Incremental Bayesian Learning of Feature Points from Natural Images

Miika Toivanen Jouko Lampinen
Department of Biomedical Engineering and Computational Science
Helsinki University of Technology, Finland
{miika, jlampine}@lce.hut.fi

Abstract

Selecting automatically feature points of an object appearing in images is a difficult but vital task for learning the feature point based representation of the object model. In this work we present an incremental Bayesian model that learns the feature points of an object from natural unannotated images by matching the corresponding points. The training set is recursively expanded and the model parameters updated after matching each image. The set of nodes in the first image is matched in the second image, by sampling the unnormalized posterior distribution with particle filters. For each matched node the model assigns a probability for it to be associated with the object, and having matched few images, the nodes with low association probabilities are replaced with new ones to increase the number of the object nodes. A feature point based representation of the object model is formed from the matched corresponding points. In the tested images, the model matches the corresponding points better than the well-known Elastic Bunch Graph Matching batch method and gives promising results in recognizing learned object models in novel images.

1. Introduction

An algorithm that matches feature points of an object in an unseen query image typically needs to be trained on images with annotated feature points on the instances of the object. Two successful examples of such an approach are based on an elastic grid of features [15] and principal components of shape and appearance [3]. However, these models typically require a rather large training set, and manually annotating the training images is a tedious and time consuming task. Furthermore, they are batch algorithms where all the training images need to be available to train the model parameters. For a system learning the environment continuously, recursive online learning by updating the internal representations after each obtained data would be more feasible than batch learning, which requires memorizing and processing all the past data.

Some models learn the object representation in a weakly supervised manner. In constellation models, multiple candidate parts are extracted from the unannotated training images, and a joint probability density for the appearance and shape of these parts is utilized. The object parts are found by computing the maximum likelihood estimate with e.g. expectation maximization methods. Constellation models or variations of them are usually implemented as batch algorithms [14, 7, 5, 8, 11, 10] with the exception of [6], where the expectation maximization algorithm is adapted to allow for incremental learning. Other approaches to weakly supervised batch learning include a hierarchical visual cortex inspired algorithms [12] or segmentation algorithms [1].

In this work we present a weakly supervised online method that automatically learns the appearance and shape of the feature points of the common object appearing in grayscale images with no annotations or pre-segmentations, processing single image at a time. The learned feature points are the corresponding points in the images, matched during the process. To our knowledge, no other method exists that aims to perform this task. A presentation of the object, formed from the learned feature points, can be used in detecting instances of the object in unseen images. The method can also be used as a precursor to training other models that require annotated training images.

In the proposed algorithm, the matching result of each image is exploited in matching the next image, thereby incrementally expanding the training set. The approach is Bayesian; the likelihood is modeled as a Gabor filter based appearance and the prior as a Gaussian distribution for the shape of the corresponding points. These are combined into the (unnormalized) posterior distribution whose main mode is searched with particle filters. The set of reference nodes, positioned heuristically in the starting image, includes both object and background nodes. The probabilistic model allows for inferring the nodes that are to be associated with the object, and nodes that are not associated with it are replaced with new ones after having processed few images. The representation of the object thus improves as more images are processed.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

2. The method

We first introduce the model in brief (see also Figure 1). The model is presented images that share instances of a common object with arbitrary location, scale and orientation, one at a time. Also images containing no object can be included. A set of nodes, which can be considered as candidate corresponding points, is positioned in the starting image in a heuristic fashion, so that they are spread fairly evenly in the image. The unnormalized posterior distribution of the node set location in the second image is obtained as a product of the likelihood of the node set location, formed from the Gabor filter responses, and the prior distribution of the node set shape, modeled as a Gaussian distribution in the mean, scale and orientation free space. The node set is matched by sampling the posterior distribution with Population Monte Carlo algorithm. For each matched node the model assigns a posterior association probability, that is, the probability for a node to be associated with the common object.

For matching the nodes in the third image, the likelihoods are formed as mixtures whose kernels are evaluated at the node locations of the starting image and at the Monte Carlo estimate of the posterior mean of the second image. Also the variances of the Gaussian shape model and the association probabilities are updated. As more images are processed, the association probabilities of the object nodes (corresponding points) evolve to unity, and nodes with low association probabilities are eliminated and new nodes are laid on locations that more probably cover the object. The system needs no pre-defined number of images as the process could be stopped at anytime and a representation of the object formed by dropping the possibly remaining background nodes. The likelihood and prior used in this paper are somewhat similar to those used in [13], which presented a batch model that accurately matches occluded objects in query images with Sequential Monte Carlo methods.

