An expectation maximization approach to the synergy between image segmentation and object categorization

Conference Paper *in* Proceedings / IEEE International Conference on Computer Vision. IEEE International Conference on Computer Vision · November 2005

DOI: 10.1109/ICCV.2005.35 · Source: IEEE Xplore

CITATIONS
15

READS
68

2 authors, including:

Petros Maragos
National Technical University of Athens
564 PUBLICATIONS **17,387** CITATIONS

SEE PROFILE

An Expectation Maximization Approach to the Synergy Between Image Segmentation and Object Categorization

Iasonas Kokkinos and Petros Maragos
Computer Vision, Signal Processing and Speech Communication Group
School of Electrical and Computer Engineering
National Technical University of Athens, Greece
jkokkin@cs.ntua.gr, maragos@cs.ntua.gr *

Abstract

In this work we deal with the problem of modelling and exploiting the interaction between the processes of image segmentation and object categorization. We propose a novel framework to address this problem that is based on the combination of the Expectation Maximization (EM) algorithm and generative models for object categories. Using a concise formulation of the interaction between these two processes, segmentation is interpreted as the E step, assigning observations to models, whereas object detection/analysis is modelled as the M-step, fitting models to observations. We present in detail the segmentation and detection processes comprising the E and M steps and demonstrate results on the joint detection and segmentation of the object categories of faces and cars.

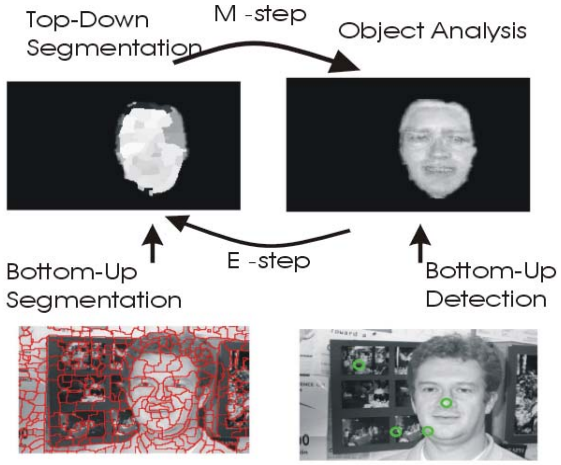

Figure 1. Overview of our approach

*This work was supported by the Greek research program HRAKLEITOS,which is co-funded by the European Social Fund (75%) and National Resources (25%)

1 Introduction

Two major computer vision problems, image segmentation and object recognition, have been traditionally dealt with using a strict, bottom-up ordering [18]: first segments are formed and subsequently recognition takes place. This typically results in suboptimal segmentations and numerous false detections; the cooperation of these two processes however can result in enhanced performance (*Synergy*).

In this paper we present a probabilistic approach to modelling and exploiting the interaction between these two problems that is based on the Expectation Maximization (EM) [10] algorithm, as shown in Fig. 1. EM is a well established algorithm for maximum likelihood parameter estimation, and as we show in this work facilitates the cooperation of bottom-up and top-down processes in an elegant and principled manner. We have systematically applied this approach to two categories of images, faces and cars, and obtained convincing results indicating its suitability for this problem.

After briefly presenting previous work, in section 2 we present our approach. In section 3 we describe the detection, segmentation and object analysis components of our system and in section 4 describe in detail the application of the EM algorithm. Experimental results are given in section 5.

1.1 Previous Work

Models that integrate high- and low-level processes of vision have been proposed during the previous decade by researchers in the biological vision community [27, 23, 25] but only recently have such models been made practically applicable to computer vision problems [30, 26, 6, 17, 9, 13, 22]. In order to clarify how our system relates to previous research on the same area, we briefly present the work in [30, 26, 6, 17] which lies closer to ours.

The approach of Tu, Chen, Yuille and Zhu [26] performs a stochastic search in the space of regions and hypotheses: hypotheses are generated, merged, split or discarded using bottom up proposals to guide the search while the regions corresponding to these hypotheses are evolved according to the region competition functional [31]. Our approach is motivated by similar ideas, since we use generative models to explain portions of the image in terms of objects, but we use the deterministic EM algorithm instead of the stochastic search procedure proposed in [26]. Even though the EM algorithm can get stuck in local minima, our experience has shown that it performs reasonably well and is significantly faster. Concerning the model's architecture, we use directly the outputs of a low-level segmentation algorithm, thereby introducing an intermediate layer that drastically reduces the computational burden. Related ideas have been proposed in [2] as being applicable to the model in [26].

