

Slightly Supervised Learning of Part-Based Appearance Models

Lexing Xie

Patrick Pérez

Dept. of Electrical Engineering
Columbia University
New York, NY 10027

Irisa/Inria
Campus de Beaulieu
F-35042 Rennes Cedex

Abstract

We extend the GMM-based approach of [17], for learning part-based appearance models of object categories, to the unsupervised case where positive examples are corrupted with clutter. To this end, we derive an original version of EM which is able to fit one GMM per class based on partially labeled data. We also allow ourselves a small fraction of un-corrupted positive examples, thus obtaining an effective, yet cheap, slightly supervised learning. Proposed technique allows as well a saliency-based ranking and selection of learnt mixture components. Experiments show that both the semi-supervised GMM fitting with side information and the component selection are effective in identifying salient patches in the appearance of a class of objects. They are thus promising tools to learn class-specific models and detectors similar to those by Weber *et al.*[6], but at a lower computational cost, while accommodating larger numbers of atomic parts.

1. Introduction

In the past four years, a number of studies have investigated various ways of learning part-based appearance models for object categories. Representative examples include [9, 3, 8, 14, 16, 6]. The idea is to extract and model the appearance, and possibly the localization, of a number of visual fragments that are specific to a given class of objects seen from one or few specified viewpoints. These approaches differ in many respects, but they all rely on the following common bases:

1. Some sort of low-level *interest point* detector, for example, an edge or corner detector, is used to focus the attention of the learning procedure.
2. The raw intensity of the image around each of these points is summarized by a *descriptor*, which might include or not various forms of invariance.
3. The large set of descriptors stemming from the image training set is then *clustered*. Each cluster can be seen as a possible appearance fragment, and can be associated, if required, to an individual part detector.

4. A global category-specific visual detector is learnt using the set of parts previously extracted or a subset of them obtained through *selection*.

The different approaches defined along these lines prove very promising. However, they are usually confined to *supervised* setups where training images are carefully labeled. This makes both learning phases 3 and 4 simpler in that positive examples are not corrupted by irrelevant clutter, and crucial information about part localization is often accessible. The high cost of manually preparing the training sets is however a major limitation.

Unsupervised scenarios are only considered by Perona *et al.* [11, 15, 5, 6]. Positive examples are images exhibiting, at an unknown location, one occurrence of the object class before a cluttered background. A very ambitious probabilistic approach is proposed to learn jointly a small set of salient parts, their relative localization, and their appearance. The formidable complexity of the resulting learning procedure requires however appropriate heuristics be designed, and, more importantly, that the number of parts involved remains extremely small (at most five).

In this paper, we propose to re-examine the generic synopsis above in an unsupervised context similar to the one of Perona *et al.*, as a first step towards lighter techniques for unsupervised learning of part-based appearance models. We shall use in particular Gaussian mixture models (GMMs) and learn them jointly on both positive and negative examples, using appropriate variants of the Expectation-Maximization (EM) algorithm. This framework will also allow us to introduce easily a small but beneficial fraction of supervision with a few “clean” positive examples. These different ingredients are presented in Section 2.

In Section 3 we discuss the ranking of the mixture components with an information-theoretic criteria and the subsequent pruning to discard both least salient fragments and those stemming, at least partly, from the clutter present in the corrupted positive examples.

We shall report a few experiments in Section 4; before turning to a number of future directions in Section 5.

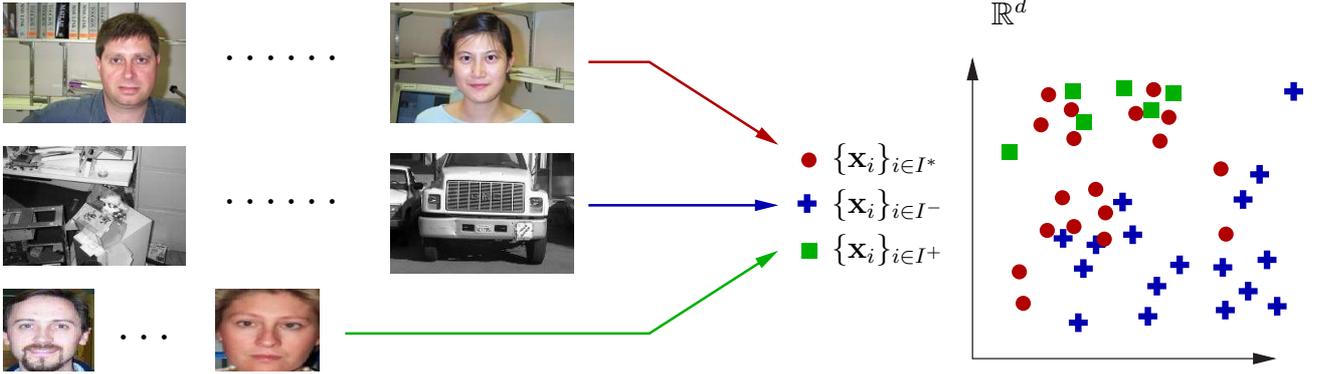


Figure 1: **Slightly supervised setup for learning part-based class-specific appearance models.** Two large collections of positive examples with clutter and negative examples, along with a small set of clutter-free positive examples are used. Generic interest-point are detected in all images and each detected point is associated with a d -dimensional descriptor based on the intensity pattern in its neighborhood. The three previous image subsets yield the three descriptor sets $\{\mathbf{x}_i\}_{i \in I^*}$, $\{\mathbf{x}_i\}_{i \in I^-}$, and $\{\mathbf{x}_i\}_{i \in I^+}$, respectively.

