International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:12, 2009

Bayesian Online Learning of Corresponding Points of Objects with Sequential Monte Carlo

Miika Toivanen Jouko Lampinen
Department of Biomedical Engineering and Computational Science
Helsinki University of Technology, P.O.Box 9203, FI-02015, TKK, Finland
{miika, jlampine}@lce.hut.fi

Abstract—This paper presents an online method that learns the corresponding points of an object from un-annotated grayscale images containing instances of the object. In the first image being processed, an ensemble of node points is automatically selected which is matched in the subsequent images. A Bayesian posterior distribution for the locations of the nodes in the images is formed. The likelihood is formed from Gabor responses and the prior assumes the mean shape of the node ensemble to be similar in a translation and scale free space. An association model is applied for separating the object nodes and background nodes. The posterior distribution is sampled with Sequential Monte Carlo method. The matched object nodes are inferred to be the corresponding points of the object instances. The results show that our system matches the object nodes as accurately as other methods that train the model with annotated training images.

Keywords—Bayesian modeling, Gabor filters, Online learning, Sequential Monte Carlo.

I. Introduction

Feature based object matching methods require typically a large set of training images with manual annotations on the corresponding points [1]–[4]. To learn the parameters of the models, the images are processed simultaneously. The computational requirements of such batch learning become huge with large training sets. Also, annotating images by hand is a time consuming job.

In this contribution a method is proposed for solving both the above-mentioned problems. The method finds the corresponding points in an image sequence without making use of annotated feature locations, by matching the images incrementally, one by one. In the algorithm, a set of node points is positioned in the starting image of the sequence in a somewhat regular grid. In the second image, a similarly shaped node set having as much correspondence with the first image as possible is located. This is realized within the Bayesian framework; the Gabor filter response based local appearance of the nodes and the shape of the node set are combined into a posterior probability distribution of the node locations, which is sampled using the Sequential Monte Carlo (SMC) sampling method to find the most probable locations for the nodes. SMC methods are usually applied in dynamic problems where new data arrive online [5]. However, they can also be used in our static setting by sampling the nodes from conditional posterior distributions, conditioned on the already sampled nodes. With some improvements on the basic SMC scheme, it turns out to be an efficient method for locating the main mode of the posterior distribution. The matching results of the second image is exploited in matching the next image, and hence the training set expands recursively as more images are processed.

The node set probably contains object nodes and nodes located in the background. Bayesian framework offers a natural way for inferring which of the matched nodes should be associated with the object and which not. As more images are processed, the uncertainty in associating the nodes decreases. A feature point based representation of the object can be formed from the object nodes by dropping the background nodes from the node set. Unlike in batch methods, the number of training images to be used for the representation can be concluded online. In the experiments the presented model demonstrates its promise by achieving a matching accuracy for the object nodes comparable to methods that learn the object model from an annotated training set. The proposed method can also be used as a precursor to build a training set for methods that require annotations on the feature points.

Slightly related with the presented model are the part based models which also utilize a joint probability density for the appearance and shape of the parts [6]–[11]. These batch methods learn a representation of an object from un-annotated training images containing instances of the object and use the representation to classify test images. The learning is usually implemented by selecting candidate parts from the training images and applying the expectation maximization method to learn the model parameters and finding the similar object parts. Some batch methods classify test images by segmenting the un-annotated training set [12]–[15]. Methods that learn the object model online are very rare; to our best knowledge, the only such method is the incremental part based method of Fei-Fei et al. [16]. Hence, this paper has an important contribution on the field.

II. The Bayesian model

Before going into details of the model, it is described here in brief. The method is given images that all contain an instance of one common object, but no further information is given. The task is to find corresponding points of the object in the images, which are processed one at the time, in random order. A set of candidate corresponding points, called nodes, are placed in the starting image of the sequence using a simple automatic procedure, which aims to place the nodes in 'interesting'

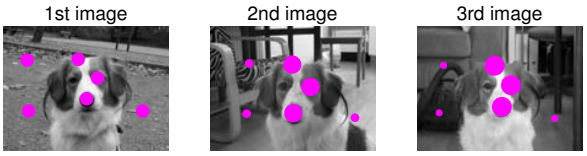

Fig. 1. An artificial example of the evaluation of a node set in a sequence of three images. The dots mark the mean of the posterior distribution of the node locations, and the size of the dots reveal the association probability (large dots are associated with the object).