2.1. Node selection

The nodes are selected in the starting image by first dividing it into small non-overlapping rectangular windows. In each window, the sum of the norm of a vector of the complex Gabor filter response magnitudes [15, 4] and the Gaussian distribution whose mean is the midpoint of the window, is maximized. As a result, the nodes are spread fairly evenly in the image with high information content (see left image of Figure 2). Basically, a node set with the shape of a regular grid could also be used, but selecting the reference nodes in (more) interesting locations alleviates the matching.

2.2. Likelihood - appearance model

The appearance of the nodes is modeled using a Gabor filter bank [4]. Denoting the magnitudes of the filter responses by a_j and phases by ϕ_j, a phase-sensitive similarity measure between a query image location x and the reference location x' is defined as

$$S_p(x, x') = \frac{\sum_j a_j(x) a_j(x') \cos(\phi_j(x) - \phi_j(x'))}{\sqrt{\sum_j a_j(x)^2 \sum_j a_j(x')^2}} , \quad (1)$$

where the index j runs over the jet coefficients [15]. By denoting the (Gabor filtered) query image with $\mathcal{I}$, (Gabor filtered) reference image with $\mathcal{I}'$, and the event "node is to be associated with the object" with V_i, the associative likelihood of node i is

$$p(\mathcal{I}|x_i, x_i', \mathcal{I}', V_i) = \exp(\beta S_p(x_i, x_i')) , \quad (2)$$

where β is a parameter controlling the steepness of the likelihood and is related to the Gaussian in-class variance of the Gabor responses. With a large enough value for β the nonlinear weighting of the similarities (2) can be such that the strongest mode of the posterior is at the correct location.

By denoting the non-association of node i with $\overline{V_i}$ we derive the likelihood by summing out the two association statuses:

$$\begin{aligned} p(\mathcal{I}|x_i, x_i', \mathcal{I}') &= p(\mathcal{I}, V_i|x_i, x_i', \mathcal{I}') + p(\mathcal{I}, \overline{V}_i|x_i, x_i', \mathcal{I}') \\ &= p(\mathcal{I}|x_i, x_i', \mathcal{I}', V_i)P(V_i) + p(\mathcal{I}|x_i, x_i', \mathcal{I}', \overline{V}_i)P(\overline{V}_i) . \end{aligned} \quad (3)$$

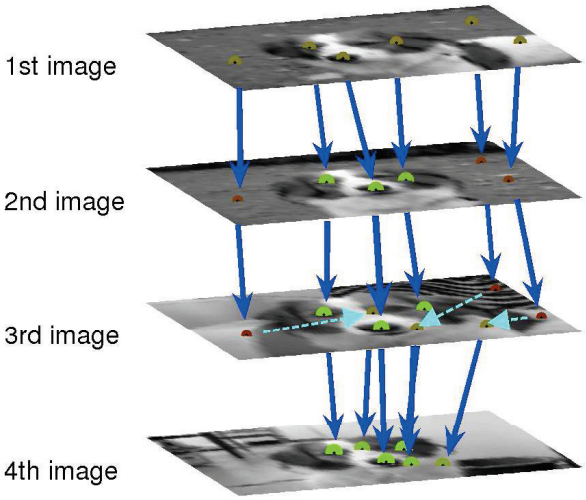

Figure 1. A schematic illustration of the model. In the first image, a simple algorithm selects the six nodes (yellow dots), which are matched in the subsequent images (solid blue arrows and dots). For each matched node, the model estimates how probably it associates with the object. The association probability is high for large green dots and low for small red dots. In the third image, three nodes with very low association probabilities are replaced by new nodes (dashed arrows) that are selected inside the convex hull of the current node set.

Figure 2. Left image: starting image and its node set. Middle image: second image in the sequence. Right image: zoomed second image, with white fields showing the modes of the likelihood of a node, marked with yellow circle in the starting image, and green contours illustrating the marginal prior distribution of the node (given scale 1.32, orientation -0.10°, and the midpoint, shown as red cross).