2. Slightly supervised GMM learning

Consider a training set composed of positive images corrupted with clutter and negative images (pure background). As opposed to Perona *et al.*, we also allow ourselves a few clutter-free examples obtained by manually framing up the object of interest in a handful of images. Some generic interest point detector (or a combination of them) is run on all these images (e.g., Harris corner detector as in [17, 8, 14], Forstner detector as in [9, 6], entropy-based Kadir-Brady blob detector as in [15], Difference-of-Gaussian blob detector as in [17], etc.).

The intensity pattern around each detected point is summarized by a d -dimensional vector $\mathbf{x} \in \mathbb{R}^d$. Examples include raw intensity patches (d is the number of pixels of the window considered) as in [9, 14, 6], subsets of wavelet or PCA coefficients as in [3, 8], or SIFT descriptors with $d = 128$ as in [17].

Let $\{\mathbf{x}_i\}_{i \in I}$ be the set of all descriptors collected on the complete training set, with the index set being partitioned as $I = I^+ \cup I^* \cup I^-$, where I^+ , I^* and I^- are respectively associated to the clean positive images, the cluttered ones, and the negative images. Figure 1 illustrates the overall setup.

Let $\{y_i\}_{i \in I}$ be the associated binary labels, where $y_i = +1$ stands for “foreground” and $y_i = -1$ stands for “background”. Two descriptor densities have to be learnt, for the positive and negative classes respectively. Following the supervised approach in [17], we resort to GMMs, as opposed to non-parametric clustering used in most of the other studies. Apart from the advantage of giving access to densities, EM-based GMM fitting is in addition well suited to the semi-supervised extension we require.

Formally, we want to fit a two-fold GMM

$$p(\mathbf{x}) = \sum_{m \in M^+ \cup M^-} \pi_m N(\mathbf{x}; \boldsymbol{\mu}_m, \Gamma_m), \text{ with } \sum_m \pi_m = 1, \quad (1)$$

where mixture subsets indexed by M^- and M^+ respectively are associated to classes -1 and $+1$ respectively. If $z \in M$ stands for the hidden variable that indicates which mixture component \mathbf{x} stems from, we have the deterministic relation:

$$y = \begin{cases} -1 & \text{if } z \in M^- \\ +1 & \text{if } z \in M^+. \end{cases} \quad (2)$$

Setting

$$p(z = m) = \pi_m \text{ and } p(\mathbf{x}|z = m) = N(\mathbf{x}; \boldsymbol{\mu}_m, \Gamma_m), \quad (3)$$

we obtain a joint model,

$$p(\mathbf{x}, y, z) = N(\mathbf{x}, \boldsymbol{\mu}_z, \Gamma_z) \pi_z \mathbf{1}_{M^{\text{sn}(y)}}(z), \quad (4)$$

over $\mathbb{R}^d \times \{-1, +1\} \times M$, whose marginal on \mathbf{x} coincides with (1). Notation $\mathbf{1}_A$ stands for the characteristic function of discrete set A . The associated graphical model is depicted in Fig. 2.

For data points \mathbf{x}_i , $i \in I^+ \cup I^-$, the class label y_i is observed, whereas it is not for $i \in I^*$. As demonstrated in other contexts, the generic EM approach to learning with incomplete data works in particular with such mixes of labeled data and unlabeled (or, similarly, noisily labeled) data [7, 4]. Our problem also falls in the class of learning problems with *side information*, where the side information is, in this case, composed of the available data labels. Other examples of side information include “equivalence” and “non-equivalence” constraints between samples as in [13, 12].

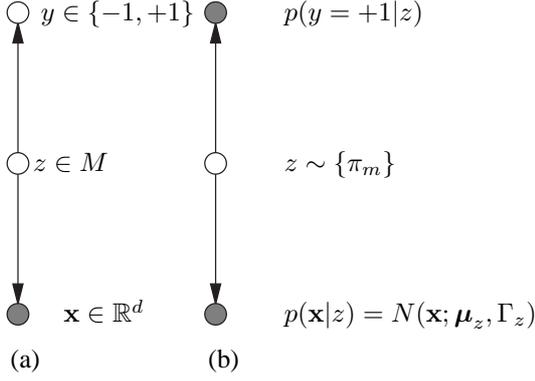


Figure 2: **Graphical model associated to the two-fold GMM to be learnt.** Learning is based on observing (a) \mathbf{x} only, on interest points from cluttered positive images, and (b) \mathbf{x} and class label y , on interest points from negative images ($y = -1$) and from a small number of clean positive images ($y = +1$). Shaded nodes correspond to observed variables.

