compact representation, we think that researchers will be able to use the rich information provided by OB in the same way they use low-level image descriptors such as SIFT.

As future work, we are planning other ways of combining OB features e.g. considering the output of all detectors at a given scale and position and combine them afterwards in a hierarchical manner. This would be a kind of dual view of the OB features. Other plausible departures could take into account the topology (e.g. spatial structure) of the pattern of detections, rather than treat the response at each location and scale as an attribute and the set of attributes as unordered. This could be done in the same spirit as in Convolutional Networks (LeCun et al., 1999), aggregating the responses for various objects detectors/locations/scales in a way that takes explicitly into account the object category, location and scale of each response, similarly to the way filter outputs at neighboring locations are pooled in each layer of a Convolutional Network.

**Acknowledgments:** We would like to thank Gloria Zen for her helpful comments. This work was supported by NSERC, CIFAR, the Canada Research Chairs, Compute Canada and by the French ANR Project ASAP ANR-09-EMER-001. Codes for the experiments have been implemented using Theano (Bergstra et al., 2010) Machine Learning library.

**REFERENCES**


Li, L.-J. and Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. ICCV.


Li-Jia Li, Hao Su, Y. L. and Fei-Fei, L. (2010b). Objects as attributes for scene classification. In European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes, Crete, Greece.