locations, approximately evenly in the image. In the following images, nodes with similar appearance are searched while keeping the shape of the node set approximately the same. For this, a Bayesian approach is adopted: the likelihood function, which is based on Gabor filter response based similarity, and the prior part, which assumes the mean shape of the node set to be similar in each image, are combined into the posterior distribution for the location of the node set. The likelihood is a mixture distribution, whose kernels are evaluated at each matched image, thereby allowing multiple appearances to be modeled for the nodes. In a successful case, the most probable node combination has the corresponding nodes at correct locations. The posterior distribution is sampled with SMC methods. From the SMC samples, the posterior mean and some integrals, needed in the computations, can be estimated. Each matched node is assigned an association probability, which reveals, how probably the node is associated with the common object. For nodes with nearly constant appearance, this probability increases to unity along the sequence. If a node is located on an image detail whose appearance has many modes, or which is occluded in some images, the increasing of its association probability is slower. The image details of the background nodes typically differ in each image, so their association probability tends to zero along the sequence. A representation of the object can be formed anytime as soon as it becomes clear, which of the nodes are object nodes. In figure 1, an illustrative example of matching incrementally a set of six nodes in a sequence of three images is shown.

A. Node selection

A simple automatism is used to select the nodes in the first image. The image is divided into small non-overlapping rectangular windows. In each window, a pixel is chosen which maximizes the sum of the magnitudes of the complex Gabor filter responses [1], [17]. An example of the selected nodes in the starting image is illustrated in figure 2. The gaps between the windows prevent many nodes from being selected at the same image detail (at neighboring pixels), which may happen if the windows touch each other. The shape of the selected node set is considered as the reference shape for the following images. It should be noted that the rectangular windows are used only for placing the nodes in the starting image and do not set any limits for where to search for the nodes in the upcoming images.

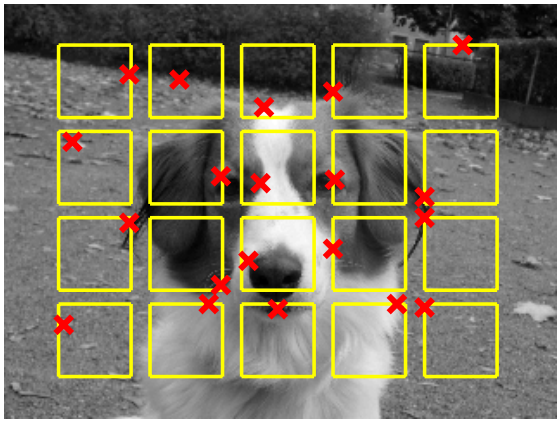

Fig. 2. An example of the automatically selected nodes in the starting image of the sequence.

B. Posterior distribution of the node locations

Let $\mathtt{I}_t$ denote the current test image, being image number $t > 1$ in the sequence. If $\mathbf{x}_t$ denotes the location of the node set in image $\mathtt{I}_t$, a posterior distribution is

$$p(\mathbf{x}_t|\mathtt{I}_{1:t},\mathbf{x}_1) \sim p(\mathtt{I}_t|\mathbf{x}_t,\mathtt{I}_{1:t-1})p(\mathbf{x}_t|\mathbf{x}_1) \ , \qquad (1)$$

where $p(\mathtt{I}_t|\mathbf{x}_t,\mathtt{I}_{1:t-1})$ is the likelihood - independent of $\mathbf{x}_1$ - and $p(\mathbf{x}_t|\mathbf{x}_1)$ is the prior distribution for the location of the node set.

C. Likelihood - appearance model

First, the joint likelihood of the node set is taken to be a product of the independent node likelihoods ('naive Bayes' assumption):

$$p(\mathtt{I}_t|\mathbf{x}_t,\mathtt{I}_{1:t-1}) = \prod_{n=1}^{d} p(\mathtt{I}_t|x_t^n,\mathtt{I}_{1:t-1}) \ , \qquad (2)$$

where n indexes the nodes of the node set, whose size is d.

Let us next drop the node index n and concentrate on an independent node likelihood. At this stage, the concept of node association is adopted. Each node of the node set is or is not associated with the common object. These cases are denoted with A and $\overline{A}$, respectively. The level of the association is naturally revealed with probabilities. The likelihood of an individual node of the node set is yielded by summing out the unknown association status:

$$\begin{aligned} p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1}) = {} & p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},A_t)P(A_t|\mathtt{I}_{1:t-1}) \\ & + p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},\overline{A}_t)P(\overline{A}_t|\mathtt{I}_{1:t-1}) \ , \end{aligned} \qquad (3)$$

where $P(A_t|\mathtt{I}_{1:t-1})$ is a prior association probability for the node, which is independent on the node location x_t. This is computed recursively as

$$P(A_t|\mathtt{I}_{1:t-1}) = \frac{1}{t-1}P(A_{t-1}|\mathtt{I}_{1:t-1}) + \frac{t-2}{t-1}P(A_{t-1}|\mathtt{I}_{1:t-2}) \ , \qquad (4)$$

that is, it is a weighted mixture of posterior and prior association probabilities of the previous image. To compute

the posterior probability for associating the node in image t, $P(A_t|\mathtt{I}_{1:t})$, the posterior association probability of the node at image location x_t is first presented, which is given by Bayes' formula:

$$p(A_t|x_t,\mathtt{I}_{1:t}) = \frac{p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},A_t)P(A_t|x_t,\mathtt{I}_{1:t-1})}{p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1})} \ . \quad (5)$$