In our specific case, the classic EM-based GMM fitting is readily extended. The updating of current parameter estimate $\theta^{(k)} = (\theta_m^{(k)})_m = (\pi_m^{(k)}, \boldsymbol{\mu}_m^{(k)}, \Gamma_m^{(k)})_m$ at step k is based on the following individual responsibilities:

$$\begin{aligned} \forall i \in I^+ \cup I^-, \xi_i^{(k)}(m) &= p(z_i = m | \mathbf{x}_i, y_i, \theta^{(k)}) \\ &\propto N(\mathbf{x}_i; \boldsymbol{\mu}_m^{(k)}, \Gamma_m^{(k)}) \pi_m^{(k)} \mathbf{1}_{M^{\text{sgn}(y_i)}} \\ \forall i \in I^*, \xi_i^{(k)}(m) &= p(z_i = m | \mathbf{x}_i, \theta^{(k)}) \\ &\propto N(\mathbf{x}_i; \boldsymbol{\mu}_m^{(k)}, \Gamma_m^{(k)}) \pi_m^{(k)}, \end{aligned} \quad (5)$$

all normalized to one over $M = M^+ \cup M^-$. The update of mean vectors, for instance, is then classically given by:

$$\boldsymbol{\mu}_m^{(k+1)} = \frac{\sum_{i \in I} \xi_i^{(k)}(m) \mathbf{x}_i}{\sum_{i \in I} \xi_i^{(k)}(m)}. \quad (6)$$

Given definitions (5), one can notice that:

- if $I^+ = I^- = \emptyset$, the problem boils down to a standard GMM fitting over $\{\mathbf{x}_i\}_{i \in I^*}$, with no particular meaning for the two subsets of mixture components;
- if $I^* = \emptyset$, the problem boils down to the independent fitting of two GMMs on data points $\{\mathbf{x}_i\}_{i \in I^+}$ and $\{\mathbf{x}_i\}_{i \in I^-}$ respectively;
- if no index subset is empty, new parameters $\theta_m^{(k+1)}$ for a positive component $m \in M^+$ depend not only on data points $\{\mathbf{x}_i\}_{i \in I^+ \cup I^*}$, but also on the full set of former parameters $\theta^{(k)}$. Hence they also depend on data points $\{\mathbf{x}_i\}_{i \in I^-}$.

The presence of the small set $\{\mathbf{x}_i\}_{i \in I^+}$ of points labeled for sure as foreground should help the unsupervised learning on mixed data $\{\mathbf{x}_i\}_{i \in I^*}$, by providing a sort of anchor for the EM-based soft labeling.

An additional way to make the most out of this certain information is to use it at initialization time, as initialization is indeed crucial for GMM fitting (as well as for other clustering techniques). The set $\{\mathbf{x}_i\}_{i \in I^+}$ being clustered beforehand with K -means, we choose to initialize a subset of the positive mixture components around these clusters.

3. Component selection

The previous step, of GMM fitting in our case, of clustering in most other approaches typically outputs a large number (e.g., from 50 to 500) of atomic visual components. Each of them is associated to one mixture component in case of GMMs.

These components are related to some extent to the parts of the objects of interest. This relation, however, is neither simple, nor one-to-one. As we shall see, different components can be associated to a same object part (seen at different resolutions or with different shifts), whereas some other components can arise from multiple object and background parts which are locally similar. This mapping between atomic visual components and semantic object parts remains a difficult and open problem that we will not try to address here.

In any case, selecting a small subset of highly *salient* components among the large set previously obtained should provide appearance models that are more compact, hence lighter to learn and manipulate, and hopefully more discriminant. Such a pruning step is however not mandatory in supervised setups: the use of interest point detectors as a preliminary sieve to discard least informative parts and the access to part localization are sometimes sufficient to learn part-based appearance models [9, 8]. However, when localization is not part of the descriptor, it is very important to separate out visual fragments that characterize the class of object of interest from those routinely seen in both positive and negative examples [17, 16]. Selection can be based on some assessment of the information content of the fragment relative to the object class, or, in other words, on the discriminant power of the fragment alone.

In the case of unsupervised or semi-supervised learning, corruption of the positive examples by clutter makes this issue of retaining only the most salient features even more important. In this adverse context, selection is also crucial to cut down the complexity of the model to be fitted.

The problem of selecting a subset of relevant parts from a larger noisy pool is analogous to the feature selection problem [2] where a subset of dimensions of the input space is selected according to its relevance to the target concept. Solutions to feature/part selection dichotomize into *filter*

and *wrapper* methods [2]: the former evaluate *relevance* from the dependency relationships between individual features and the learning target, while *relevance* in the latter is wrapped around the particular learning algorithm being used.

The part selection scheme used by Perona *et al.* [15, 6] falls into the *wrapper* class. Each subset of P candidate parts, out of a total of N , is fitted to the joint appearance-shape model, and the one with the best likelihood is eventually retained. Although the selection criteria are consistent with the overall objective of the model (likelihood), searching through all possible subsets incurs a high complexity of $O(N^P)$ on top of the exponential complexity for inferring a fully-connected shape model. Hence, greedy [6] or randomized [15] search strategies were used to render this process feasible even for a small number ($5 \sim 6$) of parts.

As done by Dorko and Schmid in a clutter-free context [17], we propose to use a *filter* method for part selection prior to learning a layout model (if any), for its ability to accommodate a much larger number of parts with much less computation. Having many fragments, hence a richer model, should improve the robustness of the object detector (e.g., higher resistance to severe occlusion) and its flexibility (e.g., the capture of multiple poses) [8].

The criteria for filtering out irrelevance can be mutual information [17, 2], likelihood ratio [17], or their variants.

