# Synergy Between Object Recognition and Image Segmentation using the Expectation Maximization Algorithm

Iasonas Kokkinos, *Member, IEEE,* and Petros Maragos, *Fellow, IEEE*

***Abstract*— In this work we formulate the interaction between image segmentation and object recognition in the framework of the Expectation Maximization (EM) algorithm. We consider segmentation as the assignment of image observations to object hypotheses and phrase it as the E-step, while the M-step amounts to fitting the object models to the observations. These two tasks are performed iteratively, thereby simultaneously segmenting an image and reconstructing it in terms of objects.**

**We model objects using Active Appearance Models (AAMs) as they capture both shape and appearance variation. During the E-step the fidelity of the AAM predictions to the image is used to decide about assigning observations to the object. For this we propose two top-down segmentation algorithms. The first starts with an oversegmentation of the image and then softly assigns image segments to objects as in the common setting of EM. The second uses curve evolution to minimize a criterion derived from the variational interpretation of EM and introduces AAMs as shape priors. For the M-step we derive AAM fitting equations that accommodate segmentation information, thereby allowing for the automated treatment of occlusions.**

**Apart from top-down segmentation results we provide systematic experiments on object detection that validate the merits of our joint segmentation and recognition approach.**

***Index Terms*— Image segmentation, object recognition, Expectation Maximization, Active Appearance Models, curve evolution, top-down segmentation, generative models.**

## I. Introduction

THE bottom-up approach to vision [28] has considered the interaction between image segmentation and object detection in the scenario where segmentation groups coherent image areas that are then used to assemble and detect objects. Due to its simplicity this approach has been widely adopted, but there is a growing understanding that the cooperation (*synergy*) of these two processes can enhance performance.

Models that integrate the bottom-up and top-down streams of information were proposed during the previous decade by researchers in cognitive psychology, biological vision and neural networks [12], [31], [33], [41], [48] where the primary concerns have been at the architectural and functional level. In this decade the first concrete computer vision approaches to the problem [7], [54] have inspired a host of more recent systems [6], [15], [21], [24], [25], [27], [32], [45], [51], [52], pursuing the exploitation of this idea.

Several of these works have been inspired from the analysis-by-synthesis framework of Pattern Theory [17], [34], [45]. In this setting a set of probabilistic, generative models are used to synthesize the observed image and the analysis task amounts to estimating the model parameters. This approach can simultaneously regularize low-level tasks using model-based information and validate object hypotheses based on how well they predict the image.

In our work we use Active Appearance Models (AAMs) as generative models and address the problem of jointly detecting and segmenting objects in images. Our main contribution, preliminarily presented in [21], is phrasing this task in the framework of the Expectation Maximization (EM) algorithm [13]. Specifically, we view image segmentation as the E-step, where image observations are assigned to the object hypotheses. Model fitting is seen as the M-step, where the parameters related to each object hypothesis are estimated so as to optimally explain the image observations assigned to it. Segmentation and fitting proceed iteratively; since we are working in the framework of EM, this is guaranteed to converge to a locally optimal solution.

To make the combination of different approaches tractable we build on the variational interpretation of EM; this phrases EM as the iterative maximization of a criterion that is a lower bound on the observation likelihood. Specifically, we consider two alternative approaches for the implementation of the E-step; the first uses initially an off-the-shelf oversegmentation algorithm and then assigns the formed segments to objects. The second uses a curve evolution-based E-step that combines AAMs with variational image segmentation. Both approaches can be seen as optimizing the criterion used in the variational interpretation of EM. Further, we combine AAM fitting and image segmentation based on this criterion. We derive modified fitting equations that incorporate segmentation information, thereby automatically dealing with occlusions.

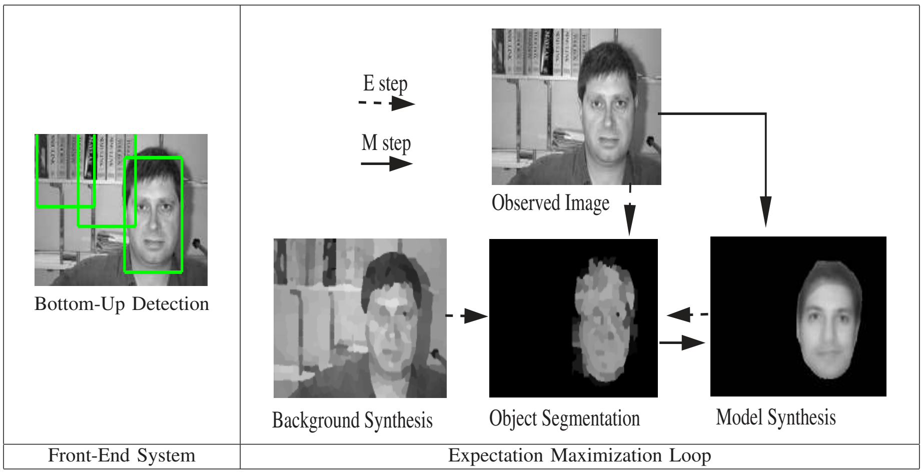
Finally, we provide systematic object detection results for faces and cars, demonstrating the merit of this joint segmentation and recognition approach.

*1) Paper Outline:* In Sec. II we introduce the basic notions of EM and give an overview of our approach. Sec. III presents the generative models we use and formulates the variational criterion optimized by EM. We present the two considered approaches for the E-step Sec. IV, and derive the M-step for AAMs in Sec. V. Experimental results are provided in Sec.

Manuscript received May 6, 2007; revised Jan. 14, 2008; accepted Apr. 30, 2008.

This work was supported by the Greek Ministry of Education research program “HRAKLEITOS”, and by the European Network of Excellence “MUSCLE”.

I. Kokkinos was with the National Technical University of Athens when this paper was first written. He is currently with the University of California at Los Angeles. P. Maragos is with the National Technical University of Athens.



System Pseudocode

```
$[\mathcal{O}_1, \dots, \mathcal{O}_K]$ = DETECT_OBJECTS($I$)
for $i = 1$ to $K$ do
    $\mathcal{A}_i$ = INITIALIZE($\mathcal{O}_i$)
    repeat
        SEG$_i$ = E_STEP($\mathcal{A}_{\mathcal{O}_i}, \mathcal{B}, I$)
        $\mathcal{A}_{\mathcal{O}_i}$ = M_STEP(SEG$_i$, $I$)
    until CONVERGENCE
    VERIFY($\mathcal{O}_i$, SEG$_i$, $\mathcal{A}_{\mathcal{O}_i}$)
end for
```

Fig. 1: Overview and pseudocode for our approach: a front-end object detection system provides a set of candidate object locations. The location of each object hypothesis $\mathcal{O}_i$ is used to initialize the parameters $\mathcal{A}_i$ of a generative model, that then enters enters an EM-loop. In the E-step the object obtains the image areas it explains better than the background and in the M-step the model parameters are updated. After convergence, the model parameters and the object segmentation are used to verify object hypotheses and prune false positives.

VI, while Sec. VII places our work in the context of existing approaches; technical issues are addressed in App. I.

## II. EM Approach to Synergy

Our work builds on the approach of generative models to simultaneously address the segmentation and recognition problems. For the purpose of segmentation we use the fidelity of the generative model predictions to the image in order to decide of the image a model should occupy. Regarding recognition, each object hypothesis is validated based on the image area assigned to the object, as well as the estimated model parameters, which indicate the familiarity of the object appearance.

This yields however an intertwined problem: on the one hand knowing the area occupied by an object is needed for the estimation of the model parameters and on the other the model synthesis is used to assign observations to the model. Since neither is known in advance, we cannot address each problem separately. We view this problem as an instance of the broader problem of parameter estimation with missing data: in our case the missing data are the assignments of observations to models. A well-known tool for addressing such problems is the EM algorithm [13], which we now briefly describe for the problem of parameter estimation for a mixture distribution [5] before presenting how it applies to our approach.

### *A. EM algorithm and Variational Interpretation*

Consider generating an observation $I_n$ by first choosing one out of $K$ parametric distributions, with prior probability $\pi_k$ and then drawing a sample from that distribution with probability $P(I_n|\theta_k)$. EM addresses the task of estimating the parameter set $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_k\}, \quad \mathcal{A}_k = (\theta_k, \pi_k)$, that optimally explains a set of observations $I = \{I_1, \dots, I_N\}$ generated this way.

The missing data are the identities of the distributions used to generate each observation; these are represented with the binary *hidden variable* vectors $\mathbf{z}_n = [z_{n,1}, \dots, z_{n,K}]^T$. $\mathbf{z}_n$ corresponds to the $n$-th observation, and its unique non-zero element indicates the component used to generate $I_n$. By summing over the unknown hidden variables $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ we can express the likelihood of the observations given the parameter set:

$$\log P(I|\mathcal{A}) = \sum_{n=1}^{N} \log P(I_n|\mathcal{A}) = \sum_{n=1}^{N} \log \sum_{\mathbf{z}_n} P(I_n, \mathbf{z}_n|\mathcal{A}) \tag{1}$$

We can write the last summand as:

$$P(I_n, \mathbf{z}_n|\mathcal{A}) = P(I_n|\mathbf{z}_n, \mathcal{A})P(\mathbf{z}_n|\mathcal{A}) = \prod_{k=1}^{K} [\pi_k P(I_n|\theta_k)]^{z_{n,k}} \tag{2}$$

Finding the optimal estimate $\mathcal{A}^*$ is intractable, since the summation over $\mathbf{z}_n$ appears inside the logarithm in (1). However, for given $\mathbf{Z}$, one can write the *full observation log likelihood*:

$$\log P(I, \mathbf{Z}|\mathcal{A}) \quad = \quad \sum_n \sum_k z_{n,k} \log\left(\pi_k P(I_n|\theta_k)\right). \tag{3}$$

The parameters in this expression can be directly estimated since the summation appears outside the logarithm.

The EM algorithm exploits this by introducing the expectation of (3) with respect to the posterior distribution of $z_{n,k}$. Denoting by $\mathbf{z}_{n,k}$ the vector $\mathbf{z}_n$ that assigns observation $n$ to the $k$-th mixture, i.e. has $z_{n,k} = 1$, we write the EM algorithm as iterating the following steps:

- E-step: derive the posterior of $\mathbf{z}$ conditioned on the previous parameter estimates, $\mathcal{A}^*$ and the observations:

$$E_{n,k} \equiv P(\mathbf{z}_{n,k}|I_n, \mathcal{A}^*) = \frac{\pi_k^* P(I_n|\theta_k^*)}{\sum_j \pi_j^* P(I_n|\theta_j^*)}, \tag{4}$$

and form the expected value of the log-likelihood under this probability mass function:

$$\langle \log P(I, \mathbf{Z}|\mathcal{A}^*) \rangle_E = \sum_n \sum_k E_{n,k} \log\left(\pi_k P(I_n|\theta_k)\right) \tag{5}$$

• M-step: maximize the expected log-likelihood with respect to the distribution parameters:

$$\pi_k^* = \frac{\sum_n E_{n,k}}{N}, \quad \theta_k^* = \text{argmax} \sum_n E_{n,k} \log P(I_n|\theta_k) \quad (6)$$

Intuitively, in the E-step the unobserved binary variables in (3) are replaced with an estimate of each mixture's 'responsibility' for the observations, which is then used to decouple parameter estimation in the M-step. This consistently increases the likelihood [13] and converges to a local maximum of (1).

EM can also be seen as a variational inference algorithm [18] along the lines of [35]. There it is shown to iteratively maximize a lower bound on the observation likelihood:

$$\begin{aligned} \log P(I|\mathcal{A}) &\geq LB(I,Q,\mathcal{A}) \\ LB(I,Q,\mathcal{A}) &= \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{P(I|\mathbf{Z},\mathcal{A})P(\mathbf{Z}|\mathcal{A})}{\log Q(\mathbf{Z})}. \end{aligned} \quad (7)$$

The bound $LB$ is expressed in terms of $Q$, an unknown distribution on the hidden variables $\mathbf{Z}$, and the parameter set $\mathcal{A}$. The form in (7) is derived from Jensen's inequality. Typically $Q$ is chosen from a manageable family of distributions; for example by choosing a factorizable distribution $Q = \prod Q_n(\mathbf{z}_n)$ computations become tractable since the summations in (7) break over $n$.

The individual distribution $Q_n(\mathbf{z}_n)$ determines the probability of assigning the $n$-th observation to one of the $K$ components. To make the relation with (4) clear, we use $Q_{n,k}$ to denote the probability of $\mathbf{z}_{n,k}$. By breaking the product in the logarithm we can thus write (7) as:

$$\begin{aligned} LB(I,Q,\mathcal{A}) &= \sum_{n,k} Q_{n,k}[\log P(I_n|\mathcal{A}_k) \\ &+ \log P(\mathbf{z}_{n,k}|\mathcal{A}_k) - \log Q_{n,k}]. \end{aligned} \quad (8)$$

Maximizing the bound in (8) with respect to $Q$ subject to the constraint that $\sum_k Q_{n,k} = 1, \forall n$ leads to $Q_{n,k} = E_{n,k}$. So, the variational approach to EM interprets the E-step as a maximization with respect to Q.