This is integrated over the posterior distribution of $\mathbf{x}_t$ to get the posterior association probability in image t,

$$p(A_t|\mathtt{I}_{1:t}) = \int p(A_t|x_t,\mathtt{I}_{1:t})p(\mathbf{x}_t|\mathtt{I}_{1:t})d\mathbf{x}_t \ . \quad (6)$$

Note that the integration is performed over the posterior distribution of the location of the whole node set. The posterior association probability in the first image, $P(A_1|\mathtt{I}_1)$, is taken to be half, and the prior probability for non-association is the complement of the prior association probability: $P(\overline{A}_t|\mathtt{I}_{1:t-1}) = 1 - P(A_t|\mathtt{I}_{1:t-1})$.

The associative likelihood model is based on Gabor filter responses [1], [17]. The used filter bank consists of six differently oriented Gabor filters ($\theta = 0, \pi/6, ..., 5\pi/6$) at three frequencies ($f = \pi/4, \pi/8, \pi/16$), with a Gaussian deviation $\sigma_g = \pi$. The images are filtered with the Gabor filters; thus, the symbol $\mathtt{I}$ actually denotes an image, filtered with the filter bank. The magnitudes and phases of the different filter responses are denoted by a_i and φ_i. Let $g' = \{\mathbf{a}', \varphi'\}$ be reference Gabor responses and $g(x_t) = \{\mathbf{a}(x_t), \varphi(x_t)\}$ be the Gabor responses at the test image location x_t. A widely used similarity measure between the two Gabor responses is [1]

$$S(g(x_t), g') = \frac{\sum_i a_i(x_t)a'_i\cos(\varphi_i(x_t) - \varphi'_i)}{\sqrt{\sum_i a_i(x_t)^2 \sum_i a'^2_i}} \ , \quad (7)$$

where the summing is over the 18 filter responses.

Next, the associative likelihood term is integrated over the distribution of $g_{1:t-1}$, which denotes the Gabor responses of the node in images $1, ..., t-1$:

$$p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},A_t) = \int p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},A_t,g_{1:t-1}) \times p(g_{1:t-1}|\mathtt{I}_{1:t-1})dg_{1:t-1} = p(\mathtt{I}_t|x_t,A_t,\hat{g}_{1:t-1}) \ , \quad (8)$$

where the dependence on $\mathtt{I}_{1:t-1}$ has been dropped and a point estimate for $g_{1:t-1}$ has been plugged in. This means that the posterior distribution of $g_{1:t-1}$ - which is independent of x_t and A_t - is taken to be a delta function at the point estimate, being for image k:

$$\hat{g}_k = \frac{\int g_k(x_k)p(A_k|x_k,\mathtt{I}_{1:k})p(\mathbf{x}_k|\mathtt{I}_{1:k},\mathbf{x}_1)d\mathbf{x}_k}{\int p(A_k|x_k,\mathtt{I}_{1:k})p(\mathbf{x}_k|\mathtt{I}_{1:k},\mathbf{x}_1)d\mathbf{x}_k} \quad (9)$$

For the first image, the Gabor responses $\hat{g}_1$ are naturally evaluated at the selected nodes. The associative likelihood is taken to be a mixture of likelihood kernels, which are one for each processed image, so that the appearance of each matched node is a reference appearance for the node in the test image:

$$p(\mathtt{I}_t|x_t,A_t,\hat{g}_{1:t-1}) = \sum_{k=1}^{t-1} p(\mathtt{I}_t|x_t,A_t,\hat{g}_k) \ . \quad (10)$$

The likelihood kernels are heuristically built from the similarity measures (7):

$$p(\mathtt{I}_t|x_t,A_t,\hat{g}_k) = \exp(\beta S(g(x_t),\hat{g}_k)) \ , \quad (11)$$

where the value of β controls the steepness of the (unnormalized) likelihood. An example of a likelihood kernel, which typically is multimodal, is illustrated in figure 3.