In our context, such criteria will rely on the joint distribution of (y, z) . However, the model used in Section 2 to fit the two-fold GMM assumes the deterministic relation (2). The derivation of the above-mentioned criteria thus requires that we loosen this relationship. The rest of the model being kept unchanged, we now set:

$$p(y = +1|z = m) = \alpha_m, \quad (7)$$

where parameters $\{\alpha_m\}_{m \in M}$ have to be learnt.

The mutual information criterion measures the statistical dependency between binary random variables y and $\mathbf{1}_{\{m\}}(z)$. As demonstrated in [17], the higher this measure for a mixture component m , the more “informative” the feature relative to class labels. In the prospect of building an object-specific detector, focusing on the discriminant power of a feature seems more appropriate. This power is better characterized by likelihood ratios [17]:

$$L_m = \frac{p(z = m|y = +1)}{p(z = m|y = -1)}. \quad (8)$$

With our model, this score reads

$$L_m \propto \frac{\alpha_m}{1 - \alpha_m} \quad (9)$$

up to a multiplicative factor independent of m . Ranking mixture components according to L_m then simply boils down to rank them according to α_m .

We are left with the problem of estimating parameters $\{\alpha_m\}_{m \in M}$. This problem can be addressed with EM, in a fashion similar to the estimation of mixture parameters in Section 2. The new definition of the individual responsibilities is:

$$\begin{aligned} \forall i \in I^+, \xi_i^{(k)}(m) &= p(z_i = m | \mathbf{x}_i, y_i = +1, \alpha^{(k)}) \\ &\propto \alpha_m^{(k)} N(\mathbf{x}_i; \boldsymbol{\mu}_m, \Gamma_m) \pi_m \\ \forall i \in I^-, \xi_i^{(k)}(m) &= p(z_i = m | \mathbf{x}_i, y_i = -1, \alpha^{(k)}) \\ &\propto (1 - \alpha_m^{(k)}) N(\mathbf{x}_i; \boldsymbol{\mu}_m, \Gamma_m) \pi_m \\ \forall i \in I^*, \xi_i^{(k)}(m) &= p(z_i = m | \mathbf{x}_i) \\ &\propto N(\mathbf{x}_i; \boldsymbol{\mu}_m, \Gamma_m) \pi_m, \end{aligned} \quad (10)$$

all normalized to one over M . The update rule is readily obtained and reads:

$$\alpha_m^{(k+1)} = \frac{\sum_{i \in I^+} \xi_i^{(k)}(m) + \alpha_m^{(k)} \sum_{i \in I^*} \xi_i^{(k)}(m)}{\sum_{i \in I} \xi_i^{(k)}(m)}. \quad (11)$$

A lighter approximate learning procedure can be obtained by freezing each hidden variable z_i to some sensible estimate \hat{z}_i independent from $\{\alpha_m\}$, e.g.,

$$\hat{z}_i = \arg \max_{m \in M} \pi_m N(\mathbf{x}_i; \boldsymbol{\mu}_m, \Gamma_m). \quad (12)$$

In this case, the EM estimate coincides with the maximum likelihood estimate on fully observed data points $\{y_i, \hat{z}_i\}_{i \in I^+ \cup I^-}$ only; that is, it is not iterative and reads

$$\alpha_m \approx \frac{\#\{i \in I^+ : \hat{z}_i = m\}}{\#\{i \in I^+ \cup I^- : \hat{z}_i = m\}}, \quad (13)$$

as in [17].

Note the GMM parameters could be jointly estimated with $\{\alpha_m\}$: we only need to replace responsibilities (5) by those in (10) with $\{\pi_m, \boldsymbol{\mu}_m, \Gamma_m\}$ set to its current value, while updates stay the same (Eqs. 6 and 11).

4. Experiments

We illustrate the semi-supervised GMM fitting and component selection algorithms for foreground/background discrimination in the following setup. Face images in heavy clutter are from the Oxford/Caltech database,¹ there is one frontal face under varying lighting conditions and scales in each image. $N = 200$ images are randomly chosen as the training set, and another 200 as the test set; $N_{\text{bg}} = 200$ background images from the same database (that contain no face) are used as the negative examples. Additional supervision, if any, is obtained by framing faces up manually in N_{rg} images out of N , which we will refer to as the “registered” images.