Hence the likelihood is a mixture of the two association possibilities weighted by their prior probabilities $P(V_i)$ and $P(\overline{V}_i) = 1 - P(V_i)$, which are considered to be independent of the node locations and the reference image. Furthermore, the probability that the query image point x_i is to be associated with the reference node point x'_i is simply the ratio of the associative likelihood and the mixture likelihood:

$$P(V_i|x_i, x'_i, \mathcal{I}, \mathcal{I}') = \frac{p(\mathcal{I}|x_i, x'_i, \mathcal{I}', V_i)P(V_i)}{p(\mathcal{I}|x_i, x'_i, V_i, \mathcal{I}')P(V_i) + p(\mathcal{I}|x_i, x'_i, \mathcal{I}', \overline{V_i})P(\overline{V_i})} . \quad (4)$$

The likelihood when the query node at x_i is not associated with the reference node at x'_i, that is, $p(\mathcal{I}|x_i, x'_i, \mathcal{I}', \overline{V_i})$, is the conditional marginal likelihood integrated over all other features except the one present at x'_i in $\mathcal{I}'$. For that, we use a constant, heuristically chosen value: $p(\mathcal{I}|x_i, x'_i, \mathcal{I}', \overline{V_i}) \equiv \kappa \; \forall i$. In [14, 7, 5] the background parts were modeled with a uniform density as well. In target tracking and surveillance, a constant value for no-detection likelihood is also typically used.

We assume independence between the filter responses at different locations, so that the total likelihood of the query node set is the product of the node likelihoods (2). Adding a positive constant κ with the associative likelihood ensures that the total likelihood is always non-zero also with non-associable nodes present.

2.3. Prior - shape model

The configuration of the nodes is controlled by setting a translated, scaled and rotated Gaussian prior distribution on the location of the nodes. The mean of the distribution is the reference shape - that is, the shape of the reference nodes - and the distribution is independent on the reference image:

$$p(\mathbf{x}|\mathbf{x}', s, \varphi) = \mathcal{N}(\mathbf{x}|m_{\mathbf{x}} + R(\varphi)s(\mathbf{x}' - E[\mathbf{x}']),\; s^2\mathbf{R}(\varphi)C) , \quad (5)$$

where $E[\cdot]$ denotes mean value, C is a diagonal covariance matrix, $m_{\mathbf{x}}$ is the midpoint of the query set, s is the scale (of the query shape compared to the reference shape), $R(\varphi)$ is the rotation matrix with angle φ and $\mathbf{R}(\varphi)$ scales the vertical and horizontal components of the covariance matrix.

2.4. Posterior distributions

The likelihood and prior parts are combined into the unnormalized joint posterior distribution:

$$p(\mathbf{x}, s, \varphi|\mathbf{x}', \mathcal{I}', \mathcal{I}) \sim p(\mathcal{I}|\mathbf{x}, \mathbf{x}', \mathcal{I}', s, \varphi)p(\mathbf{x}|\mathbf{x}', s, \varphi)p(s, \varphi) , \quad (6)$$

where $p(s, \varphi) \sim 1$ is the non-informative prior distribution for scale and rotation. Likelihood depends on s and φ if different Gabor filter jets are used for different scales and rotations. Also the posterior association probability of node i is obtained, by integrating over the posterior distribution:

$$P(V_i|\mathbf{x}', \mathcal{I}, \mathcal{I}') = \int p(V_i, \mathbf{x}, s, \varphi|\mathbf{x}', \mathcal{I}, \mathcal{I}')d\mathbf{x}\; ds\; d\varphi$$
$$= \int p(V_i|\mathbf{x}, s, \varphi, \mathbf{x}', \mathcal{I}, \mathcal{I}')p(\mathbf{x}, s, \varphi|\mathbf{x}', \mathcal{I}, \mathcal{I}')d\mathbf{x}\; ds\; d\varphi . \quad (7)$$

In Figure 2, an example of matching a query image is presented. The combination of the prior and likelihood results in the marginal posterior (not shown) whose mode is approximately at the true location.