The non-associative likelihood, $p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},\overline{A}_t)$, is difficult to compute. It is basically the probability of observing certain image detail in a test image, given that the image detail is not to be associated with the reference details. Looking at formulas (5) and (3), it can be seen that this value, multiplied by $P(\overline{A}_t|\mathtt{I}_{1:t-1})$, sets a level for how similar the Gabor responses have to be with the reference responses in order for the node to receive a high association probability. With very low value, each node is associated with the common object, while in the other extreme none is associated. A constant value is chosen for the non-associative likelihood. A heuristic setting is utilized to find an optimal value, being such that the association probabilities of the object nodes are as close to unity as possible, and those of the background nodes as low as possible. In [6]–[8] the distribution of the background parts are also modeled with a uniform density. Adding a positive constant with the associative likelihood ensures that the total likelihood - being product of the individual likelihoods - is always non-zero.

Finally, our likelihood model is condensed into a few words. The likelihood for matching a new image is a product of independent node likelihoods (equations (2) and (3)). The non-associative part $p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},\overline{A}_t)$ is a heuristically set constant. The associative part $p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1},A_t)$ is computed with (7), (10) and (11), where the reference Gabor responses $\hat{g}_k$ are computed with (9). The prior association probability $P(A_t|\mathtt{I}_{1:t-1})$ is computed with equations (4), (5) and (6).

Fig. 3. Small image: the starting image (same as in figure 2). Large image: test image, with contours of the likelihood kernel distribution ($\beta = 20$) of the reference node, marked with cross in the starting image, and manually annotated correct node location.

D. Prior - shape model

The mean shape of the node set in a test image, a priori to observing the image, is assumed to be the same as in the starting image, after put into same location and scale, with independent Gaussian deviations on the nodes. The

prior distribution, $p(\mathbf{x}_t|\mathbf{x}_1)$, is therefore defined as a Gaussian distribution in a translation and scale free space:

$$p(\mathbf{x}_t|\mathbf{x}_1) = \mathcal{N}\left(\frac{\mathbf{x}_t - E[\mathbf{x}_t]}{s(\mathbf{x}_t,\mathbf{x}_1)} \mid \mathbf{x}_1 - E[\mathbf{x}_1], \sigma^2\mathbf{I}\right) , \quad (12)$$

where $\mathbf{I}$ denotes a unit matrix, $E[\mathbf{x}]$ is the mean value of $\mathbf{x}$, σ^2 is the pre-fixed variance of the nodes and $s(\mathbf{x}_t,\mathbf{x}_1)$ is the scale of the node set $\mathbf{x}_t$ w.r.t. the node set $\mathbf{x}_1$. The scale is computed from the node locations:

$$s(\mathbf{x}_t,\mathbf{x}_1) = \frac{\sqrt{\sigma_u(\mathbf{x}_t)^2 + \sigma_v(\mathbf{x}_t)^2}}{\sqrt{\sigma_u(\mathbf{x}_1)^2 + \sigma_v(\mathbf{x}_1)^2}} , \quad (13)$$

where $\sigma_u(\mathbf{x})$ is the standard deviation of the horizontal components of the nodes in $\mathbf{x}$, and $\sigma_v(\mathbf{x})$ is that of the vertical components. Since the same Gabor filters are used with each scale, the likelihood values change with large scales. Therefore the performance of the system decreases if the size difference between the object instances in different images is large (say, more than one and half).

E. Sequential Monte Carlo sampling

The dimensionality of the posterior distribution is too huge to let the posterior to be evaluated on a dense enough grid. Also, the likelihood is such that it is difficult to approximate with any simpler function that could be integrated in closed form. Sampling methods are therefore applied to obtain weighted samples from the posterior distribution, and with these samples, the required integrals are estimated. Sequential Monte Carlo (SMC) methods are widely-used techniques for estimating the posterior distribution in dynamic state-space models [5], typical example being the tracking of a moving object. SMC algorithms can also be applied in a static setting. The static SMC scheme has the ability to handle multiple hypotheses of the posterior at the same time, making it a good choice to use for our problem as there are many possible locations for the nodes.

In our SMC implementation, all the data are available from the start, and the parameters (i.e. the coordinates of the nodes) are updated sequentially. The posterior distribution needs therefore to be defined in a conditional form so that it is possible to sample the ith node, given the locations of the already sampled nodes $\setminus i$. Since the node likelihoods are independent, it is enough to express the prior distribution in a conditional form, which is straightforward for a Gaussian distribution with a diagonal covariance matrix. For clarity, the image number index is dropped, so $\mathbf{x} \equiv \mathbf{x}_t$ and $\mathbf{x}' \equiv \mathbf{x}_1$ denote the test and reference node sets:

$$p(x_i|\mathbf{x}_{\setminus i},\mathbf{x}') = \mathcal{N}\left(\frac{x_i - E[\mathbf{x}_{\setminus i}]}{s(\mathbf{x}_{\setminus i},\mathbf{x}'_{\setminus i})} \mid x'_i - E[\mathbf{x}'_{\setminus i}], \sigma^2\mathbf{I}\right) . \quad (14)$$