Apart from providing a common criterion for the two segmentation algorithms used subsequently, this formulation makes several expressions easier. For example, by breaking the product in (7) and keeping the term $\sum_{\mathbf{Z}} Q(\mathbf{Z}) \log P(\mathbf{Z}|\mathcal{A})$, we have a quantity that captures prior information about assignments. For mixture modeling this simply amounts to the expression $\sum_n \sum_k Q_{n,k} \log \pi_k$, that favors assignments to clusters with larger mixing weights. In image segmentation however there are other forms of priors, such as small length of the boundaries between regions, or object-specific priors, capturing the shape properties of the object. We will express all of these in terms of $Q(\mathbf{Z}) \log P(\mathbf{Z}|\mathcal{A})$.

### B. Application to Synergy

In the mixture modeling problem the hidden variable vectors provide an assignment of each observation to a specific mixture component. The analogy with our problem comes by seeing the object models as the mixture components and the hidden variables as providing the image segmentation.

We apply the EM algorithm to our problem by treating segmentation as the E-step and model fitting as the M-step as shown in Fig. 1. In the E-step we determine the responsibility of the object model for image observations and in the M-step we estimate the model parameters so as to optimally explain the data that it has occupied. Intuitively we consider segmentation as determining a window through which the object is seen, with binary hidden variables determining whether the object is visible or not. Top-down segmentation decides where it is best to open this window, while model fitting focuses on the object parts seen through it.

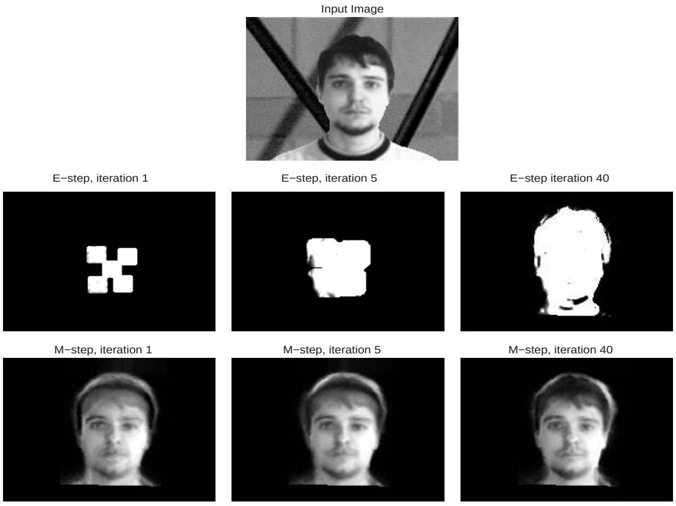
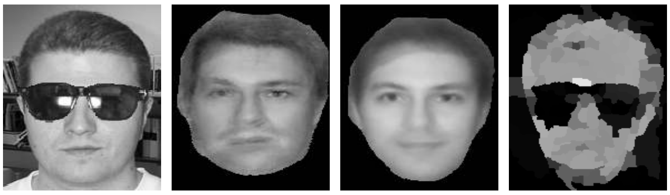


Fig. 2: Improvement of the segmentation and parameter estimates at increasing iterations of EM: The middle row shows the evolution of the face hypothesis region (E-step) and the bottom row shows object fitting results, using the above region (M-step).

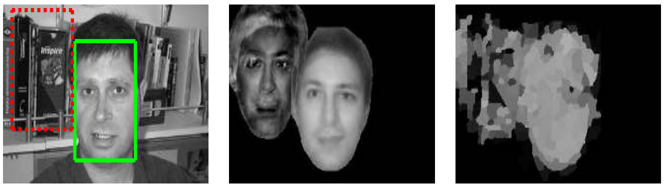
Illustrating this idea, Fig. 2 shows the result of iterating the E- and M-steps for a toy example: Starting from a location in the image proposed by a front-end detection system, the synthesis and segmentation gradually improve, converging to a solution that models a region of the image in terms of an object. The assignment of observations to a model and the estimation of the model parameters proceed in a gradual, relaxation-type fashion until convergence.

Apart from providing a top-down segmentation of the image, this idea can be useful for two more reasons: first,



(a) Input (b) Plain AAM (c) EM-based AAM (d) E-step results

Fig. 3: Dealing with occlusion: the sunglasses in (a) lead to erroneous AAM fits, as shown in (b). The EM approach leads to the more robust fit in (c) since the E-step results in (d) do not assign the sunglass region to the object.



(a) Detections (b) Syntheses (c) Segmentations

Fig. 4: Top-down information helps prune false positives: Background clutter leads to a false positive, shown with a red-dashed box in (a); this is pruned due to both the unlikely AAM parameter estimates, witnessed as a non-typical face in (b) and the lower values of the E-step results, shown by a lower gray value in (c).

we use segmentation information to deal with occlusion. The E-step can decide to assign occluded parts to the background, thereby freeing the object from explaining these areas. The fitting can therefore focus on the areas that actually belong to the object, as shown in Fig. 3: based on our approach the synthesis captures more accurately the intensity pattern of the face and gives reasonable predictions in the part that has been occluded. We address this aspect in further detail in Sec. V.

Second, we can use the E-step results as well as the AAM parameters to prune false positives, as shown in Fig. 4. The likelihood of the AAM parameters under the model's prior distribution indicates how close the observed image is to the object category, which helps discard false positives. Further, the E-step results quantify the fidelity of the model to the image data in terms of the extent of the area assigned to it. Object hypotheses generated from detections due to background clutter have a low chance of explaining a large part of the image and thereby obtain a smaller area. We systematically evaluate the merit of these ideas in Sec. VI.

Both of these uses could, in principle, be pursued with different approaches like the stochastic search over models and segmentations of [45]. However our work makes broadly accessible the use of a bottom-up/top-down loop by using a deterministic and well-studied inference algorithm. Both the EM algorithm and the system components are widely used in current research, and can be incorporated with little additional effort in existing systems.

## III. Generative Models and EM Criterion

A basic ingredient of our approach is the use of generative models; such models are popular in computer vision as they can be used to formulate in a principled manner problems like detection, tracking and in our case top-down segmentation. For object detection such models are used extensively in the setting of part-based object models. In our work we are interested in modeling the whole area occupied by an object instead of a few interest-points or features. We therefore consider global generative models for image intensity.

We now introduce the models we use for our object categories and the alternative, background hypothesis. At the end of this section we combine them in an EM criterion used in the rest of the paper. This is then maximized by the E- and M- steps of our approach.

### A. Object Model: AAMs

For Fig. 2 a PCA basis for faces [47] was used as a generative model, resulting in 'ghosting artifacts' e.g. around the hair. This is due to the absence of a registration step in typical PCA models that perplexes both the modeling and the segmentation of deformable objects.

We therefore use Morphable- Active Appearance Models (AAMs) [9], [20], [30] as models that explicitly account for shape variability and can drive both the analysis and segmentation tasks. Since we want our approach to be broadly applicable to object detection, we use AAMs learned with the approach of [23]. The only information used there is the bounding box of the object, which is used also by most unsupervised learning algorithms for object detection.

AAMs model separately shape and appearance variation using linear expressions, and combine them in a nonlinear manner. Specifically, a deformation field $S$

$$S(\mathbf{x};\mathbf{s}) \equiv (S_x(\mathbf{x};\mathbf{s}), S_y(\mathbf{x};\mathbf{s})) = \sum_{i=1}^{N_S} \mathbf{s}_i \mathcal{S}_i(\mathbf{x}) \tag{9}$$

is synthesized to bring the image pixel $(S_x(\mathbf{x};\mathbf{s}), S_y(\mathbf{x};\mathbf{s}))$ in registration with the template pixel $\mathbf{x} = (x, y)$. The appearance $T$ is synthesized on the deformation-free template grid as

$$T(\mathbf{x};\mathbf{t}) = \mathcal{T}_0(\mathbf{x}) + \sum_{i=1}^{N_{\mathcal{T}}} \mathbf{t}_i \mathcal{T}_i(\mathbf{x}). \tag{10}$$

The model parameters are the shape and texture coefficients $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{N_S})$, $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_{N_{\mathcal{T}}})$, while $\mathcal{S}$, $\mathcal{T}$ are the corresponding basis elements and $\mathcal{T}_0(\mathbf{x})$ is the mean appearance.

Given an observed image $I$, AAM fitting iteratively minimizes w.r.t. $\mathbf{s}$ and $\mathbf{t}$ a criterion defined on the template grid:

$$E(\mathbf{s},\mathbf{t}) = \sum_{\mathbf{x}} H(\mathbf{x}) \left(I(S(\mathbf{x};\mathbf{s})) - T(\mathbf{x};\mathbf{t})\right)^2, \tag{11}$$

where $H(\mathbf{x})$ is the indicator function of the object's support. Observations at locations that do not get warped to the interior of this support cannot be modeled by the AAM and therefore do not contribute to the error.

Under a white Gaussian noise error assumption the log-likelihood of $I(\mathbf{x})$ writes:

$$\log P(I(\mathbf{x})|\mathbf{s},\mathbf{t}) = -\frac{\left(I(\mathbf{x}) - T(S^{-1}(\mathbf{x};\mathbf{s});\mathbf{t})\right)^2}{2\sigma^2} - \frac{\log 2\pi\sigma^2}{2}. \tag{12}$$

Here $S^{-1}$ fetches from the template coordinate system the prediction $T(S^{-1}(\mathbf{x};\mathbf{s});\mathbf{t})$ corresponding to the observed value $I(\mathbf{x})$ and as above, this equation holds only if $H(S^{-1}(\mathbf{x};\mathbf{s})) = 1$, namely if $\mathbf{x}$ can be explained by the AAM.

If the magnification or shrinking of the template point $\mathbf{x}$ is negligible we have $P(I|\mathbf{s},\mathbf{t}) \propto \exp(-E(\mathbf{s},\mathbf{t})/(2\sigma^2))$, which interprets AAM fitting as providing a Maximum Likelihood parameter estimate. Further, we can perform Maximum-A-Posterior estimation by introducing a quadratic penalty on model parameters in (11), which equals the log-likelihood of the parameters under a Gaussian prior distribution.

### B. Background Model: Piecewise Constant Image

To determine the assignment of observations to the object we need a background model as an alternative to compete with. There are several ways to build a background model, depending on the accuracy required from it. At the simplicity extreme, for Fig. 2 we use a nonparametric distribution for the image intensity that is estimated using the whole image domain. However, for images with complex background this distribution becomes loose, and the object model may be better even around false positives. The more complex, full-blown generative approach of [45], [46] pursues the interpretation of the whole image so there is no generic background model. Practically, for the joint segmentation and detection task this could be superfluous: as we show in the experimental results a simple background model can both discard false positives and exclude occluded areas from model fitting.

The approach we take lies between these two cases. We consider that the background model is built by a set of regions, within which the image has constant intensity; this is the broadly used piecewise-constant image model. We assume that within each region $r$ the constant value is corrupted by white Gaussian noise, and estimate the parameters $(\mu_r, \sigma_r)$ from the mean and standard deviation of the region's image intensities. These, together with the prior probability $\pi_{\mathcal{B}_r}$ of assigning an observation to the region form the parameter set for background region r: $\mathcal{A}_{\mathcal{B}_r} = (\mu_r, \sigma_r, \pi_{\mathcal{B}_r})$.

We can combine all sub-models in a single background hypothesis $\mathcal{B}$, under which the likelihood of $I(\mathbf{x})$ writes:

$$\begin{aligned} P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}}) &= \prod_{r=1}^{R} [P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}_r})]^{H_r(\mathbf{x})} \\ &= N(\mu_i - I(\mathbf{x}), \sigma_i) \end{aligned} \tag{13}$$

where $\mathcal{A}_{\mathcal{B}} = (\mathcal{A}_{\mathcal{B}_1}, \ldots, \mathcal{A}_{\mathcal{B}_R})$, $H_r(\mathbf{x})$ is the support indicator for the $r$-th region and $i$ is the index of the region that contains $x$, i.e. $H_i(\mathbf{x}) = 1$. Implicitly, for (13) we assume that $\pi_{\mathcal{B}_r}$ does not depend on $r$, and condition on $I(\mathbf{x})$ belonging to the background; otherwise a $\pi_{\mathcal{B}_i}$ term would be necessary. This is an expression we will use in the following when convenient.

### C. EM criterion for Object vs Background Segmentation

We now build a lower bound on the likelihood of the image observations under the mixture of the object and background models. For the sake of simplicity we formulate it for the case of jointly segmenting and analyzing a single object; the generalization to multiple objects is straightforward.

We split the bound in (8) into object- and background- related terms. Since our models are formulated in the continuous domain but EM considers a discrete set of observations, we denote below with $\mathbf{x}_n$ the image coordinate corresponding to observation index $n$.