2.5. Sampling

For representing the distributions and computing the necessary integrals, we sample the posterior distributions with Population Monte Carlo (PMC) method [2]. In PMC the proposal distributions may differ between particles, so we let the variance of the proposal distributions decrease for associable nodes. PMC algorithm may be arbitrarily initialized without violating the convergence theorems; thus we have implemented the global move step of Elastic Bunch Graph Matching (EBGM) method [15] as heuristic means to get probable node locations. The first round of our PMC

1. Initialize

- Pre-compute a large filter bank consisting of five scales $\sqrt{2}\pi/\{2,4,8,16,32\}$ and 12 orientations $\{0,\pi/6,...,11\pi/6\}$
- Initialize the particles by scanning (with interval of few pixels) the query image with a rigid reference node set using five scales $s = 2^{\{-0.8,-0.4,0,0.4,0.8\}}$ and three angles $\varphi = \{-10°, 0°, +10°\}$. For each match (there are typically thousands of them), compute phase-sensitive similarities (1) and phase-insensitive similarities (by dropping the phase term off), using Gabor jets consisting of three frequencies and six orientations, by interpolating the filter responses with four neighboring filters of the pre-computed filter bank. Give each match a score, defined as the mean value of the most similar 25 % of the nodes, using the phase-insensitive similarity. Take 150 best scores (matches) and initialize the particles with the corresponding locations ($\mathbf{x}^0$), scales (s^0), orientations (φ^0) and phase-sensitive similarities.
- Set $t = 1$.

2. Sample each particle from the proposal distributions

- Compute the posterior association probabilities with (4) and set $T_i = P(V_i|x_i, x_i', \mathcal{I}, \mathcal{I}')$
- Set $q(\log[s]|s^{t-1}) = \mathcal{N}(\log[s^{t-1}], \sigma^2_{\log[s]})$, where $\sigma_{\log[s]} = \log[1.06]/(1 + E[T])$
- Set $q(\varphi|\varphi^{t-1}) = \mathcal{N}(\varphi^{t-1}, \sigma^2_\varphi)$, where $\sigma_\varphi = 2/(1 + E[T])$ (in radians)
- Sample scale and angle: $\log[s^*] \sim q(\log[s]|s^{t-1})$, $\varphi^* \sim q(\varphi|\varphi^{t-1})$
- For $i = 1...N_{\text{nodes}}$
  - Set $q(x_i|\mathbf{x}^{t-1}, s^*, \varphi^*) = \mathcal{N}(x_i|m_\mathbf{x} + R(\varphi^*)s^*(x_i^{t-1} - E[\mathbf{x}^{t-1}]),\ C_i)$, where $m_\mathbf{x}$ is the $\mathbf{T}$ weighted midpoint of $\mathbf{x}^{t-1}$ and $C_i = \begin{pmatrix} \sigma^2_{x_i} & 0 \\ 0 & \sigma^2_{y_i} \end{pmatrix}$, where $\sigma_{x_i} = s^*(\sigma_{x_i,prior})^{\sqrt{2-T_i}}$ and $\sigma_{y_i} = s^*(\sigma_{y_i,prior})^{\sqrt{2-T_i}}$
  - Set $q(\mathcal{I}|x_i, x_i', \mathcal{I}', s^*, \varphi^*) = P(V_i)\tilde{L}_i(x) + (1 - P(V_i))\kappa$, where $P(V_i)$ is the prior association probability and $\tilde{L}_i(x) = \exp(\beta S_p)$, where S_p is the pre-computed phase-sensitive similarity with scale and orientation closest to s^* and φ^*
  - Set $\alpha_i = \sum_{x_i \in A} q(x_i|\mathbf{x}^{t-1}, s^*, \varphi^*)q(\mathcal{I}|x_i, x_i', \mathcal{I}', s^*, \varphi^*)$, where the size of the search window A is $[3\ \ 3]C_i$ pixels around the prior mean
  - Set $q(x_i|\mathbf{x}^{t-1}, s^*, \varphi^*, \mathcal{I}, \mathcal{I}') = \frac{1}{\alpha_i} q(x_i|\mathbf{x}^{t-1}, s^*, \varphi^*)q(\mathcal{I}|x_i, x_i', \mathcal{I}', s^*, \varphi^*)$
  - Sample location (numerically): $x_i^* \sim q(x_i|\mathbf{x}^{t-1}, s^*, \varphi^*, \mathcal{I}, \mathcal{I}')$