¹<http://www.robots.ox.ac.uk/~vgg/data/>

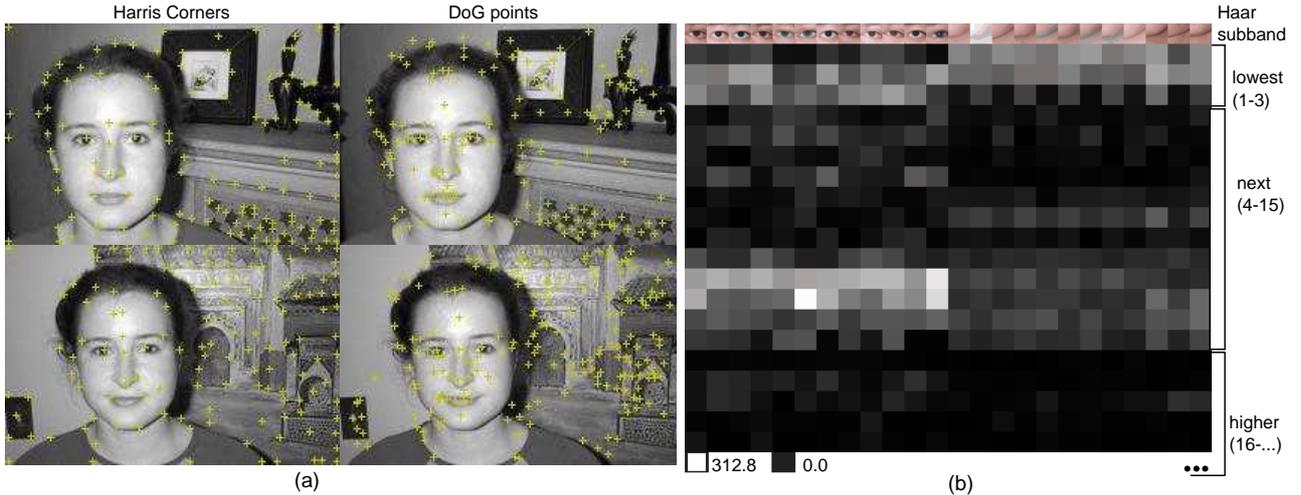


Figure 3: **Interest points and descriptors:** (a) Comparison of the Harris corner and DoG [18] interest point detectors. (b) The magnitudes of Haar wavelet coefficients for different 16×16 images patches from two clusters (“eye” and “chin”); the first fifteen are retained to build patch descriptors.

We first apply Harris corner detector on each image. As illustrated in Figure 3, this interest operator exhibits satisfactory sparsity and stability across objects in the same broad class. Haar wavelet decomposition is then performed on a 16×16 patch around the interest point, and we retain the lowest $d = 15$ AC coefficients as the descriptor for each patch. The wavelet descriptors are chosen for their ability to capture the perceptually salient features of the patch while ignoring the noise (Fig. 3). Compared to alternative descriptors, it is computationally lighter than PCA and is of much lower dimensionality than SIFT [18]. We have also empirically found that the lowest 15 coefficients are good to characterize the different patch components while generalizing well. They allow a 94% data reduction to be achieved from the raw 256-dimensional patch.

4.1. GMM fitting

We perform the semi-supervised GMM fit with side information (Section 2) on the wavelet-based descriptors. We use $|M^+| = 40$ components for the “object” class, initialized using the interest points lying within the bounding box in the registered images (I^+); and $|M^-| = 80$ components for the “background”, initialized using the negative patches (I^-) lying both in the N_{bg} negative images and outside the bounding boxes in the registered images (if any). The number of components of the Gaussian mixture is empirically chosen to ensure that the mixture captures well the various patch appearances within the object, and that it does not include similar components that would each be associated to a very small number (< 5) of training patches. The Gaussian components are initialized with K -means, and the EM algorithm usually halts within 10 iterations when the improvement in the log-likelihood falls below 0.1%.

To assess the quality of estimated models, we build the following simple classifier

$$\hat{y} = \begin{cases} +1 & \text{if } \frac{p(\mathbf{x}|y=+1)}{p(\mathbf{x}|y=-1)} > \text{threshold} \\ -1 & \text{otherwise,} \end{cases} \quad (14)$$

where label likelihoods are computed according to the model defined in Section 2; e.g.,

$$p(\mathbf{x}|y = +1) = p(\mathbf{x}|z \in M^+) = \frac{\sum_{m \in M^+} \pi_m N(\mathbf{x}; \boldsymbol{\mu}_m, \Gamma_m)}{\sum_{m \in M^+} \pi_m}. \quad (15)$$

Various thresholds are used on the likelihood ratio in order to produce the ROC curves in Figure 4, wherein the false positive rate (recall) vs. true positive rate are measured as follows.

$$\text{True positive rate} = \frac{\#\text{positive returns}}{\#\text{all positive samples}}$$

$$\text{False positive rate} = \frac{\#\text{false returns}}{\#\text{all negative samples}}$$

Note that, for the purpose of retrieving a few top-ranked patches that lies on the object, only the lower left portion of the ROC curve is useful. Also, one should bear in mind that these simple GMM-based classifiers are only meant to serve as the basis for a more sophisticated object-specific classifier where part layout is taken into account. Figure 4(a) shows that the constrained GMM fit yields superior results than ordinary GMM fit upon equivalent initialization: the first $|M^+|$ and the next $|M^-|$ components in both mixture models are initialized using the labeled subset I^+ and I^- , respectively. The only difference between the two fits is that the side information is not used in the second case.