We first consider the part of the EM bound in (8) that involves the object hypothesis, $\mathcal{O}$. This can be expressed in terms of the column of $Q_{n,k}$ that relates to $\mathcal{O}$, $Q_{\mathcal{O}}$ and the object parameters $\mathcal{A}_{\mathcal{O}} = (\mathbf{s}, \mathbf{t}, \pi_{\mathcal{O}})$ that include the AAM parameters $\mathbf{s}, \mathbf{t}$ and the prior probability $\pi_{\mathcal{O}}$ of assigning an observation to the object if it falls within its support. Using these we write the related part of the bound as:

$$LB(I, Q_{\mathcal{O}}, \mathcal{A}_{\mathcal{O}}) = \sum_n Q_{n,\mathcal{O}} \left[\log P(I_n|\mathcal{A}_{\mathcal{O}}) + \log P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{O}})\right]. \tag{14}$$

Here $P(I_n|\mathcal{A}_{\mathcal{O}}) = P(I(\mathbf{x}_n)|\mathbf{s}, \mathbf{t})$ is the observation likelihood under the appearance model of (12) and $\mathbf{z}_{n,\mathcal{O}}$ is the hidden variable vector that assigns the observation $n$ to hypothesis $\mathcal{O}$.

The term $P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{O}})$ equals the prior probability of $\mathbf{z}_{n,\mathcal{O}}$ under the AAM model and constrains the AAM to only model observations in the template interior. Specifically, we have:

$$P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{O}}) = H(S^{-1}(\mathbf{x}_n, \mathbf{s}))\pi_{\mathcal{O}}. \tag{15}$$

In words, hypothesis $\mathcal{O}$ can take hold of observation $n$ only if $S^{-1}$ brings it inside the object's interior. In that case, the prior probability of obtaining it is $\pi_{\mathcal{O}}$. This brings shape information directly in segmentation without introducing additional terms to a segmentation criterion as is done e.g. in [11], [43]. We therefore see AAMs as providing a natural means to introduce shape-related information in segmentation.

For the background model we adopt the mixture modeling approach described in the previous subsection and write:

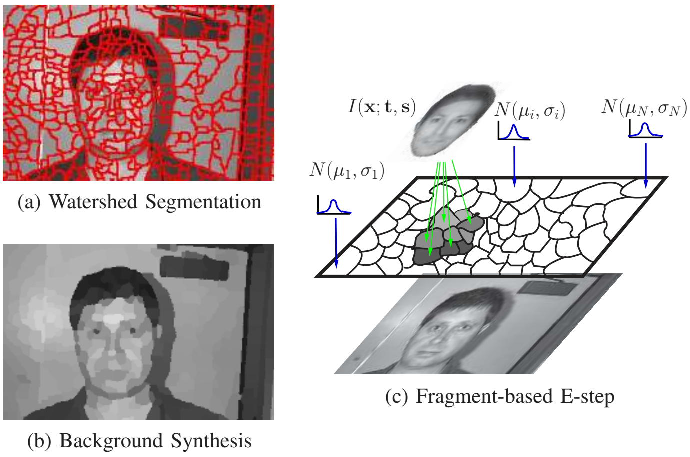
$$\begin{aligned} LB(I, Q_{\mathcal{B}}, \mathcal{A}_{\mathcal{B}}) &= \sum_{n,r} Q_{n,\mathcal{B}_r}[\log P(I_n|\mathcal{A}_{\mathcal{B}_r}) \\ &\quad + \log P(\mathbf{z}_{n,\mathcal{B}_r}|\mathcal{A}_{\mathcal{B}_r})]. \end{aligned} \tag{16}$$

As in (14), $Q_{\mathcal{B}}$ are the columns of $Q_{n,k}$ related to the background hypotheses and $\mathcal{A}_{\mathcal{B}}$ are the corresponding parameters. The first summand is the likelihood of the observations under the $r$-th background sub-model. The second summand is a prior distribution over the assignments that we use to balance the complexity of the fore- and background models. Specifically, the AAM has often larger reconstruction error than the background model, since it explains an heterogenous set of observations with a varying set of intensities. Instead, the background regions are determined using bottom-up cues and have almost constant intensity, thereby making it easier to model their interiors. We therefore assign observations to the object model more easily by setting $P(\mathbf{z}_{n,\mathcal{B}_r}|\mathcal{A}_{\mathcal{B}_r}) = \pi_{\mathcal{B}_r}$ to a low value; this gives rise later to 'MDL' or 'balloon' terms.

We combine these two terms with a scaled version of the entropy-related term of (7) and obtain the following lower bound on the log-likelihood of the data:

$$\begin{aligned} LB(I, Q, \mathcal{A}) &= \sum_n \sum_{h \in \{\mathcal{O}, \mathcal{B}_1, \ldots, \mathcal{B}_R\}} Q_{n,h} \Bigg[\log P(I_n|\mathcal{A}_h) \\ &\quad + \log P(\mathbf{z}_{n,h}|\mathcal{A}_h) - \frac{1}{\alpha}\log Q_{n,h}\Bigg] \end{aligned} \tag{17}$$

where $Q = \{Q_{\mathcal{O}}, Q_{\mathcal{B}}\}$ and $\mathcal{A} = \{\mathcal{A}_{\mathcal{O}}, \mathcal{A}_{\mathcal{B}}\}$. The last summand favors high-entropy distributions and leads to soft assignments. Since $-\sum_{n,h} Q_{n,h} \log Q_{n,h} \geq 0$, for all $\alpha \geq 1$ we have a lower bound on the log-likelihood: for $\alpha = 1$ we have the original EM bound of (7), while in the winner-take-all version of EM described in [35] we set $\alpha \to \infty$, so the entropy term vanishes and all assignments become hard. This is also the common choice for image segmentation.



(a) Watershed Segmentation

(b) Background Synthesis

(c) Fragment-based E-step

Fig. 5: Fragment-based E-step: We break the image into fragments using the watershed algorithm as shown in (a). The background model uses a Gaussian distribution within each fragment and its prediction, shown in (b), is constant within each fragment. During the E-step the occupation of fragments is determined based on whether the object synthesis, $I(\mathbf{x};\mathbf{s},\mathbf{t})$ reconstructs the image better than the background model. The gray value indicates the degree to which a fragment is assigned to the object.

We can now proceed to the description of the E- and M-steps; they are both derived so as to minimize (17) with respect to $Q$ and $\mathcal{A}$ respectively.

## IV. E-step: Object-Based Segmentation

In what follows we present two alternatives to implementing the E-step; each constitutes a different approach to finding the background regions and minimizing the EM criterion of (17).

Our initial approach of [21], described in Sec. IV-A, utilizes an initial oversegmentation to both determine the background model and implement the E-step. This is efficient and modular, since any image segmentation algorithm can be used at the front-end. Still, it does not fully couple the segmentation and analysis tasks, since the initial segmentation boundaries cannot be modified. We therefore subsequently propose an alternative in Sec. IV-B that utilizes curve evolution for the E-step, incorporating smoothness priors and edge information. This yields superior segmentations but comes at the cost of increased computation demands; these can be overcome using efficient algorithms such as [38].

### *A. Fragment-based E-step*

As suggested in [2], [32] an initial oversegmentation of the image can efficiently recover most object boundaries. Adopting this approach, in our work we use the morphological watershed algorithm [4]. Specifcially, we use the Brightness-Gradient boundary strength function of [29] to obtain both edges and markers; we extract the latter from the local minima of the boundary strength function. As shown in Fig. 5, this gives us a small set of image fragments that we use in two complementary ways.

First, we define a background distribution by modeling the image intensities within each fragment with a normal distribution. We thereby build our piecewise-constant background model with a set of fixed regions.

Second, since these regions are highly cohesive, we treat them as 'bundled' observations - or 'atomic regions' in [2] and 'superpixels' in [32]. We thus use a fragment-based E-step that uniformly assigns an image fragment to either the object or the background hypothesis. This reduces the number of assignment variables considered from the number of pixels to the number of fragments.

We now consider the part of the EM criterion involving observations in region $R_r$, by limiting the summation in (17) to $n \in R_r$. We can simplify its expression by noting first that only the background sub-model $\mathcal{B}_r$ built within region $r$ is active, and second by using a common value $Q_{r,k}$ for the related assignment variables $Q_{n,k}, n \in R_r$. Further, since only the object and a single background hypothesis are entailed, we set $q_r = Q_{r,\mathcal{O}} = 1 - Q_{r,\mathcal{B}_r}$ for simplicity. We can thus rewrite the considered part of (17) as:

$$\begin{aligned} LB(I, q_r, \mathcal{A}) = \sum_{n \in R_r} & q_r \left[\log P(I_n|\mathcal{A}_\mathcal{O}) + \log P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_\mathcal{O})\right] \\ & + (1-q_r)\left[\log P(I_n|\mathcal{A}_{\mathcal{B}_r}) + \log P(\mathbf{z}_{n,\mathcal{B}_r}|\mathcal{A}_{\mathcal{B}_r})\right] \\ & - \frac{1}{\alpha}\left[q_r \log q_r + (1-q_r)\log(1-q_r)\right] \end{aligned}$$

Substituting from (15) and maximizing with respect to $q_r$ gives:

$$\frac{1}{\alpha}\log\frac{q_r}{1-q_r} - \beta = \frac{1}{|R_r|}\sum_{n \in R_r}\log\frac{P(I_n|\mathcal{A}_\mathcal{O})H(S^{-1}(\mathbf{x}_n,\mathbf{s}))}{P(I_n|\mathcal{A}_{\mathcal{B}_r})}, \tag{18}$$

where $\beta = \log\frac{\pi_\mathcal{O}}{\pi_{\mathcal{B}_r}}$ and $|R_r|$ is the cardinality of region $r$. We treat $\beta$ as a design parameter that allows us to determine how easily we assign fragments to the object. Finally, we use the notation $\overline{\log\frac{P(I|\mathcal{O})}{P(I|\mathcal{B})}}$ for the right hand side of (18) so the optimal $q_r$ is given by a sigmoidal function:

$$q_r = \frac{1}{1 + \exp\left(-\alpha\left[\overline{\log\frac{P(I|\mathcal{O})}{P(I|\mathcal{B})}} + \beta\right]\right)} \tag{19}$$

For all experiments we use the values $\alpha = 10, \beta = 1$, estimated by tuning the system's performance on a few images. We note that a different front-end segmentation algorithm might require different values for $\alpha$ and $\beta$. For example if the segments returned were significantly smaller, a lower value for $\beta$ would be needed: as argued in Sec. III-C, in that case the background model would generally be more accurate, so we would need to make it even easier for the foreground model to acquire a part. To avoid manual tuning, one can therefore use the simple learning-based approach we had initially used in [21] to estimate $\alpha$ and $\beta$ from ground truth data.

On the left of each column pair in Fig. 7 we demonstrate top-down segmentation results for faces and cars that validate our system's ability to segment objects of varying shape and appearance. We show the border of the region that is obtained by thresholding the results of the E-step for the object corresponding to the strongest bottom-up hypothesis.

The segmentations are generally appealing, correctly capturing the pose of the object categories considered, while excluding unpredictable locations like beards for faces or pedestrians for cars. However, jagged boundaries can occur,

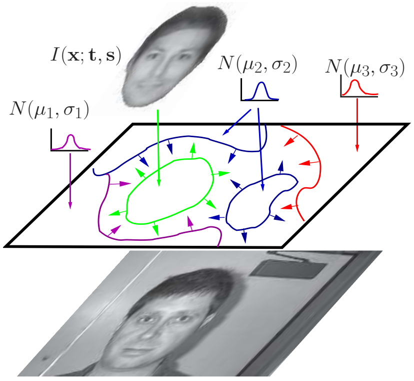


Fig. 6: Curve evolution-based E-step: we represent the object region as the interior of an evolving contour. To occupy image observations the object region changes its boundary by competing with a set of deformable background hypotheses.

due to the E-values of some fragment falling below threshold. Further, inaccuracies of front-end segmentation propagate to the top-down segmentation as is more prominent for the car images where the low-level cues are unreliable; these problems led us to consider the segmentation scheme presented next.

### B. Curve Evolution-based E-step

In this second approach to implementing the E-step a small set of deformable regions constitute our background model, as shown in Fig. 6. Their boundaries evolve so that each region occupies a homogeneous portion of the image while at the same time the boundary of the object region evolves to occupy the parts explained by it. This is the common curve evolution approach to image segmentation [8], [53] that is typically driven by the the minimization of variational criteria. These criteria can incorporate smoothness and edge-based terms, thereby addressing the problems of the previous method.

Our contributions consist in using the variational interpretation of EM to justify the use of such methods in our setting, and introducing AAMs as shape priors for segmentation.

*1) Region Competition and EM Interpretation:* Region Competition is a variational algorithm that optimizes a probabilistic criterion of segmentation quality. Using $K$ regions $R_k$ and assuming the observations within region $k$ follow a distribution $P(\cdot|\mathcal{A}_k)$, the likelihood of the observations for the current segmentation is considered as a term to be maximized. Combining the observation likelihood with a prior term that penalizes the length of the region borders, $\Gamma = \{\Gamma_1, \ldots, \Gamma_K\}$ gives rise to the Region Competition functional [53]:

$$J(\Gamma, \mathcal{A}) = \sum_{k=1}^{K} \frac{\mu}{2} \int_{\Gamma_k} ds - \iint_{R_k} \log P(I(\mathbf{x})|\mathcal{A}_k) d\mathbf{x}, \quad (20)$$

where $\mu$ controls the prior's weight. Calculus of variations yields the evolution law:

$$\frac{\partial \Gamma_k}{\partial t} = -\mu \kappa \mathcal{N} + \log \frac{P(I(\mathbf{x})|\mathcal{A}_k)}{P(I(\mathbf{x})|\mathcal{A}_m)} \mathcal{N} \quad (21)$$

where $P(I(\mathbf{x})|\mathcal{A}_m)$ is the log-likelihood of $I(\mathbf{x})$ under the competing neighboring hypothesis $m$, $\kappa$ is the $k$-th border curvature and $\mathcal{N}$ its outward normal unit vector. A region boundary moving according to (21) assigns observations to the region that predicts them better while maintaining the borders smooth, as it minimizes the functional (20).