3. Compute the particle weights for each particle

- Compute $w = \frac{p(\mathcal{I}|\mathbf{x}^*, \mathbf{x}', \mathcal{I}', s^*, \varphi^*)p(\mathbf{x}^*|\mathbf{x}', s^*, \varphi^*)}{[\prod_{i=1}^{N_{\text{nodes}}} q(x_i^*|\mathbf{x}^{t-1}, s^*, \varphi^*, \mathcal{I}, \mathcal{I}')]q(\log[s^*]|s^{t-1})q(\varphi^*|\varphi)}$, where the value of the likelihood has been computed by interpolating between four neighboring filter responses, according to s^* and φ^*.
- If $t = 1$, set the particle weight to $w^{1/3}$, otherwise set the particle weight to w

4. Resample and move the particles

- Resample the particles with replacement according to the particle weights using deterministic resampling
- Set the components of the particles $\{\mathbf{x}^t, s^t, \varphi^t\}$ to the resampled values
- Move each particle with Langevin Monte Carlo sampling step
- If $t > 1$ and the variance of $\mathbf{x}^t$ is below a threshold, or if $t = t_{\text{max}}$, stop. Otherwise, set t = t+1 and go to step 2

Algorithm 1: Population Monte Carlo implementation. Particle indexing has been dropped for clarity, hence steps 2 and 3 are performed individually for each particle.

implementation is special as the particle weight is taken to be $w^{1/3}$ to prevent degeneration - theoretically this does not produce samples from the posterior but acts as an initialization for the second round. The PMC implementation is presented in a greater detail in Algorithm 1, where indexing for different particles is omitted for the sake of clarity. The number of best scores in the initialization step, and hence the number of particles, was chosen to be 150 to have a sensible computational time. The hypothesis is that the correct approximative match is included in these 150 best scores.

2.6. Incremental processing of the image set

Next we describe how the images are processed one by one (see also Algorithm 2). After being matched with the starting image, the second image turns into a reference image for the third image. The similarities and the likelihoods are mixtures of two kernels whose Gabor coefficients are evaluated at the nodes of the starting image and at the mean

1. Actions for the starting image:
   (a) Gabor transform the starting image
   (b) Select the nodes in the starting image as explained in the text and store the Gabor responses at each node
2. Match the next image
   (a) Gabor transform the next image
   (b) Match the node set with PMC (**Algorithm** 1)
3. Update the parameter values
   (a) Estimate the posterior association probabilities of the nodes with (7)
   (b) Compute the magnitudes of the Gabor responses with (8), and likewise the phases, for the mixture likelihood
   (c) Update the prior association probabilities of the nodes with (10)
   (d) Update the prior node variance with (11)
4. Modify the node set
   (a) For each node with $P(V_i) < P_{th}$, reposition the node as explained in the text
   (b) Update the reference shape with (12) and (14)
5. Go to step 2

Algorithm 2: Incremental algorithm for matching the corresponding points

Gabor responses of the second image, weighted by the posterior association probabilities. In general, the average Gabor magnitudes of node i of kernel (image) $k > 1$ are computed as

$$(\mathbf{a}_i)_k = \frac{\int \mathbf{a}(x_i, s, \varphi)\Psi(\mathbf{x}, s, \varphi)\, d\mathbf{x}\, ds\, d\varphi}{\int \Psi(\mathbf{x}, s, \varphi)\, d\mathbf{x}\, ds\, d\varphi}\,, \qquad (8)$$

where

$$\Psi(\mathbf{x}, s, \varphi) = P(V_i|x_i, x_i', \mathcal{I}, \mathcal{I}', s, \varphi)_k\, p(\mathbf{x}, s, \varphi|\mathbf{x}', \mathcal{I}', \mathcal{I})_k\,, \qquad (9)$$

where the subscript k refers to the kth kernel, and likewise for phase $(\boldsymbol{\phi}_i)_k$. The prior association probabilities for the next image $K + 1$ are updated as

$$P(V_i)_{K+1} = \frac{K-1}{K} P(V_i)_K + \frac{1}{K} P(V_i|\mathbf{x}', \mathcal{I}, \mathcal{I}')_K\,. \qquad (10)$$

Initially, all the prior association probabilities are half.