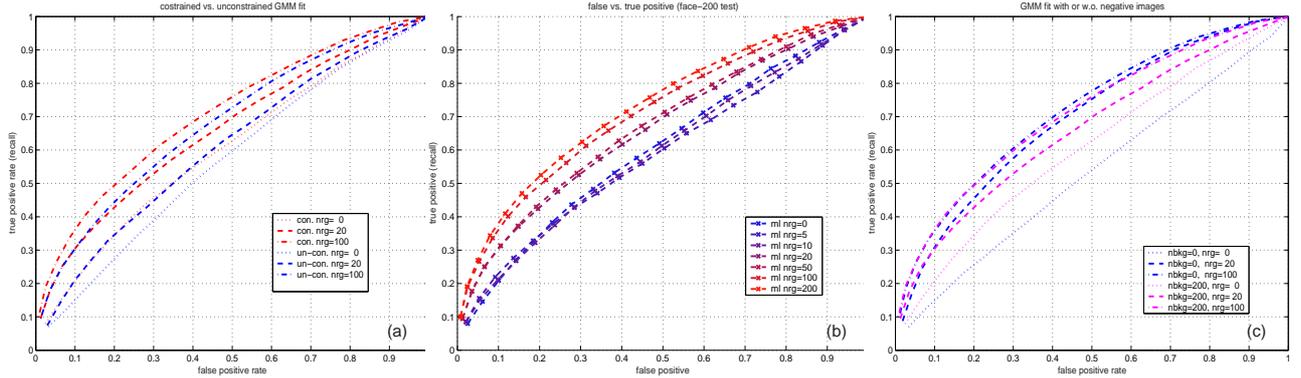


Figure 4: **Semi-supervised GMM fitting.** (a) constrained GMM fit with side information vs. ordinary GMM fit. (b) constrained GMM fit with different numbers N_{rg} of registered images. (c) constrained GMM fit with or without the set of $N_{\text{bg}} = 200$ background images.

Figure 4(b) illustrates that having more registered images does improve the accuracy, as we vary from a totally unsupervised approach with no object bounding boxes provided ($N_{\text{rg}} = 0$) to a supervised approach with registration information in every training image ($N_{\text{rg}} = N$, hence $I^* = \emptyset$). We have noticed that on the unsupervised end, having no registered images at all may do better than having very few of them (≤ 10), since it is difficult to reliably initialize the “object” mixture components on too few data points.

From the comparison in Fig. 4(c) on using or not the $N_{\text{bg}} = 200$ background images, we can see that using correlated negative examples alone (mostly clutter in indoor scenes from the same image as the objects) is better than using additional *un*-correlated negative images (random outdoor scenes). For instance, $N_{\text{rg}} = 20$ and $N_{\text{bg}} = 0$ outperforms $N_{\text{rg}} = 20$ and $N_{\text{bg}} = 200$, while the latter has a similar performance as $N_{\text{rg}} = 100$ and $N_{\text{bg}} = 20$.

4.2. Component selection

We select a subset of $|K|$ components with the largest conditional probability of being in the positive class (Section 3); i.e., choose $K \subset M$ such that $\forall k \in K, m \in M \setminus K, \alpha_k \geq \alpha_m$. We obtain the values of α_m with EM after constrained GMM fitting (Eq. 11) or by joint EM on all the parameters in the graphical model as explained at the end of Section 3. We compare the components chosen by these two (*GMM* and *Joint*) learning and selection methods to those obtained by the following baselines: (1) *Random*, where $|K|$ components are chosen at random from the $|M^+|$ components in the GMM obtained in Section 2, and the results are averaged over five independent runs; (2) *Tight*, where the $|K|$ components with the smallest 2-norm in the covariance matrix are chosen; (3) *MAP* [17], where α_m is approximated with the empirical percentage of positive patches among all maximum *a posteriori* patches as-

signed to component m in the training set (Eq. 13).

Similar to the previous subsection, we traverse the ROC curve of each selection algorithm with the likelihood ratio $\frac{P(\mathbf{x}|z \in K)}{P(\mathbf{x}|z \in M \setminus K)}$ of the $|K|$ -component subset over all the other components in the GMM. Figure 5(b) shows an example that selection strategies *GMM* or *joint* identify most of the foreground patches with as few as half (20) of the components in the positive mixture components M^+ . From Figure 5(c) we can see that *GMM*, *Joint*, or *MAP* outperform *Random* or *Tight*, while *GMM* is still better than the other two. The difference between *Joint* and *GMM* is seen in Figure 5(a): starting from identical initial conditions, the joint EM results in slightly *looser* mixture components and more skewed values of α_m . While the data likelihood values are very close in either case, the performance difference may be due to the fact that addressing two related but different learning objectives (*tight* fit vs. *relevant* components) at once may not be better than two separate steps addressing each objective in turn. Note the baseline *random* or *tight* caused sub-diagonal performance in the upper right part of the ROC curve because many of the salient foreground patches were erroneously thrown into the background class.

One alternative strategy for traversing the ROC curve is to return all the MAP patches under component $k \in K$ before returning any in $M \setminus K$. Compared to the use of likelihood ratios, this is a stricter criterium in that it makes each component in the subset K compete with the full model M^+ . Intuitively, the MAP patches are all treated equally in this respect, while those lying further away from the component centers are more likely to be the false positives. The component-wise ROC obtained with the *GMM* and *MAP* ranking is compared with that of the likelihood ratio, and this experiment is conducted on the UIUC car dataset [9, 17]. The training set contains 400 positive images of size 40×100 showing the side views of sedans, with approximately the same scale, as well as 400 negative