There is an intuitive link between Region Competition and EM: the E-step is similar to curve evolution, where observations are assigned to region hypotheses and the M-step to updating the parameters of the region distributions. The difference is that instead of a generic EM clustering scheme that treats an image as an unordered set of pixels, Region Competition brings in useful geometric information and considers only hard assignments of observations to hypotheses.

The formal link we build relies on using the variational interpretation of EM to restrict the distributions considered during the minimization of (17) with respect to $Q_{n,k}$. Specifically, we consider only binary, winner-take-all [35] distributions over assignments. Denoting the set of observations that are assigned to hypothesis $k$ as $R_k = \{n : Q_{n,k} = 1\}$ the first term of (17) writes:

$$\sum_n \sum_k Q_{n,k} \log P(I_n|\mathcal{A}_k) = \sum_k \sum_{n \in R_k} \log P(I_n|\mathcal{A}_k) \quad (22)$$

which is a discretization of the area integral in (20).

Further, we can introduce the arclength penalty of (20) into our EM criterion by appropriately constructing the prior on the hidden variables, i.e. the second term in (8). For this we introduce a boolean function $b(\mathbf{z}_{\mathcal{N}_n})$ whose argument is the window of assignment vectors in the neighborhood $\mathcal{N}_n$ of $n$. $b$ indicates whether observations around $n$ are assigned to different hypotheses, i.e. if $n$ is on a boundary; we use $b$ to write the length-based prior

$$P(\mathbf{Z}) = \frac{1}{Z} \prod_n \exp(-b(\mathbf{z}_{\mathcal{N}_n})), \quad (23)$$

where $Z$ is a normalizing constant. We could also consider object specific terms, but we assume $P(\mathbf{Z}|\mathcal{A}) = P(\mathbf{Z})$ for simplicity. Since $Q$ is factorizable and $\sum_k Q_{n,k} = 1$, we have

$$\begin{aligned} -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \log P(\mathbf{Z}|\mathcal{A}) &= \sum_n \sum_k Q_{n,k} b(\mathbf{z}_{\mathcal{N}_n}) + c \\ &= \sum_n b(\mathbf{z}_{\mathcal{N}_n}) + c, \end{aligned}$$

which is, apart from the constant $c = \log Z$ a discretized version of the arc-length penalty used in Region Competition.

Finally, the entropy term $-\sum_{\mathbf{Z}} Q(\mathbf{Z}) \log Q(\mathbf{Z})$ of (17) generally favors smooth assignments of observations to the available hypotheses; since the Region Competition scheme by design assigns in a hard manner image observations to regions this term always equals zero and does not affect the EM bound. We note that we would end up with the same result if we set $\alpha = \infty$ in (17) from the start; then the entropy term would vanish and the optimal distributions would be binary.

Summing up we can see Region Competition as minimizing a version of (17) that utilizes specific expressions for $P(\mathbf{Z}|\mathcal{A})$ and $Q(\mathbf{Z})$. Even though mostly technical, this link allows us to use well studied segmentation algorithms in our system without straying from the original EM-based formulation.

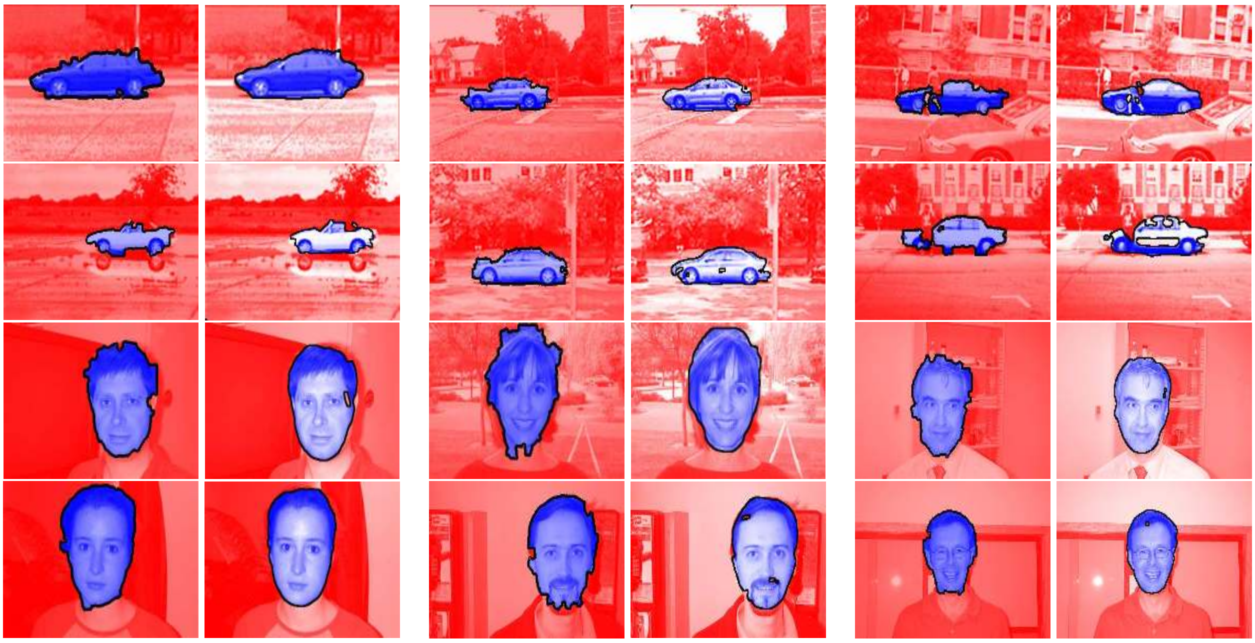


Fig. 7: Top-down segmentations of car and face images using fragment-based (left) and curve evolution-based (right) segmentation. For display, all background hypotheses are merged in a single region; For the fragment-based segmentation we threshold the E-step results at a fixed value. We observe that the curve evolution-based results provide smoother segmentations, that accurately localize object borders.

*2) AAMs as Shape Priors:* Coming to our case, the data fidelity terms for both the object and background hypotheses break into sums over the image grid, so they directly fit the setting of Region Competition. A variation stems from the $P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{O}})$ and $P(\mathbf{z}_{n,\mathcal{B}}|\mathcal{A}_{\mathcal{B}})$ terms that enforce prior information on the assignment probabilities. As mentioned in the previous section, $P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{O}})$ prevents the object from obtaining observations that do not fall within the template support; $P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{B}})$ can be a small constant, that acts as a penalty on the background model and helps the foreground model obtain observations more easily.

By taking into account the $P(\mathbf{z}_{n,\mathcal{O}}|\mathcal{A}_{\mathcal{O}})$ and $P(\mathbf{z}_{n,\mathcal{B}}|\mathcal{A}_{\mathcal{B}})$ terms we have the following evolution law for the front $\Gamma$ that separates the the object, $\mathcal{O}$ and the background $\mathcal{B}$ hypotheses:

$$\frac{\partial \Gamma}{\partial t} = -\mu\kappa\mathcal{N} + \log\frac{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{O}})P(\mathbf{z}_{n(\mathbf{x}),\mathcal{O}}|\mathcal{A}_{\mathcal{O}})}{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}})P(\mathbf{z}_{n(\mathbf{x}),\mathcal{B}}|\mathcal{A}_{\mathcal{B}})}\mathcal{N}$$
$$\overset{(15)}{=} \left[-\mu\kappa + \log\frac{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{O}})H(S^{-1}(\mathbf{x},\mathbf{s}))}{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}})} + \beta\right]\mathcal{N}.$$

Above $\beta = \log\frac{\pi_{\mathcal{O}}}{\pi_{\mathcal{B}_r}}$, $\mathbf{x}$ is an image location through which the front passes and $n(\mathbf{x})$ the corresponding observation index. The term $H(S^{-1}(\mathbf{x},\mathbf{s}))$ gates the motion due to the observation likelihood ratio term, $\log\frac{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{O}})}{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}})}$. Specifically, it lets the object compete only for observations that fall within its support, i.e. if $H(S^{-1}(\mathbf{x},\mathbf{s})) = 1$. Otherwise the observation is assigned to the background.

This constrains the object region to respect the shape properties of the corresponding category and introduces shape knowledge in the segmentation. Contrary to other works, such as [11], [43], this does not require additional shape prior terms but comes naturally from the AAM modelling assumptions.

Further, as in the previous subsection, we use a positive balloon force $\beta$ which favors the object region over the background.

We also use terms that result in improved segmentations, even if they do not stem from a probabilistic treatment. Specifically, as in [39], an edge-based term is utilized that pushes the segment borders towards strong intensity variations:

$$\frac{\partial \Gamma}{\partial t} = \Bigg[-\mu\kappa + \log\frac{P(I(\mathbf{x})|\mathbf{s},\mathbf{t})H(S^{-1}(\mathbf{x},\mathbf{s}))}{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}})}$$
$$+\beta - \nabla G(|\nabla I|)\cdot\mathcal{N}\Bigg]\mathcal{N}, \qquad (24)$$

where $G(|\nabla I|)$ is a decreasing function of edge strength $|\nabla I|$.

Curve evolution is implemented using level-set methods [37], [44] which are particularly well-suited for our problem; their topological flexibility allows holes to appear in the interior of regions, thereby excluding occluded object areas. Two competing background fronts are introduced, which form two large clusters for bright and dark regions. Initialization is random for all but the object fronts that are centered around the bottom-up detection results. Finally, we smooth $H$ with a Gaussian kernel of $\sigma = 2$ for stability.

In Fig. 7 where we compare the top-down segmentations offered by the two approaches, we observe that curve evolution yields superior results. The curvature term results in smooth boundaries, the edge force accurately localizes object borders, the shape of the objects is correctly captured, while occluded areas are discarded. Some partial failures, as e.g. the bottom-left car image can be attributed to the limited expressive ability of the AAM, that could not capture the specific illumination pattern. In that respect the modularity offered by the EM algorithm is an advantage, since any better generative model can be incorporated in the system once available.

## V. M-STEP - PARAMETER ESTIMATION

In the M-step the model parameters are updated to account for the observations assigned to the object during the E-step. The generative models we use assume a Gaussian noise process so that parameter estimation amounts to weighted least squares minimization, where the weights are provided by the E-step: higher weights are given to observations assigned with high confidence to the object and vice versa.

This approach faces occlusions by discounting them during model fitting. The typical AAM approach, e.g. [40] either considers occluded areas are known or utilizes a robust norm to reduce their effect on fitting. Instead, viewing AAMs in the generative model/EM setting tackles this problem by allowing alternative hypotheses to explain the observations, without modifying the AAM error norm.

### A. *EM-based AAM fitting Criterion*

In order to derive the update equations for the object parameters $\mathcal{A}_{\mathcal{O}} = (\mathbf{s}, \mathbf{t})$ we ignore the entropy-related term of the EM criterion (17) since it does not affect the final update. Further, the support-related term $H(S^{-1}(\mathbf{x}, \mathbf{s}))$ of (11) is hard to deal with inside the logarithm: it can equal zero and introduce infinite values in the optimized criterion. To avoid this we notice that any observation falling outside the support cannot be assigned to the object, by default. Therefore, we multiply the object weights delivered by the E-step with the indicator function which has the desired effect of taking the object support into account. The quantity maximized is thus:

$$C_{EM}(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{x}} E(\mathbf{x}) H(S^{-1}(\mathbf{x}; \mathbf{s})) \log P(I(\mathbf{x})|\mathcal{A}_{\mathcal{O}}) + \left(1 - E(\mathbf{x}) H(S^{-1}(\mathbf{x}; \mathbf{s}))\right) \log P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}}) \quad (25)$$

where $E(\mathbf{x}) = Q_{n(\mathbf{x}),\mathcal{O}}$ are the results of the previous E-step, obtained according to one of the two schemes in the previous section. Introducing the constant $c = \sum_{\mathbf{x}} \log P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}})$ and gathering terms we rewrite (25) as

$$C_{EM}(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{x}} E(\mathbf{x}) H(S^{-1}(\mathbf{x}; \mathbf{s})) \log \frac{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{O}})}{P(I(\mathbf{x})|\mathcal{A}_{\mathcal{B}})} + c. \quad (26)$$

Ignoring $c$, which is unaffected by the optimization of the foreground model and working on the template coordinate system this criterion writes:

$$C_{EM}(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{x}} E(\mathbf{x}_{\mathbf{s}}) H(\mathbf{x}) D(\mathbf{x}; \mathbf{s}) \log \frac{P(I(\mathbf{x}_{\mathbf{s}})|\mathcal{A}_{\mathcal{O}})}{P(I(\mathbf{x}_{\mathbf{s}})|\mathcal{A}_{\mathcal{B}})}, \quad (27)$$

where we introduce the notation $\mathbf{x}_{\mathbf{s}} = S(\mathbf{x}; \mathbf{s})$. Since the deformation $\mathbf{x} \rightarrow S(\mathbf{x})$ locally rescales the template domain, the determinant of its Jacobian, $D(\mathbf{x}; \mathbf{s})$, commeasures (26),(27) which are viewed as discretizations of area integrals. Finally, modeling both the fore- and background reconstruction errors as a white Gaussian noise process we write (27) as:

$$C_{EM}(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{x}} E(\mathbf{x}_{\mathbf{s}}) H(\mathbf{x}) D(\mathbf{x}; \mathbf{s}) \left[ (I(\mathbf{x}_{\mathbf{s}}) - T(\mathbf{x}, \mathbf{t}))^2 - (I(\mathbf{x}_{\mathbf{s}}) - B(\mathbf{x}_{\mathbf{s}}))^2 \right], \quad (28)$$

where $T$ is the object-based synthesis, and $B$ is the image reconstruction using the background model. The multiplicative factor from the standard deviation of the noise process is omitted, since it does not affect the final parameter estimate.