The SMC implementation is illustrated in algorithm 1. Apart from the first component, the sampling order of the particles is sampled deterministically according to the prior association probabilities, so that the nodes that probably associate with the object tend to be matched first. Each particle thus samples the nodes in different order. The first component of the particles is matched from the likelihood. For the following components, the proposal distribution is a mixture of likelihood and prior terms, with the mixture coefficient φ heuristically determined on the basis of the node association probability. The purpose is that the associative nodes are sampled from the likelihood and the non-associative nodes from the prior. In the algorithm, only local area R around the prior mean is used. During the process the particles containing high likelihood nodes will

1. **Initialization,** $m = 1$
   for j = 1 to N **do**
   - Assign indices of first component for each particle, $\mathbf{J}_1^{(j)} = j \mod d$
   - Sample $x_i^* \sim p(\mathtt{I}_t|x_i,\mathtt{I}_{1:t-1},A_t)$ using $i = \mathbf{J}_1^{(j)}$
   - Set $\theta_1^{(j)} = x_i^*$ and $w_i^{(j)} = 1$

   end for
   - Set $m = 2$
2. **Importance sampling**
   for j = 1 to N **do**
   - Assign $\mathbf{J}_m^{(j)}$ from $1,...,d$ according to the prior association probabilities $P(\mathbf{A}_t|\mathtt{I}_{1:t-1})$ so that $\mathbf{J}_m^{(j)} \neq \mathbf{J}_{1:m-1}^{(j)}$
   - Assign $i = \mathbf{J}_m^{(j)}$ and $\mathbf{x}_{\setminus i} = \theta_{1:m-1}^{(j)}$
   - Sample $x_i^* \sim q(x_i|\mathbf{x}_{\setminus i},\mathbf{x}',\mathtt{I}) =$
   $\varphi \frac{p(\mathtt{I}_t|x_i,\mathtt{I}_{1:t-1})}{\sum_{x_i\in R} p(\mathtt{I}_t|x_t,\mathtt{I}_{1:t-1})} + (1-\varphi)\frac{p(x_i|\mathbf{x}_{\setminus i},\mathbf{x}')}{\sum_{x_i\in R} p(x_i|\mathbf{x}_{\setminus i},\mathbf{x}')}$, where
   $\varphi = 1 - \exp\left(1 - \frac{\sum_{x_i\in R}[p(\mathtt{I}_t|x_i,\mathtt{I}_{1:t-1})]}{\sum_{x_i\in R} p(\mathtt{I}_t|x_i,\mathtt{I}_{1:t-1},\overline{A}_i)P(\overline{A}_i|\mathtt{I}_{1:t-1})}\right)$
   - Set $\tilde{\theta}_{1:m}^{(j)} = (\theta_{1:m-1}^{(j)}, x_i^*)$
   - Set $w_m^{(j)} = \sqrt{w_{m-1}^{(j)} \frac{p(x_i^*|\mathbf{x}_{\setminus i},\mathbf{x}',\mathtt{I}_{1:t})}{q(x_i^*|\mathbf{x}_{\setminus i},\mathbf{x}',\mathtt{I}_{1:t})}}$

   end for
3. **Reducing the particle number**
   - Denote the current number of particles as $N' = N$
   - Eliminate M particles with lowest weights, where $M/N \ll 1$
   - Set the new number of particles to be $N = N' - M$
4. **Resampling with Langevin MOVE step**
   - Resample with replacement N particles $(\theta_{1:m}^{(j)}, j = 1,...,N)$ from the set $(\tilde{\theta}_{1:m}^{(j)}, j = 1,...,N)$ according to the importance weights $w_{1:m}^{(j)}$

   for j = 1 to N **do**
   - For $l = 1,...,m$, assign $i = \mathbf{J}_l^{(j)}$ and set $E_l = p(\mathtt{I}|x_i,\mathtt{I}_{1:t-1})$ with probability $p(A_t|\theta_l^{(j)},\mathtt{I}_{1:t})$, otherwise set $E_l = p(x_i|\theta_{1:l-1,l+1:m}^{(j)},\mathbf{x}')$
   - For $n = 1 : N_{iter}$, assign $\setminus i = \mathbf{J}_{1:m}^{(j)}$ and sample $\theta_{1:m}^{(j)} \sim \theta_{1:m}^{(j)} - \frac{\varepsilon^2}{2}\frac{\partial E_{1:m}}{\partial x_{\setminus i}}(\theta_{1:m}^{(j)}) + \varepsilon\mathcal{N}(0,1)$

   end for
   - If $m < d$ set $m = m+1$ and go to step 2

Algorithm 1: The Sequential Monte Carlo implementation. The number of particles and nodes are N and d, and θ denotes the parameter vector. Also, following notation is adopted: $\mathbf{x} \equiv \mathbf{x}_t$, $\mathbf{x}' \equiv \mathbf{x}_1$.

survive the resampling and thus the partially matched node sets will contain different sets of nodes that probably associate with the object.