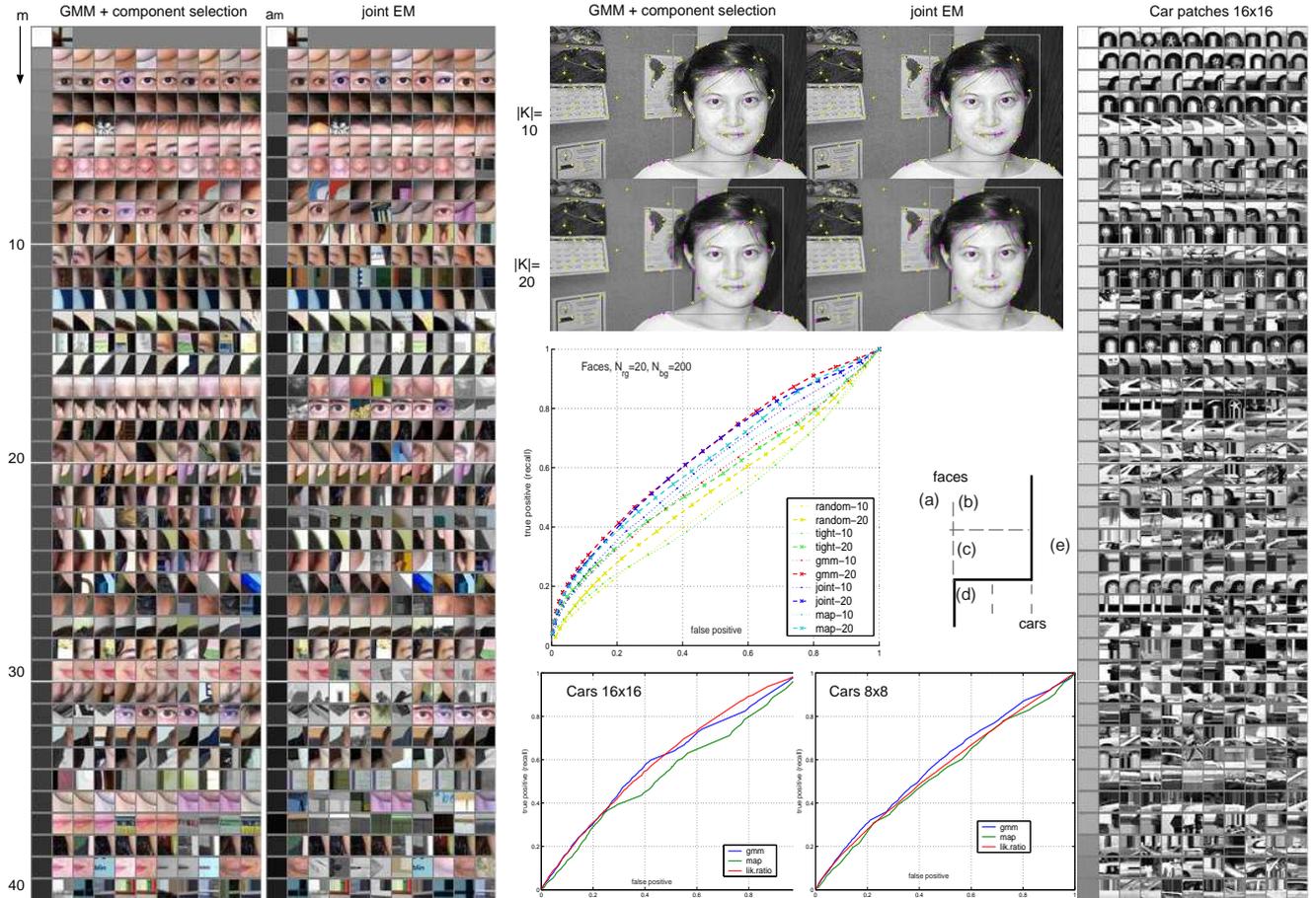


Figure 5: **Results for component selection.** (a) Ten most-likely patches from the test set in the 40 top mixture components where GMM fitting and component selection are performed separately or jointly. The joint model uses identical initialization as the constrained GMM fitting, and the resulting clusters are ranked in the order of decreasing α_m (shown in gray levels) learnt separately using Eq. 11. (b) Examples of foreground/background patches by the *GMM* and *joint* model, using $|K| = 10$ or 20 top-ranked mixture components. The foreground points are marked with magenta and the background points are in yellow. (c) The false positive vs. true positive rate of the top 10 or 20 mixture components under different selection strategies: *GMM*, *Joint*, *MAP*, *Random*, or *Tight* (see text for detail). (d) ROC curves for *GMM*, *MAP*, and *likelihood ratio* on the car dataset using 16×16 and 8×8 patches, respectively. (e) Ten most-likely patches for each component in the car dataset with 16×16 patches.

40×100 images of natural or urban scenes without cars. The testing set has 150 images that contain one or more sedans (of size 40×100 pixels) in natural environments. The location of cars are known in this case, hence leading to a supervised task where $|M^+|$ and $|M^-|$ components are learnt from the positive and negative images, respectively (Eq. (5)). From Figure 5(d) we can see that GMM performs comparably or even better than likelihood ratio while both outperform *MAP*.

5. Summary and future directions

In this paper, we investigated the scenario of semi-supervised learning of a Gaussian-mixture patch appearance model for object-class recognition. This scenario re-

quires much less effort in preparing the dataset than the supervised approaches [9, 17, 8, 10], yet involves much lighter computation than a totally unsupervised one [11, 15, 6]. We proposed an algorithm for semi-supervised fitting of appearance GMMs using side information, and addressed the problem of mixture component selection by estimating the conditional distributions of the patch class labels with respect to the mixture components. Experiments show that both the constrained GMM fit and the component selection improves the performance in finding patches lying on the objects of interest despite the presence of heavy clutter in the training images.