The standard, least squares, AAM criterion of (11) can be transcribed using this notation as:

$$C_{LS}(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{x}} H(\mathbf{x}) \left(I(\mathbf{x}_{\mathbf{s}}) - T(\mathbf{x}, \mathbf{t})\right)^2 . \quad (29)$$

Comparing (28) to (29) we observe three main deficiencies of the latter: First, the segmentation information of $E(\mathbf{x}_{\mathbf{s}})$ is discarded, forcing the model to explain potentially occluded areas. Second, the fidelity of the foreground and background models to the data are not compared; in the absence of strong edges this leads to mismatches of the image and model boundaries. Third, the magnification or shrinking of template points due to the deformation is ignored, while it is formally required by the generative model approach.

### B. *Shape fitting equations*

In the following we provide update rules for AAM fitting going from (29) to (28), by gradually introducing more elaborate terms. As in [30] we derive the optimal update based on a quadratic approximation to the cost; we provide details in App. I.

Perturbing the shape parameters by $\Delta\mathbf{s}$ we have:

$$I(S(\mathbf{x}; \mathbf{s} + \Delta\mathbf{s})) \simeq I(S(\mathbf{x}; \mathbf{s})) + \sum_{i=1}^{N_S} \frac{dI}{d\mathbf{s}}_i(\mathbf{x}; \mathbf{s}) \Delta\mathbf{s}_i \quad (30)$$

$$\frac{dI}{d\mathbf{s}}_i(\mathbf{x}; \mathbf{s}) = \frac{\partial I(S(\mathbf{x}; \mathbf{s}))}{\partial x} \frac{\partial S_x}{\partial \mathbf{s}_i} + \frac{\partial I(S(\mathbf{x}; \mathbf{s}))}{\partial y} \frac{\partial S_y}{\partial \mathbf{s}_i}, \quad (31)$$

where $N_S$ the number of shape basis elements. To write (30) concisely we consider raster scanning the image whereby $I$ becomes a $N \times 1$ vector, where $N$ is the number of observations, $\frac{dI}{ds}$ becomes a $N \times N_S$ matrix, while $\Delta\mathbf{s}$ is treated as a $N_S \times 1$ vector. We can thus write (30) as:

$$\mathbf{I}(\mathbf{s} + \Delta\mathbf{s}) = \mathbf{I}(\mathbf{s}) + \frac{\mathbf{dI}}{\mathbf{ds}} \Delta\mathbf{s}, \quad (32)$$

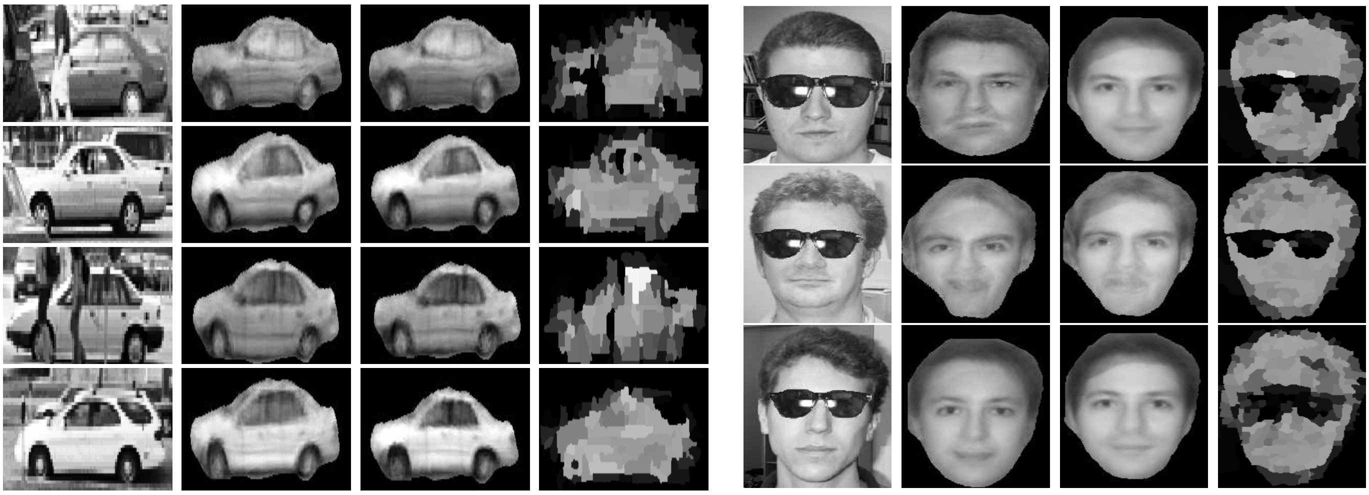
where $\mathbf{I}(\mathbf{s})$ denotes the vector formed by raster scanning $I(S(\mathbf{x}; \mathbf{s}))$; this is a notation we use in the following for all quantities appearing inside the criteria being optimized. For simplicity we also omit the $\mathbf{s}$ argument from $\mathbf{I}(\mathbf{s})$.

To write the quadratic approximation to the perturbed cost $C_{LS}(\mathbf{s} + \Delta\mathbf{s}, \mathbf{t})$ we introduce $\mathcal{E} = \mathbf{I} - \mathbf{T}$ and denote by $\circ$ the Hadamard product, $(a_{ij}) \circ (b_{ij}) = (a_{ij} b_{ij})$. We thereby write:

$$C_{LS}(\mathbf{s} + \Delta\mathbf{s}, \mathbf{t}) = C_{LS}(\mathbf{s}, \mathbf{t}) + \mathcal{J} \Delta\mathbf{s} + \frac{1}{2} \Delta\mathbf{s}^T \mathcal{H} \Delta\mathbf{s},$$

$$\mathcal{J} = 2 \left[\mathbf{H} \circ \mathcal{E}\right]^T \frac{\mathbf{dI}}{\mathbf{ds}}, \quad \mathcal{H} = 2 \left(\mathbf{H} \circ \frac{\mathbf{dI}}{\mathbf{ds}}\right)^T \frac{\mathbf{dI}}{\mathbf{ds}}, \quad (33)$$

where $\mathcal{J}$ is the Jacobian of the cost function, and $\mathcal{H}$ its Hessian. For terms like $\mathbf{H} \circ \frac{\mathbf{dI}}{\mathbf{ds}}$ where $\mathbf{H}$ is $N \times 1$ and $\frac{\mathbf{dI}}{\mathbf{ds}}$ is $N \times N_S$, $\mathbf{H}$ is replicated $N_S$ times horizontally. From (33) we get the update of the forward additive method [30]: $\Delta\mathbf{s}^* = -\left[\mathcal{J}\mathcal{H}^{-1}\right]^T$.



(a) Input (b) Plain LS Fit (c) EM-based Fit (d) E-step (a) Input (b) Plain LS Fit (c) EM-based Fit (d) E-step

Fig. 8: Differences in AAM fitting using the EM algorithm: (a) Input image, (b) plain least squares (LS) fit, (c) EM-based fit and (d) E-step results. The EM-based fit outperforms the typical LS fit as the E-results robustify the AAM parameter estimation. This is accomplished by discounting occlusions or areas with unprecedented appearance variations, such as the third window and the hair fringe in the bottom row.

Further, introducing the E-step results yields the criterion:

$$\sum_{\mathbf{x}} E(\mathbf{x_s})H(\mathbf{x})\left(I(\mathbf{x_s}) - T(\mathbf{x},\mathbf{t})\right)^2 = \left[\mathbf{E}\circ\mathbf{H}\circ\mathcal{E}\right]^T\mathcal{E} \quad (34)$$

for which the Jacobian and Hessian matrices become:

$$\mathcal{J} = 2\left(\mathbf{H}'\circ\mathcal{E}\right)^T\frac{\mathbf{dI}}{\mathbf{ds}} + \mathcal{E}^T\left(\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathbf{H}\circ\mathcal{E}\right) \quad (35)$$

$$\mathcal{H} = 2\left[\mathbf{H}'\circ\frac{\mathbf{dI}}{\mathbf{ds}} + 2\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathbf{H}\circ\mathcal{E}\right]^T\frac{\mathbf{dI}}{\mathbf{ds}} \quad (36)$$

where $\mathbf{H}' = \mathbf{E}\circ\mathbf{H}$. Multiplication with $\mathbf{E}$ forces the fitting scheme to lock onto the areas assigned to the object and results in the new terms $\mathcal{E}^T(\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathbf{H}\circ\mathcal{E})$, $2\left(\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathbf{H}\circ\mathcal{E}\right)^T\left(\frac{\mathbf{dI}}{\mathbf{ds}}\right)$. These account for the change caused by $\Delta\mathbf{s}$ in the probability of assigning observations to template points.

A more elaborate expression results from incorporating the deformation's Jacobian in the update; as it does not critically affect performance we only report it in App. I.

Finally, we consider the reconstruction error of the background model, $\mathcal{E}_B = \mathbf{I} - \mathbf{B}$, where $\mathbf{B}$ is the matrix formed by raster-scanning the background synthesis $B(\mathbf{x_s})$. We thus obtain the cost function and Jacobian and Hessian matrices for the original EM criterion (28):

$$C_{EM}(\mathbf{s},\mathbf{t}) = \left[(\mathbf{H}\circ\mathbf{E})\circ\mathcal{E}\right]^T[\mathcal{E}] - \left[(\mathbf{H}\circ\mathbf{E})\circ\mathcal{E}_B\right]^T[\mathcal{E}_B] \quad (37)$$

$$\mathcal{J} = \mathcal{J}_\mathcal{E} - \mathcal{J}_{\mathcal{E}_B}, \mathcal{H} = \mathcal{H}_\mathcal{E} - \mathcal{H}_{\mathcal{E}_B}, \quad (38)$$

where $\mathcal{J}_\mathcal{E}, \mathcal{H}_\mathcal{E}$ are as in (35),(36) and $\mathcal{J}_{\mathcal{E}_B}, \mathcal{H}_{\mathcal{E}_B}$ are their background model counterparts. Since the minimized term is no longer convex instabilities may occur. An optimal scaling of the update vector is therefore chosen with bisection search, starting from one.

### *C. Appearance fitting equations*

The appearance parameters are estimated by considering the part of the EM criterion that depends on the model prediction:

$$C_{EM}(\mathbf{s},\mathbf{t}) = \sum_{\mathbf{x}} W(\mathbf{x})\left[I(\mathbf{x_s}) - \mathcal{T}_0(\mathbf{x}) - \sum_{i=1}^{N_\mathcal{T}}\mathbf{t}_i\mathcal{T}_i(\mathbf{x})\right]^2 \quad (39)$$

where $N_\mathcal{T}$ is the number of appearance basis elements and $W(\mathbf{x})$ combines all scaling factors: $W(\mathbf{x}) = D(\mathbf{x};\mathbf{s})H(\mathbf{x})E(S(\mathbf{x};\mathbf{s}))$. This yields the weighted least squares error solution:

$$\mathbf{t}^* = \left[\left[\mathbf{W}\circ(\mathbf{I}-\mathbf{T_0})\right]^T\mathbf{T}\right]\left[\mathbf{T}^T(\mathbf{W}\circ\mathbf{T})\right]^{-1} \quad (40)$$

where $\mathbf{T}$ is the $N\times N_\mathcal{T}$ array formed by the appearance basis elements.

Finally, a prior distribution learned during model construction is introduced in the updates of both the $\mathbf{s}$ and $\mathbf{t}$ parameters. For an independent Gaussian distribution the Jacobian and Hessian matrices are modified as:

$$\mathcal{J}'_i = \mathcal{J}_i + \lambda\frac{p_i}{\sigma_i^2}, \quad \mathcal{H}'_{i,i} = \mathcal{H}_{i,i} + \lambda\frac{1}{\sigma_i^2}, \quad (41)$$

where $i$ ranges over the number of parameter vector elements, $p_i$ is the $i$-th element of the parameter estimate at the previous iteration, $\sigma_i$ its standard deviation on the training set and $\lambda$ controls the tradeoff between prior knowledge and data fidelity.