As compared to the basic SMC algorithm, some modifications have been made that take into account the static nature of our problem. The resampling scheme with square weights is applied, which keeps the particle weights of the previous nodes and decreases the variance of the importance weights. After sampling each node, a Langevin Monte Carlo [18] step is applied to improve the sampled parameter values. Also, it is desired to save the computation time without impoverishing the results. Since the hypothesis of the correct mode is assumed to get stronger as more components are matched, the particle number is reduced along the sampling procedure by removing some proportion of particles with lowest weights before resampling.

F. The incremental processing

1) Gabor transform the starting image $\mathtt{I}_1$
2) Select the nodes in the starting image and store the Gabor responses at each node. Set $t = 2$
3) Gabor transform the next image $\mathtt{I}_t$
4) Compute the likelihood with equation (2) and the succeeding equations
5) Sample the image with SMC algorithm 1
6) Estimate and store the posterior association probabilities of the nodes with equations (5) and (6)
7) Compute the mean Gabor responses with equation (9) and store them
8) Set $t = t + 1$ and go to step 3

Algorithm 2: The online algorithm for matching the corresponding points in the images.

Finally, algorithm 2 shows our online method for learning the corresponding points in an algorithmic form.

III. Experiments

The method was tested on images of two databases, with resolution 240×320 pixels. The first database contains 64 digital camera images of a dog. The second database is the DTU-IMM database [19], containing 37 images of human faces, into which random background using Caltech background images (http://www.robots.ox.ac.uk/∼vgg/data3.html) was added. For both databases, the same parameter values were used. The steepness parameter of the likelihood was set to $\beta = 20$, the variance of the Gaussian prior was $\sigma^2 = 20$, and the number of nodes was 20. The Langevin equation was iterated 10 times with leapfrog stepsize $\varepsilon = 1$. The number of SMC particles was 400 in the beginning, and it was reduced by one fourth at each iteration until it reached 50.

Figure 4 shows few sample runs. The object nodes seem to be correctly located and are clearly associated with the object, whereas the background nodes are clearly associated with the background. The association probabilities of some nodes, such as those near the ear of the dog or the hear nodes of the faces, stay close to 0.5 since they contain multiple different appearances. Note that some lines on the graphs are superimposed because the prior association probabilities of most of the object nodes similarly approach unity. It should also be noted that although for a human observer the added background textures of the face images form an evident pattern, the proposed method is based only on the node points with no segment analysis. Therefore, the regular shaped background patterns give no advantage for the method; in contrary, the background segment boundaries appearing in the same locations in each image complicates the matching as they produce false candidates for object nodes. An example of an SMC particle set is shown in figure 5. Note how the variance of the posterior seems to be less for the object nodes than for the background nodes, as expected due to the peaked nature of the likelihood.

Fig. 5. An example of SMC representation. Left: the first image. Right: the second image, with SMC particles superimposed. The particle weights are not shown.

Measuring the numerical performance of the proposed method in finding the common corresponding points in the set of images is not straightforward, as the resulting corresponding points differ for each image set. A laborious solution would be to manually annotate the locations of the object points found by the method in all the test images, and repeated test runs. An approximation of this error measure can be computed by morphing the images according to pre-annotated points (12 in the dogs images and 58 in the DTU database), and measuring the distance between the reference location of the node in the morphed image and the location set by the proposed method. For each matched image the weighted Euclidean distance between the object nodes in the starting image and the matched object nodes is computed, weights being the posterior association probabilities. To reduce the errors due to morphing, the mean of the distances was computed in both directions (the starting image morphed to match the test image and vice versa), and an average of those was taken.