Appearance modeling is the first step towards a patch-based constellation object model than can be used for class-

specific detection. Moving toward such a global part-based appearance model poses the problem of the spatial relationship between the parts. This can be addressed with full covariance Gaussian models as in [6] and sequel, or with more tractable and yet very powerful tree-based layout models [3, 8, 16], with the star model based on relative positioning w.r.t. reference frame as a useful particular case.

Other open issues, left untouched in the literature, include the following aspects:

- Better capture of the available side-information: the noisy/incomplete labeling learning framework does not model all the available information. In particular, the soft labeling of training inputs in descriptor space is independent from one point to another, including for points arising from the same training image. This mechanism is thus unable to enforce the fact that one instance of the object of interest is present for sure in each positive image. As done with other types of side information [13, 12], new EM mechanisms should be designed to capture this particular type of side information.
- Earlier (and hopefully better) use of the negative examples in the learning process. Techniques that follow the generic synopsis presented in introduction usually make use of negative examples only at selection time. Indeed, the candidate pool of visual components is extracted from positive examples only, whereas negative examples are invoked later on to assess the discriminant power of these various components. The technique we propose goes one step further by using both corrupted positive images and negative images as soon as the extraction of the visual fragments. If this permits the jointly fitting of Gaussian mixtures to both populations while taking label noise into account, it also favors the emergence of mixture components with high discriminant power. At the moment, this assessment of discriminant power is only performed afterwards, at selection time. A more integrated learning scheme could however be considered, based on discriminant learning tools popular in speech analysis [1].
- Discriminant power and uniqueness: inner salient features of an object appearance are likely to correspond to specific parts of the object (e.g., the nose or the eye for faces). Hence, as opposed to texture features, the features of interest are likely to be *rare* in the images where one occurrence of the object class lies. This rarity makes them prone to be lost in each of the different learning steps. Mechanisms, in both the GMM fitting and the component selection, should be devised to protect, and favor, image fragments that appear consistently in the positive images but with low intra-image frequency.

Acknowledgement

The authors would like to thank: Microsoft Research Cambridge for the support and Prof. Andrew Blake for his insights during the inception of this work; Prof. Shih-Fu Chang for his support and Dongqing Zhang for helpful discussions in the later stages.

References

- [1] Ephraim, Y. and Rabiner, L. On the relations between modeling approaches for speech recognition. *IEEE Trans. Information Theory*, 36(2):372–380, 1990.
- [2] Daphne Koller and Mehran Sahami. Toward optimal feature selection. pages, 284–292, 1996.
- [3] Felzenszwalb, P. and Huttenlocher, D. Efficient matching of pictorial structures. In *Proc. Conf. Comp. Vision Pattern Rec.*, Hilton Head, SC, June 2000.
- [4] Nigam, K. and McCallum, A. and Thrun, S. and Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- [5] Weber, M. and Welling, M. and Perona, P. Towards automatic discovery of object categories. pages, II: 101–108, Hilton Head, SC, June 2000.
- [6] Weber, M. and Welling, M. and Perona, P. Unsupervised learning of models for recognition. pages, I: 18–32, Dublin, Ireland, June 2000.
- [7] Lawrence, N. and Scholkopf, B. Estimating a kernel Fisher discriminant in the presence of label noise. In *Proc. Int. Conf. Machine Learning*, San Francisco, CA, 2001.
- [8] Ioffe, S. and Forsyth, D. Mixtures of trees for object recognition. pages, II:180–185, Kauai, Hawaii, December 2001.
- [9] Agarwal, S. and Roth, D. Learning a sparse representation for object detection. pages, IV: 113–130, 2002.
- [10] Leibe, B and Schiele, B. Analyzing appearance and contour based methods for object categorization. In *Proc. Conf. Comp. Vision Pattern Rec.*, Madison, WI, June 2003.
- [11] Fei-Fei, L. and Fergus, R. and Perona, P. A Bayesian approach to unsupervised one-shot learning of object categories. pages, 1134–1141, Nice, France, October 2003.
- [12] Shental, N. and Zomet, A. and Hertz, T. and Weiss, Y. Computing Gaussian mixture models with EM using equivalence constraints. In *Int. Conf. Neural Inf. Proc. Systems*, Vancouver, Canada, December 2003.
- [13] Hertz, T. and Shental, N. and Bar-Hillel, A. and Weinshall, D. Enhancing image and video retrieval: Learning via equivalence constraints. pages, II: 668–674, Madison, WI, June 2003.
- [14] Leibe, B and Schiele, B. Interleaved object categorization and segmentation. In *Proc. British Machine Vision Conf.*, Norwich, UK, Sept. 2003.
- [15] Fergus, R. and Perona, P. and Zisserman, A. Object class recognition by unsupervised scale-invariant learning. pages, II: 264–271, 2003.
- [16] Vidal-Naquet, M. and Ullman, S. Object recognition with informative features and linear classification. pages, 281–288, Nice, France, October 2003.
- [17] Dork, G. and Schmid, C. Selection of scale-invariant parts for object class recognition. pages, 634–640, Nice, France, October 2003.
- [18] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 2004.