The improvements in fitting quality attained with the EM-based scheme are shown in Fig. 8. These examples either have actual occlusions, or locally have appearances that cannot be extrapolated from the training set. The plain least squares criterion of (29) is forcing the model to explain the whole of its interior, and therefore results in a suboptimal fit.

Instead, in the EM-based setting, even though the AAM predicts the appearance for the whole object domain, certain regions may not get assigned to the model if its prediction there does not match the image observations. As the lower values of the E-step results reveal, the model is thereby freed from explaining occluded regions.

The price to pay for this increased flexibility is that informative areas like nostrils, teeth, etc. may be discounted if not modeled adequately well. Still, as the following section shows, the robustness of the estimated parameters is in practice more important for the detection task.

## VI. Synergetic Object Category Detection

Our goal in this section is to explore how the synergy between segmentation and recognition improves detection performance. This is a less explored side of the bottom-up/top-down idea compared to top-down segmentation and as we show with the object categories of faces and cars, it is equally practical and useful.

### A. Detection Strategy

*1) Bottom-Up Detection:* We use a front-end object detection system to provide us with all object hypotheses by setting its rejection threshold to a conservative value. As in [45], we treat these detections as proposals that are pruned via the bottom-up/top-down loop. We rely on the point-of-interest based system of [22], which represents objects in terms of a codebook of primal sketch features. This system builds object models by clustering blobs, ridges and edges extracted from the training set and then forming a codebook representation. During detection the extracted features propose object locations based on their correspondences with the codebook entries. Since any other bottom-up system could be used instead of this one, we refer to [22] for further details as well as to related literature on this quickly developing field, e.g. [1], [7], [14], [25], [50].

*2) Top-Down & Bottom-Up combination:* For object detection we complement the bottom-up detection results with information obtained by the parameters of the fitted AAM models and the segmentation obtained during the E-step, as illustrated in Fig. 4. We thus have three different cues for detection: first, the bottom-up detection term $C_{BU}$ quantifies the likelihood of interest point features given the hypothesized object location [22].

Second, the AAM parameters are used to indicate how close the observed image is to the object category. We model the AAM parameter distributions as Gaussian density functions, estimated separately on foreground and background locations during training. We thereby obtain a simple classifier:

$$C_{AAM} = \log \frac{P(\mathbf{s}, \mathbf{t}|\mathcal{O})}{P(\mathbf{s}, \mathbf{t}|\mathcal{B})}, \tag{42}$$

that decides about the presence of the object based on the estimated AAM parameters.

Third, we quantify how well the object hypothesis predicts the image data using the E-step results that give the probability $E(\mathbf{x})$ of assigning observation $\mathbf{x}$ to the object. We build the segmentation-based classifier by computing the average of $E(\mathbf{x})$ over the area that can be occupied by the object:

$$C_{SEG} = \frac{\sum_{\mathbf{x}} H(S^{-1}(\mathbf{x}))E(\mathbf{x})}{\sum_{\mathbf{x}} H(S^{-1}(\mathbf{x}))}. \tag{43}$$

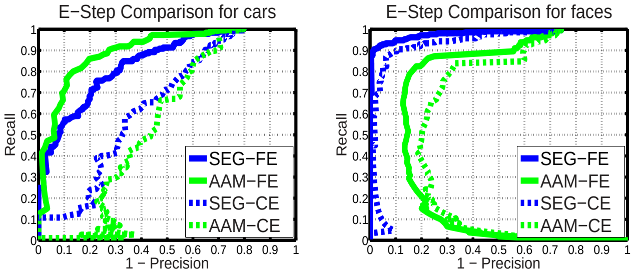


Fig. 9: Comparison of the Curve Evolution-based E-step (CE) and the Fragment-based E-step (FE) based on the detection of faces [14] and cars [1]. For both categories the classifiers using segmentation ('SEG') and AAM parameter ('AAM') information perform better if the Fragment-based E-step is used.

The summation is over the whole image domain, and $H(S^{-1}(\mathbf{x}))$ indicates whether $\mathbf{x}$ can belong to the object. Using the E-step results in this way prunes false positives, around which the AAM cannot explain a large part of the image, thereby resulting in a low value of $C_{SEG}$.

We combine the three classifiers using the supra-Bayesian fusion setting [19]. The output $C_k$ of classifier $k$ is treated as a random variable, following the distributions $P(C_k|\mathcal{O}), P(C_k|\mathcal{B})$ under the object and background hypotheses, respectively. Considering the set of classifier outputs as a vector of independent random variables, $\mathbf{C} = (C_1, \ldots, C_k)$ we use their individual distributions for classifier combination:

$$\frac{P(\mathcal{O}|\mathbf{C})}{P(\mathcal{B}|\mathbf{C})} = c\frac{P(\mathbf{C}|\mathcal{O})}{P(\mathbf{C}|\mathcal{B})} = c\prod_{k=1}^{K} \frac{P(C_k|\mathcal{O})}{P(C_k|\mathcal{B})} \tag{44}$$

where $c = P(\mathcal{O})/P(\mathcal{B})$.

### B. Experimental Results

*1) Performance Evaluation and Experimental Settings:* We use Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves to evaluate a detector: ROC curves consider deciding whether an image contains an object, irrespective of its location. PR curves evaluate object localization, comparing the ratio $R$ of retrieved objects (recall) to the ratio $P$ of correct detections (precision); both curves can be summarized using their Equal Error Rate, namely the point where the probability of a false hit equals the probability of a missed positive.

In order to compare our results with prior work, we have used the setup of [14] for faces and that of [1] for cars. Cars are rescaled by a factor of 2.5, and flipped to have the same direction during training, while faces are normalized so that the eye distance is 50 pixels; a $50 \times 30$ box is used to label a detected face a true hit. Further, nonmaximum suppression is applied on the car results as in [1], allowing only the strongest hypothesis to be active in a $100 \times 200$ box.

Regarding system tuning, we determine the parameters that influence segmentation and model fitting using a few images from the training set of each category; during testing we use the same parameter choices for both categories, on all of the subsequent detection tasks.

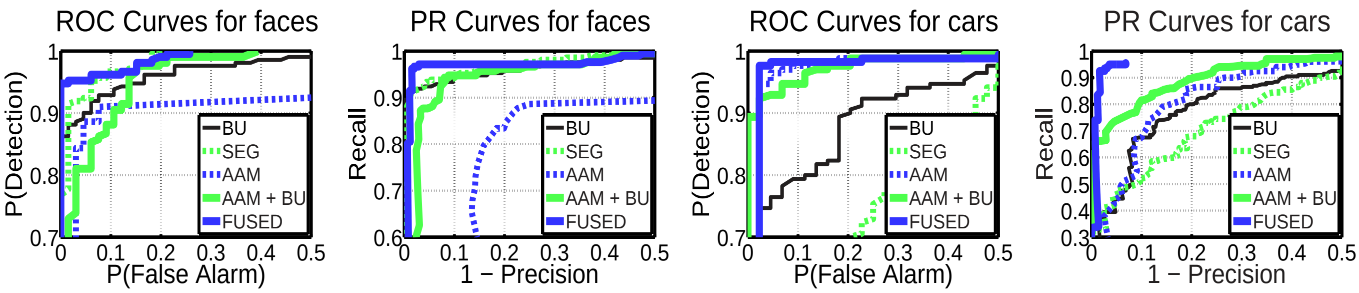


Fig. 10: Performance of the individual and combined classifiers on the face [14] and car [1] datasets. *Please note the ranges of the axes; preferably see in color.* The individual classifiers use bottom-up ('BU'), AAM parameter ('AAM') and segmentation ('SEG') cues. The others combine different cues using (44): the 'AAM+BU' classifier uses AAM and bottom-up information while the 'FUSED' classifier combines the AAM, BU and segmentation cues. From both PR and ROC curves we see that the 'FUSED' classifier outperforms the rest.

*2) Comparison of Alternative E-Step Implementations:* We initially compare the Fragment-based E-step (FE) and Curve Evolution-based E-step (CE) approaches in terms of their appropriateness for object detection. Specifically, we have applied the EM approach to both object categories considered, using identical settings for the detection front-end, the EM system components and the classifier combination.

In Fig. 9 we provide the Precision Recall curves of the individual $C_{SEG}$ and $C_{AAM}$ classifiers for the two approaches. We observe that the CE approach is outperformed by the FE approach on the detection task. In our understanding this is because the CE approach makes a hard assignment using local information while the FE approach takes soft decisions and uses the information within a whole image fragment. We note that the CE approach uses a balloon force in (24) that largely influences the performance of the segmentation-based classifier; we therefore experimented with different values and present the best results we obtained.

Since the FE approach performs systematically better on the detection task we use it for the subsequent, more detailed detection experiments. The CE approach could be used after a decision has been made about the presence of an object as it provides more appealing top-down segmentations by enforcing smoothness and drawing boundaries close to image edges. In this setting, the thresholded FE results could serve for initialization.

*3) Joint Bottom-Up and Top-Down Object Detection:* In Fig. 10 we provide PR and ROC curves for the different detectors operating in isolation and their combinations according to the combination rule (44). Even though the bottom-up detector is already performing well, the individual detectors behave in a complementary manner in different areas and their combinations yield systematically improved results. In specific we note that the car dataset is harder than the face dataset, at least for bottom-up detection; still, the final classifier fusing bottom-up and top-down cues performs equally well for both categories.

Equal Error Rates

| Method | Cars | Faces |
|---|---|---|
| Ours | 5.5 | 4.7 |
| Fergus [14] | 11.6 | 8.3 |
| Leibe [25] | 3.0 | - |
| Opelt [36] | 17.0 | 6.5 |

TABLE I: EER of our system compared to that of other researchers on the same datasets; for cars we report the Precision-Recall EER measurement, as the other references.

Comparing our results to those reported by others in Table I we observe that our system is outperformed only by that of [25] on the car dataset. However, our bottom-up system [22] uses 80 codebook clusters and is significantly simpler than that of [25], where more than 500 codebook entries are used. Further, our top-down validation stage takes approximately 2 sec. per hypothesis, which is approximately two orders of magnitude less than that of [25]. We should note here that flipping the car images during training and fixing the scale of the faces may have introduced some small positive bias in favor of our method. We consider it however more important that systematic improvement in performance is obtained by combining top-down and bottom-up information via the EM algorithm.

After validating the usefulness of top-down information we address the question whether the joint treatment of the two tasks is really necessary. One particular concern has been whether this improvement is exclusively due to the AAM classifier; if this is so, this would render the EM approach superfluous for detection. The first answer comes from comparing the results obtained by combining all cues ('Fused') with the ones using only the AAM and Bottom-Up classifiers. For both cases considered we observe an improvement both in the ROC and PR curves, which is due to the additional information provided by the Segmentation-based classifier. Still, what we consider more important and now demonstrate is that EM allows for the use of generative models in hard images, by discounting image variation that was not observed in the training set.

*4) Occluded Object Detection:* We argue here that segmentation helps obtain robust estimates of the model parameters, and thereby supports the performance of the AAM classifier in images where the objects are occluded. Since all objects are fully observed in the dataset of [14], this point cannot be clearly made for faces; we therefore repeat the previous classifier combination experiment after artificially adding sunglasses

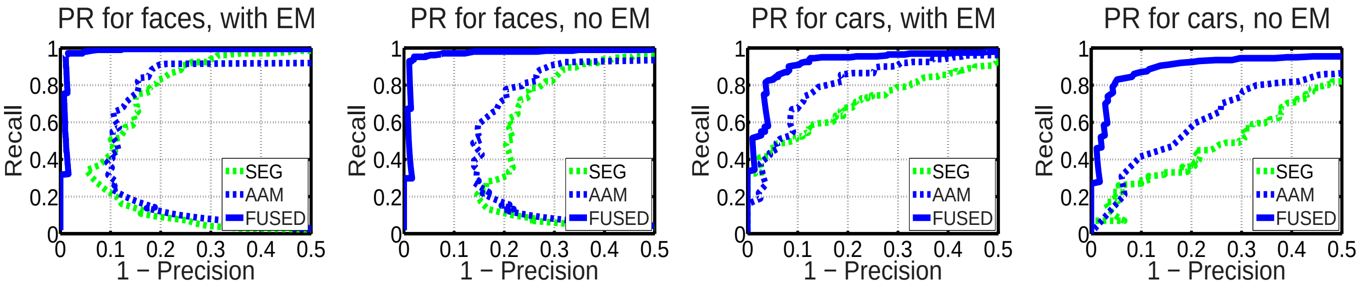


Fig. 11: Influence of using the EM algorithm on detection performance: Segmentation helps exclude occluded areas from AAM fitting and results in more robust parameter estimates. This is reflected in improved performance of both the individual AAM and the FUSED classifier that combines the bottom-up, segmentation and AAM classifiers.

to the faces as in Fig. 8. To deconvolve evaluation, we assume the bottom-up detection system is insensitive to occlusion, and use the results it furnishes with the fully observed images. For cars there are substantial occlusions in the test set, so we use the original images.

The top-down classifiers are evaluated in two distinct scenarios, by (a) ignoring segmentation and setting the segmentation weights $E(\mathbf{x})$ to one everywhere and (b) using results furnished by EM. The input to the segmentation-based classifier in the first case is obtained by fitting the AAM with $E(\mathbf{x}) = 1$, and using after convergence the fitted AAM parameters to estimate $E(\mathbf{x})$ anew. It is this $E(\mathbf{x})$ that is then used in (43).