The errors were computed using sequences of 10 images. To average over the influence of random processing order of the images, and especially over the effect of the starting image, the method was given multiple sequences, so that each image of the database was three times as the starting image, and the remaining images were chosen randomly. This resulted in $3 \cdot 64 = 192$ test runs for the dog database and $3 \cdot 37 = 111$ test runs for the DTU-IMM database. Nodes were classified as object nodes if the prior association probability exceeded a threshold of 0.85 in the end of the sequence. In the sample runs of figure 4, these nodes are marked with crosses. The threshold was merely chosen on the basis that it results in a reasonable number of object nodes which in average was 7. The errors are not very sensitive to the threshold since the distances are

Fig. 4. Two examples of matching sequences of ten images. The dots depict the Monte Carlo estimate of the posterior mean of the node sets. The green values of the dots are proportional to the posterior association probabilities of the nodes. The sizes of the dots are proportional to the prior association probabilities of the nodes which are also shown on the graphs on the right. The nodes whose prior association probability exceeds 0.85 in the end of the sequence are marked with plus signs in the images.

TABLE I
THE MATCHING ERRORS FOR THE TWO DATABASES. COLUMNS FROM LEFT THE RIGHT: MEAN ERROR AND STANDARD DEVIATION OF THE PROPOSED METHOD; MEDIAN ERROR; AN ESTIMATE OF THE MEASURING ERROR; DOUBLED MEAN ERROR; DOUBLED ESTIMATE OF THE TRUE MEAN ERROR OF THE MODEL; THE ERROR OF EBGM METHOD [1]; THE ERROR OF AAM METHOD [19]; THE ERROR OF THE BAYESIAN OBJECT MATCHING METHOD [3]. THE CITED METHODS ARE BATCH METHODS AND USED DOUBLED IMAGE SIZES SO THE FIGURES IN THE FIVE RIGHTMOST COLUMNS, I.E. RIGHT TO THE SECOND VERTICAL LINE, CAN BE COMPARED.

Database	M $\pm$ std	Med	E_{morph}	(M $\pm$ std) $\times 2$	$\hat{E} \times 2$	EBGM	AAM	BOM
Dogs	5.74 $\pm$ 4.79	4.47	4.57 $\pm$ 2.05	11.5 $\pm$ 9.57	11.1	-	-	-
DTU-IMM	6.25 $\pm$ 1.89	5.97	5.13 $\pm$ 1.72	12.4 $\pm$ 3.95	7.32	6.16 $\pm$ 1.75	5.74 $\pm$ 1.18	5.52 $\pm$ 1.46

weighted with the posterior association probabilities.

The Euclidean node-to-node errors, averaged over all the test runs, are tabulated in table I, together with the errors of the DTU-IMM database of other published methods. These batch methods are the Elastic Bunch Graph Matching (EBGM) method [1], Active Appearance Model (AAM) implementation of Stegmann [19] and the Bayesian Object Matching (BOM) method [3]. As the image size in our experiments was half of the original, doubled matching errors are also reported for comparison. The DTU-IMM errors of our system exceeds the errors of the other two methods for several reasons. The other methods use leave-one-out cross-validation with a large number (36) of manually annotated images in the training set, while the proposed method uses only the information from the images analyzed during the sequence. The manually annotated feature points are likely to be more informative than those automatically set in our system, there are more of them (58), and they are all located on the object, whereas the proportion of the object nodes in our node sets is typically less than half.

Finally, the measurement error caused by morphing the images produces additional errors due to morphing defects. A rough estimate for this error was formed in the following way. One of the three series in which each of the images was once the starting image was chosen. For each run, one of the remaining nine images was randomly chosen. It was then morphed to match the starting image, the nodes selected by the proposed method were manually annotated in the morphed image and the distance between the two was computed. Assuming that the matching errors and morphing errors are uncorrelated (so the mean squared errors are additive) a rough estimate of the true model error $\hat{E}$ is achieved. There are two obviously false dog database matches whose contribution to the mean square error is huge; by leaving these out, the figure drops from 11.1 to 7.17.

IV. DISCUSSION

This paper has proposed an online method for learning the corresponding points of an object, whose instances appear in

un-annotated grayscale images in arbitrary location and scale. This is a demanding task, compared to the task the existing feature point based matching algorithms are trying to solve. For instance, the batch method of [4] uses 600 annotated training images to match test images, which are simpler (faces with similar scale on a uniform background) than the ones used here. In our system, a reference node set is automatically located in the starting image. The node set is matched in the subsequent images using Bayesian framework. The likelihood is a mixture distribution whose kernels are evaluated at the mean Gabor responses of each matched image. The prior is a Gaussian distribution whose mean, after removing the translation and scale effects, is the reference node set, and whose covariance matrix is diagonal. Combining the likelihood and prior results in the posterior distribution for the locations of the nodes, which is sampled with SMC methods. Each node is assigned a probability, with which the node is associated with the object. Since the model contains parameters whose values are set more or less heuristically the framework cannot be considered as being 'pure' Bayesian.