Having an inverse-gamma prior distribution on the node variance of the shape model leads to an analytical form for the conditional posterior distribution [9]. Using the mode of the posterior distribution and setting the observed variance in horizontal (x) direction v_{x_i} to the point estimate at the posterior mean, the prior node variance is updated as

$$(\sigma_{x_i,prior})^2_{K+1} = \frac{\nu_0 + K}{\nu_0 + K + 2}\, \frac{\nu_0\sigma_0^2 + K v_{x_i}}{\nu_0 + K}\,, \qquad (11)$$

where ν_0 and σ_0 are the hyperparameters of the prior distribution for σ_{x_i}, and similarly for y direction. Updating the node variance allows the system to learn the level of rigidity of the object being matched.

With many object nodes and few background nodes the representation of the object model improves, making it easier to match the node set in the next image. Therefore, the node set is modified by replacing the nodes whose prior association probability goes below a threshold with new nodes so that the total number of the nodes stays constant. Because in the starting image the nodes are distributed over the whole image, the new nodes are positioned inside the convex hull of the existing nodes so that the object is approached 'outside' by gradually shrinking the area of the node set. For non-convex objects, this approach leads to trial end error type procedure, as a new node may not get correspondence in the subsequent images and is again moved to a new position. The new nodes are positioned by maximizing the information content while penalizing the vicinity of other nodes, as in selecting the nodes in the starting image. For these new nodes, the prior associations are set to half, the previous posterior association probabilities are set to zero, the prior node variances are set to the initial values and the previous kernels in their likelihood mixtures are removed. To prevent some image to have too much control over the new node locations, an upper limit can be set on the number of the modifications.

The reference shape for the next image is a mixture of the previous mean free shapes, weighted by their mean posterior association probabilities:

$$\mathbf{x}' = \frac{\sum_{k=1}^{K} E_i[P(V_i|\mathbf{x}', \mathcal{I}, \mathcal{I}')_k](\mathbf{x} - M_{\mathbf{x}})_k}{\sum_{k=1}^{K} E_i[P(V_i|\mathbf{x}', \mathcal{I}, \mathcal{I}')_k]} \,. \quad (12)$$

For this, the midpoints of the matched images $M_{\mathbf{x}}$ are recomputed to minimize the weighted variance of the x and y components of the distances of the nodes to the midpoints:

$$E_x = \sum_i \sum_k V_i^k \left((\Delta x)_i^k - \frac{\sum_{k'} V_i^{k'} (\Delta x)_i^{k'}}{\sum_{k'} V_i^{k'}} \right)^2 / \sum_{k'} V_i^{k'} \,, \quad (13)$$

where $V_i^k \equiv P(V_i|\mathbf{x}', \mathcal{I}, \mathcal{I}')_k$ is the posterior association probability of node i in image k, and $(\Delta x)_i^k \equiv x_i^k - m_x^k$ is the coordinate of x_i^k in m_x^k mean coordinate system, and likewise for the y component. By denoting with V a matrix with elements V_i^k, with X a matrix with elements x_i^k, and by differentiating E_x w.r.t. $M_x = \{m_x^1, m_x^2, ..., m_x^K\}$ and setting the result to zero, we get

$$M_X = (Z)^{-1} B_x \,, \quad \text{where} \quad (14)$$

$$Z = \mathrm{d}\left[\left(\frac{V}{V I_{K\times K}} \right)^T I_{N\times 1} \right] - 2 \left(\frac{V}{(V I_{K\times K})^2} \right)^T V + \left(\frac{V}{(V I_{K\times K})^3} \right)^T \mathrm{d}[V I_{K\times 1}] V \quad (15)$$

and

$$B_x = \mathrm{D}\left[\left(\frac{V}{V I_{K\times K}} \right)^T X \right] - 2 \left(\frac{V}{(V I_{K\times K})^2} \right)^T \mathrm{D}[V X^T] + \left(\frac{V}{(V I_{K\times K})^3} \right)^T \mathrm{d}[V I_{K\times 1}] \mathrm{D}[V X^T] \,, \quad (16)$$

where $\mathrm{d}[a]$ means constructing a diagonal matrix of vector a, $\mathrm{D}[A]$ means taking the main diagonal of matrix A and I is a matrix of ones. The divisions and involutions are made element-wise.