As shown in Fig. 11, the parameter estimates derived in scenario (a) yield significantly worse performance, since the occluded region affects AAM fitting, while in scenario (b) performance degrades gracefully. This behavior is propagated to the fused classifier performance, where in scenario (b) consistently better performance is observed.

These results indicate the importance of a joint treatment of the segmentation and detection tasks in the practical situation where faces are occluded by glasses, scarfs, beards or cars are occluded by pedestrians, boxes, etc. The gain is not only due to the validation information offered by a top-down segmentation, but also due to the robust model fitting that sustains the performance of the classifier that uses the AAM parameters.

## VII. Survey and Discussion

Herein we briefly discuss and compare related work on this relatively new problem, to place our contributions in a broader context.

### A. *Previous work on joint detection and segmentation*

We can classify most of the existing works on joint segmentation and detection based on whether they use global or part-based object representations. Global approaches [21], [45], [51] assume that a monolithic object model can account for the shape and appearance variation of the object category, and thereby take hold of all the image observations. Part-based models such as [7], [24], [25], [27], [52] offer a modular representation that is used to build the top-down segmentation in a hybrid fashion, using high-level information wherever available, and low-level cues to bring the rest of the object together [24], [52].

At a more detailed level, the approach of [45], [54] performs a stochastic search in the space of regions and hypotheses, by extending the Data-Driven MCMC scheme of [46] to include global generative models for object categories. During search object and generic region hypotheses are generated, merged, split or discarded while their borders are localized by curve evolution using Region Competition [53]. Even though this approach is elegant, in a practical object detection application one typically only needs the probability of an object being present, which as we show here can be efficiently and reliably estimated using the observation window containing the object and EM instead of stochastic search.

Following a non-generative approach, codebook representations are used for joint detection and segmentation in [6], [7] and [25]. Figure-ground maps associated with the codebook entries are stored during training and used during detection to assemble a segmentation of the object. Even though good performance is demonstrated in [6], [25], the segmentation depends on the ability to cover a large area of the object using overlapping patches, necessitating complex models. In another approach using a part-based representation in [52] an object-sensitive affinity measure is introduced, and pairwise clustering methods are used to find a global minimum of the data partitioning cost. The affinity measure used leads to a grouping of pixels based on both low-level cues (absence of edges, similarity) and high-level knowledge. However, the lack of a probabilistic interpretation impedes the cooperation with other processes while the detection task is not considered.

Coming to work involving the EM algorithm, we note first that the use of the EM algorithm for image segmentation problems is certainly not novel; it has been used previously for low-level problems such as feature-based image segmentation [3] or layered motion estimation [49]. Further, in [24] a part-based object representation is combined with the graph-cut algorithm to derive a top-down segmentation, yielding accurate results for articulated object segmentation. The EM algorithm is used there as well, but in an optimization rather than a generative model fitting task: the shape parameters are treated as hidden variables and the E-step constructs a nonparametric shape distribution. The M-step then amounts to the maximization via graph cuts of a segmentation quality cost that entails the distribution constructed in the E-step. This

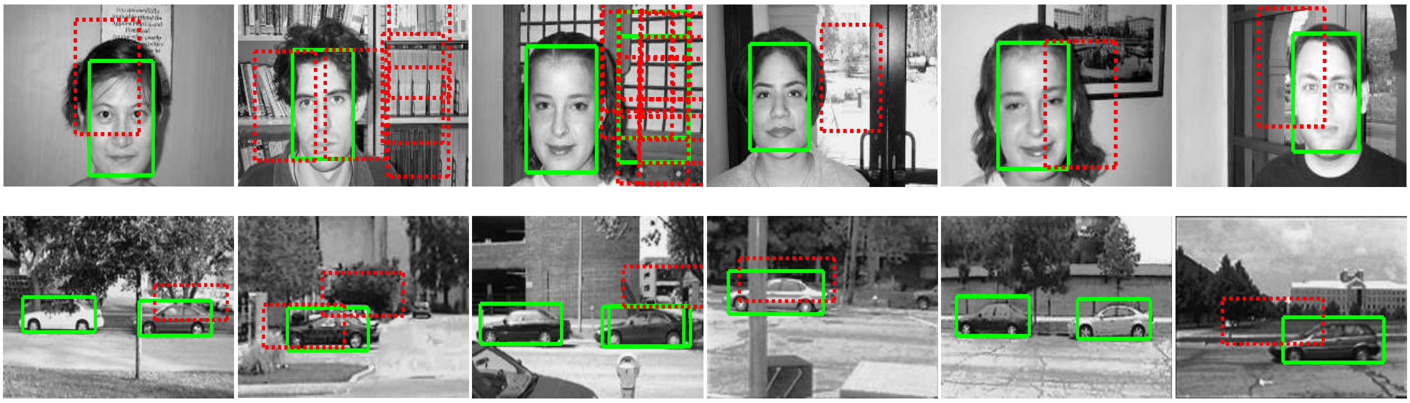


Fig. 12: Sample detection results on the datasets of [1], [14]. The locations proposed by a bottom-up detection system are used to initialize an EM-loop which brings in additional information from segmentation and the AAM parameters. Based on these, the initial hypotheses are either pruned (red dashed boxes) or validated (green boxes).

deviates from the use of EM in the generative model setting, where parameter estimation is accomplished in the M-step. As we show here, the generative model approach allows the principled combination of different methodologies, like curve evolution and AAMs, while minimizing the choices that a generic optimization approach requires.

Further, in the work of [16], [51] the EM algorithm is used to perform an object-specific segmentation of an image using the 'sprites & layers' model where the E-step assigns observations to objects ('sprites') and the M-step updates the object parameters. Intuitively this approach is similar to ours, but the interaction of the two processes is not explored: The background model is estimated from a fixed set of images, thereby introducing strong prior knowledge that is not available for the general segmentation problem, while it is not actually determined whether an object is present in the image, based on either bottom-up or top-down cues. Further, the deformations used do not model the object category shape variation, since they are either restricted to affine transformations [16] or use an MRF prior on a piecewise constant deformation field [51].

### B. Previous Work on Shape Prior segmentation

Complementary to research on top-down/bottom-up integration, progress has been made during the last years in the use of object-specific shape prior knowledge for image segmentation e.g. in [10], [11], [26], [42], [43]. By focusing on the object boundaries these approaches efficiently exploit shape knowledge to overcome noise and partial occlusions.

Most shape prior methods rely on the implicit representation of level-set methods, where a curve is represented in terms of an embedding function, such as the distance transform. This allows for a convenient combination with curve evolution methods: the variational criterion driving the segmentation is augmented with a shape-related penalty term that is expressed in terms of the embedding function of the evolving curve. This allows for the combination of shape-based and image-based cues in a common curve evolution framework.

Even though such methods do not model object aspects like appearance or deformation statistics, they have been proven particularly effective in tasks such as medical image segmentation [42] or tracking a detected person [10]. On the one hand this can be seen as an advantage, since less demanding object models are used, on the other we believe that they do not provide a complete solution to the bottom-up/top-down combination problem.

Specifically, part of the object may be occluded so the boundaries of the object and the region assigned to it do not have to be related. For example, if we consider a person wearing a hat, or sunglasses, a shape prior-driven segmentation will force the curve corresponding to the object hypothesis to include the occluded parts of the head, as most heads are roughly ellipsoidal. Even though one can argue that this indicates robustness to occlusion, in our understanding, a top-down segmentation should indicate the image regions occupied by an object. This can be accomplished with our approach, where a generative model like an AAM can still fit the shape of the object, but in the E-step the occluded parts are not assigned to the object.

We should mention that the shape prior-based technology has made advances in a broader range of problems, like articulated object tracking and tracking under severe occlusion using limited appearance information, cf. e.g. [10], so this added functionality of our system can be seen as being complementary. However, the EM/generative model approach has no fundamental limitation in addressing these problems as well. Part-based deformation models can be used for articulated objects, while temporal coherence for tracking can be enforced by using a dynamical model for the generative model parameters. Having proved the merit of the EM approach on a more constrained problem, we would like to explore these more challenging directions in future research.

## VIII. Conclusions - Future Work

In this paper we have addressed the problem of the joint segmentation and analysis of objects, casting it in the framework of the Expectation-Maximization algorithm. Apart from a concise formulation of bottom-up/top-down interaction, this has facilitated the principled combination of different computer vision techniques. Based on the EM algorithm we have built a system that can segment in a top-down manner images of objects belonging to highly variable categories, while also significantly improving detection performance. Summing up, the EM framework offers a probabilistic formulation for a recently opened problem and deals with its multifaceted nature in a principled manner.

An essential direction for rendering this approach applicable to a broader set of problems is the automated construction of models for generic objects; recent advances [7], [14], [50] have initiated a surge of activity on simple part-based representations, e.g. [1], [22], [25] but little work has been done for global generative models [23], [51]. Further, a point that deserves deeper inspection is the combination of low-level cues with part-based and global generative models for joint object segmentation, which has only partially been tackled [22], [24], [52]. In future work we intend to address these issues in the framework of generative models with the broader goal of integrating different computer vision problems in a unified and practical approach.

## ACKNOWLEDGEMENTS

We thank G. Papandreou for discussions on AAM fitting and for indicating related references. We wish to thank the Reviewers for their constructive comments that helped improve the quality of the paper.

## APPENDIX I
## DERIVATION OF THE EM/AAM UPDATE RULES

Using the notation introduced in Sec. V, the Jacobian and Hessian in the typical update are obtained by approximating the perturbed cost function (29) as:

$$
\begin{aligned}
C_{LS}(\mathbf{s}+\Delta\mathbf{s},\mathbf{t}) &= \sum_{\mathbf{x}} H(\mathbf{x})(I(\mathbf{x};\mathbf{s}+\Delta\mathbf{s}) - T(\mathbf{x},\mathbf{t}))^2 \\
&\simeq \left[\mathbf{H}\circ\left(\mathbf{I}+\frac{\mathbf{dI}}{\mathbf{ds}}\Delta\mathbf{s}-\mathbf{T}\right)\right]^T\left[\mathbf{I}+\frac{\mathbf{dI}}{\mathbf{ds}}\Delta\mathbf{s}-\mathbf{T}\right] \\
&= [\mathbf{H}\circ\mathcal{E}]^T\mathcal{E} + 2\,[\mathbf{H}\circ\mathcal{E}]^T\frac{\mathbf{dI}}{\mathbf{ds}}\Delta\mathbf{s} + \Delta\mathbf{s}^T\left[\mathbf{H}\circ\frac{\mathbf{dI}}{\mathbf{ds}}\right]^T\frac{\mathbf{dI}}{\mathbf{ds}}\Delta\mathbf{s},
\end{aligned}
$$

where in the third line we use $\mathcal{E} = \mathbf{I} - \mathbf{T}$. We thereby get the expressions in (33). The criterion $[\mathbf{H}\circ\mathbf{E}\circ\mathcal{E}]^T[\mathcal{E}]$ in (34) incorporates segmentation information and its perturbation is written as:

$$
\left[\mathbf{H}\circ\left(\mathbf{E}+\frac{\mathbf{dE}}{\mathbf{ds}}\Delta\mathbf{s}\right)\circ\left(\mathcal{E}+\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}}\Delta\mathbf{s}\right)\right]^T\left(\mathcal{E}+\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}}\Delta\mathbf{s}\right).
$$

Keeping the first and second order product terms we have:

$$
\begin{aligned}
\mathcal{J} &= 2\,[\mathbf{H}\circ\mathbf{E}\circ\mathcal{E}]^T\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}} + \mathcal{E}^T\left[\mathbf{H}\circ\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathcal{E}\right] \\
\mathcal{H} &= 2\left[\mathbf{H}\circ\mathbf{E}\circ\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}} + 2\mathbf{H}\circ\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathcal{E}\right]^T\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}}.
\end{aligned}
$$

These are identical to the expressions in (35,36), as $\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}} = \frac{\mathbf{dI}}{\mathbf{ds}}$.

To incorporate the determinant of the deformation's Jacobian we express it using the shape synthesis relation of (9):

$$
D(\mathbf{x};\mathbf{s}) = \sum_k \mathbf{s}_k\frac{\partial\mathcal{S}_{k,x}}{\partial x}\sum_j \mathbf{s}_j\frac{\partial\mathcal{S}_{j,y}}{\partial y} - \sum_k \mathbf{s}_k\frac{\partial\mathcal{S}_{k,y}}{\partial x}\sum_k \mathbf{s}_y\frac{\partial\mathcal{S}_{k,x}}{\partial y} \tag{45}
$$

In matrix notation, $\mathbf{D}(\mathbf{s}) = [\mathbf{s}S_x^x]\circ[\mathbf{s}S_y^y] - [\mathbf{s}S_y^x]\circ[\mathbf{s}S_x^y]$. We can write the following linear approximation to $\mathbf{D}(\mathbf{s}+\Delta\mathbf{s})$:

$$
\mathbf{D}(\mathbf{s}+\Delta\mathbf{s}) = \mathbf{D}(\mathbf{s}) + \frac{\mathbf{dD}}{\mathbf{ds}}\Delta\mathbf{s} + O(\Delta\mathbf{s}^2), \tag{46}
$$

where $\frac{\mathbf{dD}}{\mathbf{ds}} = S_x^x\circ\mathbf{s}S_y^y + S_y^y\circ\mathbf{s}S_x^x + S_y^x\circ\mathbf{s}S_x^y + S_x^y\circ\mathbf{s}S_y^x$.