Due to the loosely supervised nature of the problem, the efficiency of the method for locating the corresponding points in the images is difficult to measure. It was estimated by morphing the images according to pre-annotated points. The measuring error this analysis carries was estimated as well, and assuming that the measuring errors and model errors are uncorrelated, the model error of the DTU-IMM images is on a par with those of other published methods that use simultaneously 36 training images with 58 manually annotated object features to train the model [1], [3], [19]. Taking into account the incremental nature of our matching scheme and the existence of nodes selected in the added background, the results can be considered as surprisingly good. As to the unsuccessful matches, two error sources may contribute on these: either the posterior distribution is such that the strongest mode is not the correct location, or the SMC algorithm fails to converge to the strongest mode. Separation of these errors is, in practice, impossible.

The used likelihood is not robust to large scale changes. However, since scaling results in a linear shift of the response pattern of the Gabor filter bank along the spatial frequencies, scale invariance could be added at the cost of increased computing time by finding the highest likelihood over different scales. Likewise, orientation invariance could be added. Another (straightforward) improvement would be to update the variances of the prior model. The method is also capable of dealing with missing data, as the object instance lacking in an image can be considered as being totally in occlusion. However, the later such background images appear in the sequence, the easier it is for the method to match upcoming images, as the object representation is then stronger than at the beginning. Especially, it is desired that the instance in the starting image is a typical representative of the object, as the node appearances are initialized based on that.

The computational complexity of the proposed method is currently rather high. With unoptimized Matlab code, processing one image takes about one minute on a latest desktop computer. As the SMC particle number is proportional to the number of nodes d, the sampling is $\mathcal{O}(d^2)$ complex (not taking into account the decreasing of the particle size), and the computation of the likelihood fields is $\mathcal{O}(dt)$ complex, since the likelihoods contain one kernel for each of the t processed images. These - especially the sampling - are the two computational bottlenecks of the method. Luckily, the SMC algorithm is parallel by nature, as the particles are independent on each other, apart from the resampling step. Also, instead of carrying the likelihood kernels of all the matched images, some kernel selection method could be applied to select only the most informative kernels. The used likelihood could also be replaced by simpler models. Thus, with an efficient implementation, the processing time of an image could be dropped to a fraction.

REFERENCES

[1] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE TPAMI*, vol. 19, pp. 775–779, 1997.
[2] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE TPAMI*, vol. 23, no. 6, pp. 681–685, 2001.
[3] T. Tamminen and J. Lampinen, "Sequential Monte Carlo for Bayesian matching of objects with occlusions," *IEEE TPAMI*, vol. 28, pp. 930–941, 2006.
[4] J. Kamarainen, M. Hamouz, J. Kittler, P. Paalanen, J. Ilonen, and A. Drobchenko, "Object localisation using generative probability model for spatial constellation and local image features," in *Proc. ICCV*, 2007, pp. 1–8.
[5] C. Doucet, J. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.
[6] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. ECCV*, 2000, pp. 18–32.
[7] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. CVPR*, 2003, pp. 264–271.
[8] L. Fei-Fei, R. Fergus, and P. Perona, "A Bayesian approach to unsupervised one-shot learning of object categories," in *Proc. ICCV*, 2003, pp. 1134–1141.
[9] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *Proc. CVPR*, 2006, pp. 26–36.
[10] S. Lazebnik, C. Schmid, and J. Ponce, "A discriminative framework for texture and object recognition using local image features," *Lecture notes in computer science*, vol. 4170, p. 423, 2006.
[11] R. Fergus, P. Perona, and A. Zisserman, "A sparse object category model for efficient learning and complete recognition," in *Toward Category-Level Object Recognition*, ser. LNCS. Springer, 2007, vol. 4170, pp. 443–461.
[12] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 17–32, 2004.
[13] E. Borenstein, E. Sharon, and S. Ullman, "Combining top-down and bottom-up segmentation," in *Proc. CVPR Workshop*, 2004, pp. 46–53.
[14] J. Winn and N. Jojic, "Locus: Learning object classes with unsupervised segmentation," in *Proc. ICCV*, vol. 1, 2005.
[15] N. Ahuja and S. Todorovic, "Learning the taxonomy and models of categories present in arbitrary images," in *Proc. ICCV*, 2007, pp. 1–8.
[16] 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*, vol. 106, no. 1, pp. 59–70, 2007.
[17] 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]*, vol. 36, no. 7, pp. 1169–1179, 1988.
[18] R. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Department of Computer Science, University of Toronto, Tech. Rep., 1993.
[19] M. B. Stegmann, "Analysis and segmentation of face images using point annotations and linear subspace techniques," Informatics and Mathematical Modelling, Technical University of Denmark, Tech. Rep., 2002.