3. Experiments

In this section we give both qualitative and quantitative results using the following parameter values: likelihood steepness was $\beta = 50$, hyperparameters of the shape model were $\nu_0 = 5$ and $\sigma_0^2 = 10$, threshold for removing the nodes was $P_{th} = 0.24$, maximum number of PMC iterations was 5, the number of nodes was $N_{\text{nodes}} = 30$ and the maximum number of nodes to be modified in a singly image was set to 6. These parameter values were found by testing and were used throughout the experiments. The method is, however, not very sensible for these parameters - the value for P_{th} was chosen to prevent the node set to be modified yet in the second image. The image databases we used include images of traffic signs and a dog taken with a digital camera, as well as the face and leaves databases of the widely used Caltech images [7] (see Figure 3). The average CPU time to process an image is about one minute in an unoptimized Matlab environment and up-to-date desktop computer, being proportional to the number of nodes. Due to the parallel nature of the PMC sampling, the computation time could be reduced substantially with parallel computing.

Figure 3. Examples of the images used in experiments. From top to bottom: SIGNS, DOGS, Caltech FACES and Caltech LEAVES.

In Figure 4, examples on matching sequences of 10 SIGNS and DOGS images are illustrated. Object nodes are accurately matched, and the number of the object nodes increases along the sequence.

The Euclidean matching errors for SIGNS and DOGS sequences were measured by manually picking in the next matched image the corresponding points of the nodes whose prior association probability was higher than 0.9 in the last image (average number of such nodes was five), and taking the average over the images and nodes. For comparison, we

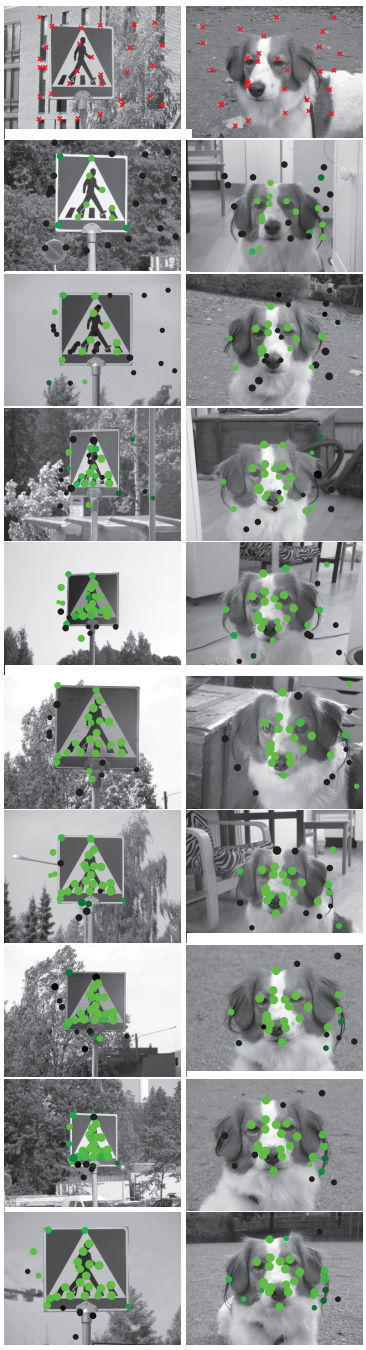

Figure 4. Illustrative examples of the matching results. The locations of the dots reveal the Monte Carlo estimate of the posterior mean. The size of the dots is proportional to the prior association probability and the green value to the posterior association probability. The image sequence proceeds from top to bottom. The green dots form the representation of the object.

tested the same images with EBGM method [15], by first annotating the objects (6 annotations in SIGNS images and 12 in DOGS images, in fiducial locations such as corners) and defining the edges as Delaunay triangulations of the annotated nodes. The median errors of 100 image sequences of 10 images are tabulated in Table 1 for both methods. Our method outperforms EBGM method although it processes the images recursively and uses no object annotations.

Recognition tests were performed with Caltech images. The object model was first learned (recursively) from object image sequences. For FACES database, the system was then given five background images (from the Caltech background images database) and five object images, the learned object model was matched to the images and the number of correct and false detections were counted, with detection measured if the mean posterior node association probability exceeded a threshold. ROC curves are formed by varying the threshold. The numerical figures are shown in Table 2. Having more training images, dropping the background nodes before matching the query image, as well as using a common value for κ seem to improve the results. The incremental constellation model of [6] was tested with different databases (Caltech 101) and cannot thus be compared with these results; however, it is worth while mentioning that the average areas under the ROC curves they report are 71 % with one training image and 75 % with three training images. LEAVES database was tested by first extracting four subsets from it according to the type of the leaves, learning each leave model with five images, and matching the learned models to three representatives of each of the four leave types. The query image with the largest mean posterior association probability was inferred to contain the learned object. The confusion matrix, averaged over 75 such processes, is depicted in Table 3.