The perturbed cost and the Jacobian and Hessian obtained by retaining first- and second- order terms then become:

$$
\left[\mathbf{H}\circ\left(\mathbf{D}+\frac{\mathbf{dD}}{\mathbf{ds}}\Delta\mathbf{s}\right)\circ\left(\mathbf{E}+\frac{\mathbf{dE}}{\mathbf{ds}}\Delta\mathbf{s}\right)\circ\left(\mathcal{E}+\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}}\Delta\mathbf{s}\right)\right]^T\left(\mathcal{E}+\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}}\Delta\mathbf{s}\right) \tag{47}
$$

$$
\begin{aligned}
\mathcal{J} &= 2\,[\mathbf{H}'\circ\mathbf{D}\circ\mathcal{E}]^T\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}} + \mathcal{E}^T\left[\mathbf{H}\circ\mathbf{D}\circ\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathcal{E}\right] + \mathcal{E}^T\left[\mathbf{H}'\circ\frac{\mathbf{dD}}{\mathbf{ds}}\circ\mathcal{E}\right] \\
\mathcal{H} &= 2\left[\mathbf{H}'\circ\mathbf{D}\circ\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}} + 2\mathbf{H}\circ\mathbf{D}\circ\frac{\mathbf{dE}}{\mathbf{ds}}\circ\mathcal{E} + 2\mathbf{H}'\circ\frac{\mathbf{dD}}{\mathbf{ds}}\circ\mathcal{E}\right]^T\frac{\mathbf{d}\mathcal{E}}{\mathbf{ds}}
\end{aligned}
$$

## REFERENCES

[1] S. Agarwal and D. Roth, "Learning a Sparse Representation for Object Detection," in *Proc. Eur. Conf. on Comp. Vision*, 2002.

[2] A. Barbu and S. C. Zhu, "Graph partition by Swendsen-Wang cuts," in *Proc. Int.l Conf. on Comp. Vision*, 2003.

[3] S. Belongie, C. Carson, H. Greenspan, and J. Malik, "Color- and Texture-Based Image Segmentation using EM and its Application to Content-Based Image Retrieval," in *Proc. Int.l Conf. on Comp. Vision*, 1998.

[4] S. Beucher and F. Meyer, "The Morphological Approach to Segmentation: The Watershed Transformation," in *Mathematical Morphology in Image Processing*, E. R. Dougherty, Ed. New York: Marcel Dekker, 1993, pp. 433–481.

[5] C. Bishop, "Latent Variable Models," in *Learning in Graphical Models*, M. Jordan, Ed. MIT Press, 1998.

[6] E. Borenstein, E. Sharon, and S. Ullman, "Combining Top Down and Bottom-Up Segmentation," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2004.

[7] E. Borenstein and S. Ullman, "Class-Specific, Top-Down Segmentation," in *Proc. Eur. Conf. on Comp. Vision*, 2002.

[8] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *Int.l. J. of Comp. Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[9] T. Cootes, G. J. Edwards, and C. Taylor, "Active Appearance Models," *IEEE Trans. Pat. Anal. and Mach. Intel.*, vol. 23, no. 6, pp. 681–685, 2001.

[10] D. Cremers, "Dynamical statistical shape priors for level set based tracking," *IEEE Trans. Pat. Anal. and Mach. Intel.*, vol. 28, no. 8, pp. 1262–1273, August 2006.

[11] D. Cremers, N. Sochen, and C. Schnorr, "Towards Recognition Based Variational Segmentation Using Shape Priors and Dynamic Labelling," in *Proc. Intl. Conf. Scale Space*, 2003.

[12] P. Dayan, G. Hinton, R. Neal, and R. Zemel, "The Helmholtz Machine," *Neural Computation*, vol. 7, pp. 889–904, 1995.

[13] A. Dempster, N. Laird, and D. Rudin, "Maximum Likelihood from Incomplete Data via the EM algorithm," *Journal of The Royal Statistical Society, Series B*, 1977.

[14] R. Fergus, P. Perona, and A. Zisserman, "Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition," *Int.l. J. of Comp. Vision*, vol. 71, no. 3, pp. 273–303, 2007.

[15] V. Ferrari, T. Tuytelaars, and L. V. Gool, "Simultaneous Object Recognition and Segmentation by Image Exploration," in *Proc. Eur. Conf. on Comp. Vision*, 2004.

[16] B. Frey and N. Jojic, "Estimating Mixture Models of Images and Inferring Spatial Transformations Using the EM Algorithm," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 1999.

[17] U. Grenander, *General Pattern Theory*. Oxford University Press, 1993.

[18] T. Jaakkola, "Tutorial on Variational Approximation Methods," in *Advanced Mean Field Methods: Theory and Practice*. MIT Press, 2000.

[19] R. Jacobs, "Methods for Combining Experts' Probability Assesements," *Neural Computation*, no. 7, pp. 867–888, 1995.

[20] M. Jones and T. Poggio, "Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes," *Int.l. J. of Comp. Vision*, vol. 29, no. 2, pp. 107–131, 1998.

[21] I. Kokkinos and P. Maragos, "An Expectation Maximization Approach to the Synergy between Object Categorization and Image Segmentation," in *Proc. Int.l Conf. on Comp. Vision*, 2005.

[22] I. Kokkinos, P. Maragos, and A. Yuille, "Bottom-Up and Top-Down Object Detection Using Primal Sketch Features and Graphical Models," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2006.

[23] I. Kokkinos and A. Yuille, "Unsupervised Learning of Object Deformation Models," in *Proc. Int.l Conf. on Comp. Vision*, 2007.

[24] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Obj cut," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2005.

[25] B. Leibe, A. Leonardis, and B. Schiele, "Combined Object Categorization and Segmentation with an Implicit Shape Model," in *Proc. Eur. Conf. on Comp. Vision*, 2004, SLCV workshop.

[26] M. Leventon, O. Faugeras, and E. Grimson, "Statistical Shape Influence in Geodesic Active Contours," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2000.

[27] A. Levin and Y. Weiss, "Learning to Combine Bottom-Up and Top-Down Segmentation," in *Proc. Eur. Conf. on Comp. Vision*, 2006.

[28] D. Marr, *Vision*. W.H. Freeman, 1982.

[29] D. Martin, C. Fowlkes, and J. Malik, "Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues," *IEEE Trans. Pat. Anal. and Mach. Intel.*, vol. 26, no. 5, pp. 530–549, 2004.

[30] I. Matthews and S. Baker, "Active Appearance Models Revisited," *Int.l. J. of Comp. Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[31] R. Milanese, H. Wechsler, S. Gil, J. M. Bost, and T. Pun, "Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 1994.

[32] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering Human Body Configurations: Combining Segmentation and Recoginiton," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2004.

[33] D. Mumford, "Neuronal Architectures for Pattern Theoretic Problems," in *Large Scale Theories of the Cortex*. MIT Press, 1994.

[34] ——, "Pattern Theory: A Unifying Approach," in *Perception as Bayesian Inference*, 1996.

[35] R. Neal and G. Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse and Other Variants," in *Learning in Graphical Models*, M. Jordan, Ed., 1998.

[36] A. Opelt, A. Fussenegger, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition." in *Proc. Eur. Conf. on Comp. Vision*, 2004.

[37] S. Osher and J. Sethian, "Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations," *Journal of Computational Physics*, vol. 79, pp. 12–49, 1988.

[38] G. Papandreou and P. Maragos, "Multigrid Geometric Active Contour Models," *IEEE Trans. Im. Proc.*, vol. 16, no. 1, pp. 229–240, 2007.

[39] N. Paragios and R. Deriche, "Geodesic Active Regions: A new Paradigm to Deal with Frame Partition Problems in Computer Vision," *Journal of Visual Communication and Image Representation*, vol. 13, pp. 249–268, 2002.

[40] R. Gross and I. Matthews and S. Baker, "Active Appearance Models With Occlusion," *Image and Vision Computing*, vol. 24, no. 6, pp. 593–604, 2006.

[41] R. Rao and D. Ballard, "Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex," *Vision Research*, vol. 9, pp. 721–763, 1997.

[42] M. Rousson and D. Cremers, "Efficient kernel density estimation of shape and intensity priors for level set segmentation," in *Medical Image Computing and Computer Assisted Intervention*, 2005.

[43] M. Rousson and N. Paragios, "Shape Priors for Level Set Representations," in *Proc. Eur. Conf. on Comp. Vision*, 2002.

[44] J. Sethian, *Level Set Methods*. Cambridge University Press, 1996.

[45] Z. W. Tu, X. Chen, A. Yuille, and S. C. Zhu, "Image Parsing: Unifying Segmentation, Detection, and Recognition," *Int.l. J. of Comp. Vision*, vol. 63, no. 2, pp. 113–140, 2005.

[46] Z. W. Tu and S. C. Zhu, "Image Segmentation by Data-Driven MCMC," *IEEE Trans. Pat. Anal. and Mach. Intel.*, vol. 24, no. 5, pp. 657–673, 2002.

[47] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[48] S. Ullman, "Sequence Seeking and Counterstreams," in *Large Scale Theories of the Cortex*. MIT Press, 1994.

[49] Y. Weiss and E. Adelson, "Perceptually Organized EM: a Framework for Motion Estimation that Combines Information about Form and Motion," in *Proc. Int.l Conf. on Comp. Vision*, 1995.

[50] M. Welling, M. Weber, and P. Perona, "Unsupervised Learning of Models for Recognition," in *Proc. Eur. Conf. on Comp. Vision*, 2000.

[51] J. Winn and N. Jojic, "LOCUS: Learning Object Classes with Unsupervised Segmentation," in *Proc. Int.l Conf. on Comp. Vision*, 2005.

[52] S. Xu and J. Shi, "Object Specific Figure-Ground Segregation," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2003.

[53] S. C. Zhu and A. Yuille, "Region Competition: Unifying Snakes, Region Growing and Bayes/MDL for Multiband Image Segmentation," *IEEE Trans. Pat. Anal. and Mach. Intel.*, vol. 18, no. 9, pp. 884–900, 1996.

[54] S. C. Zhu, R. Zhang, and Z. W. Tu, "Integrating Top-Down/Bottom-Up for Object Recognition by Data-Driven Markov Chain Monte Carlo," in *Proc. IEEE Conf. on Comp. Vision & Pat. Rec.*, 2000.

**Iasonas Kokkinos** (S '02, M '06) received the Diploma in Electrical and Computer Engineering and the Ph.D. degree from the National Technical University of Athens, Greece, in 2001 and 2006, respectively.

During 2004 he visited with the Odyssee team at Sophia-Antipolis, France, and in 2006 he joined the Center for Image and Vision Sciences at UCLA as a post-doctoral researcher.

His research interests are in computer vision, pattern recognition, and image and signal processing. He has worked on texture segmentation, biologically motivated vision and nonlinear dynamical modeling, while his current research focus is on the combination of top-down approaches with bottom-up information for the problems related to object detection. He has served as a reviewer for the IEEE Transactions on PAMI and Image Processing and on the program committee of EMMCVPR '07 and ICCV '07.

**Petros Maragos** (S'81-M'85-SM'91-F'95) received the Diploma in electrical engineering from the National Technical University of Athens in 1980, and the M.Sc.E.E. and Ph.D. degrees from Georgia Tech, Atlanta, USA, in 1982 and 1985.

In 1985 he joined the faculty of the Division of Applied Sciences at Harvard University, Massachusetts, where he worked for 8 years as professor of electrical engineering. In 1993 he joined the faculty of the ECE School at Georgia Tech. During parts of 1996-1998 he was on sabbatical and academic leave working as director of research at the Institute for Language and Speech Processing in Athens. Since 1998 he has been working as professor at the NTUA School of ECE. His research and teaching interests include signal processing, systems theory, communications, pattern recognition, and their applications to image processing and computer vision, speech processing and recognition, and multimedia. He has served as associate editor for IEEE Transactions (*ASSP* and *PAMI*), and editorial board member for the journals *Signal Processing* and *Visual Communications and Image Representation*; general chairman or co-chair of conferences (VCIP'92, ISMM'96, VLBV'01, MMSP'07); and member of IEEE DSP committees.

His research has received several awards, including: a 1987 NSF Presidential Young Investigator Award; the 1988 IEEE SP Society's Young Author Paper Award for the paper 'Morphological Filters'; the 1994 IEEE SP Senior Award and the 1995 IEEE Baker Award for the paper 'Energy Separation in Signal Modulations with Application to Speech Analysis'; the 1996 Pattern Recognition Society's Honorable Mention Award for the paper 'Min-Max Classifiers'; 1995 election to IEEE Fellow; the 2007 EURASIP Technical Achievements Award for contributions to nonlinear signal processing and systems theory, image processing, and speech processing.