Table 1. Median point-to-point errors of 100 image sequences for our method and EBGM method, with the number of training images in parentheses, in pixels. The errors in our method describe the average match over the sequence with 1,...,9 training images.

	OUR	EBGM(1)	EBGM(5)	EBGM(9)
SIGNS	1.7	16.9	11.6	6.3
DOGS	4.5	7.3	3.4	3.4

Table 3. Confusion table for the Caltech LEAVES. For instance, 95% of type 4 query images were correctly classified whereas 5% were misclassified as other types.

	Type 1	Type 2	Type 3	Type 4
Model 1	84	14	0	1
Model 2	14	81	1	1
Model 3	0	1	92	3
Model 4	1	4	8	95

Table 2. ROC error figures, averaged over 100 learned models. AURx refers to the area under ROC curve and EERx refers to the equal error rate, using x training images. The two left columns indicate, whether the background nodes (with $P(V_i) < 0.5$) were dropped before matching the query image and whether the same κ (such value that maximizes the figure in question) was used for all the association probabilities. The EER figures of other reported results are 96.4 [7] and 98.2 [12], which are batch methods and used training sets with 225 images.

Drop BG	Same κ	AUR1	AUR3	AUR6	AUR9	EER1	EER3	EER6	EER9
		65.8	73.2	83.6	87.5	63.0	67.2	76.5	79.8
X		65.8	67.2	82.7	90.1	63.0	63.2	75.4	84.4
	X	76.1	84.0	87.8	88.4	71.2	77.4	80.7	81.0
X	X	76.1	70.9	82.5	89.7	71.2	66.2	75.8	84.8

4. Conclusions

We have presented a translation-, scale- and rotation-invariant incremental method for finding a set of representative feature points for an unknown object from a series of images with no manually selected feature locations. The method is based on Bayesian learning; a set of nodes is selected in the starting image, and for each new image, the posterior distributions of the node locations, and the probability of a node being associated with the object are estimated. The appearance and shape models are updated with information from the new image. The mixture likelihood allows for large variations in the node appearance in the corresponding location. The experiments suggest that our method is able to accurately match the corresponding points and to learn the object representations that make it possible to detect and recognize the learned objects in new images.

Although all the images used in the experiments contained an instance of the object, the proposed model is not very sensitive to the number of background images included in the sequence. Also, the initialization of the nodes in the starting image is of minor relevance, whilst the ordering of the images has a bigger role - an outlier late in the sequence causes less harm than in the beginning of the sequence. The current shape model is restricted to objects or classes of objects with similar shape, such as human faces. Adding additional metric transformations, or generic affine transformation, would be straightforward, allowing for association of wider class of objects in one model. For classifying objects with highly varying shapes in a same category (for example a class "chair") could be implemented as a higher level clustering or supervised classification process, where the low level models recognize the basic shapes.

References

[1] N. Ahuja and S. Todorovic. Learning the taxonomy and models of categories present in arbitrary images. In *Proc. ICCV*, pages 1–8, 2007.

[2] O. Cappe, A. Guillin, J. Marin, and C. Robert. Population Monte Carlo. *J. Comput. Graph. Statist*, 13(4):907–929, 2004.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001.

[4] J. Daugman. Complete discrete 2-d Gabor transforms by neural networks for imageanalysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, 36(7):1169–1179, 1988.

[5] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. ICCV*, pages 1134–1141, 2003.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, pages 264–271, 2003.

[8] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and complete recognition. In *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 443–461. Springer, 2007.

[9] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *Proc. BMVC*, volume 2, pages 959–968, 2004.

[11] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Proc. CVPR*, pages 26–36, 2006.

[12] T. Serre, L. Wolf, S. Bileschi, and M. Riesenhuber. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE TPAMI*, 29(3):411–426, 2007.

[13] T. Tamminen and J. Lampinen. Sequential Monte Carlo for Bayesian matching of objects with occlusions. *IEEE TPAMI*, 28:930–941, 2006.

[14] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, pages 18–32, 2000.

[15] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE TPAMI*, 19:775–779, 1997.