An approach that uses the EM algorithm to perform an object-specific segmentation of an image is the 'sprites & layers' model of [14], where an E-step assigns observations to objects ('sprites') and updates the transformation of a prototypical object and the M-step updates the object parameters. In this work it is not finally decided whether an object exists in the image, while the background model is estimated from a fixed set of images, thereby introducing strong prior knowledge that may not be available in the general setting.

The object representation of Borenstein & Ullman [6, 5] and Leibe et al [17] uses codebooks of local appearance, which are brought together to build a segmentation map as shown in Fig. 2(b). It is known from the training stage which pixels in the patch belong to the object and which to the background, so this serves as a point of reference for all segmentations that are compliant with the existence of an object at a specific location. Codebook representations are typically high dimensional, while the segmentation depends on the ability to cover a large area of the object using overlapping patches, rather than fitting a model to an image.

In another approach using the architecture of Fig. 2(b), Xu and Shi [30] propose using an object-sensitive affinity measure, and 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 absence of a probabilistic interpretation impedes the cooperation with other processes while the system does not eventually help determine whether there is an object in the image.

2. EM Approach to Synergy

As in most top-down models of vision [24, 25, 26], in our approach scene analysis is formulated as the estimation of the parameters of a set of hypotheses that explain an observed image. According to this analysis-by-synthesis framework, models of objects are fit to the image, and to each object is assigned that part of the image that it best explains. This is a typical 'chicken-and-egg' problem for which we use the EM-algorithm. We note here that the use of the EM algorithm for image segmentation problems is certainly not novel; it has been used previously for low level problems like feature-based image segmentation or layered motion estimation [28]. In this work we introduce the EM algorithm as a natural and well-founded framework to model the high level problem of the interaction between object categorization and image segmentation.

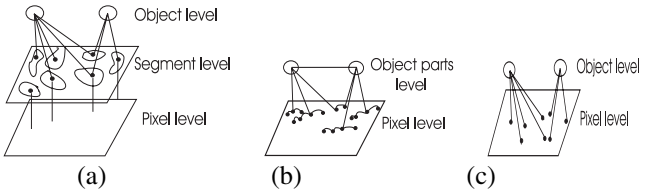

Figure 2. Architectures for synergy (a) Our model (b) Models of [30, 6, 17] (c) Models of [26, 14]. Edges denote interactions leading to grouping.

Before presenting our approach we introduce notation and the basic concepts with a brief review of the application of the EM algorithm to the problem of parameter estimation for a mixture distribution along the lines of [4].

2.1. EM algorithm for mixture modelling: Basic Concepts

Assume we are given a set of N independent observations $X = \{X_1, \ldots, X_N\}$ which can be modelled by a mixture of K distributions $P_k(X|\theta_k)$ with mixing weights $\pi_1 \ldots \pi_K$, considered known for simplicity. The log-likelihood of the observations is given by:

$$\log(P(X)) = \sum_n \log\left(\sum_k \pi_k P_k(X_n|\theta_k)\right) \qquad (1)$$

Maximizing this sum with respect to θ_k is intractable, since a summation appears inside the logarithms. We introduce for each observation n a vector of mutually exclusive hidden binary variables $Z_n = \{z_{1,n}, \ldots z_{K,n}\}$ s.t. $\sum_k z_{k,n} = 1$, with $z_{l,n} = 1$ if X_n is due to cause l. The logarithm of $P(X, Z)$ for a fixed Z is then given by:

$$\log(P(X, Z)) = \sum_n \sum_k z_{k,n} \log\left(\pi_k P_k(X_n|\theta_k)\right) \qquad (2)$$

For a set of parameters θ^* the posterior distribution of Z is:

$$R_{n,k} = P(z_{n,k} = 1|X_n, \theta_k^*) = \frac{\pi_k P_k(X_n|\theta_k^*)}{\sum_j \pi_j P_j(X_n|\theta_j^*)} \qquad (3)$$

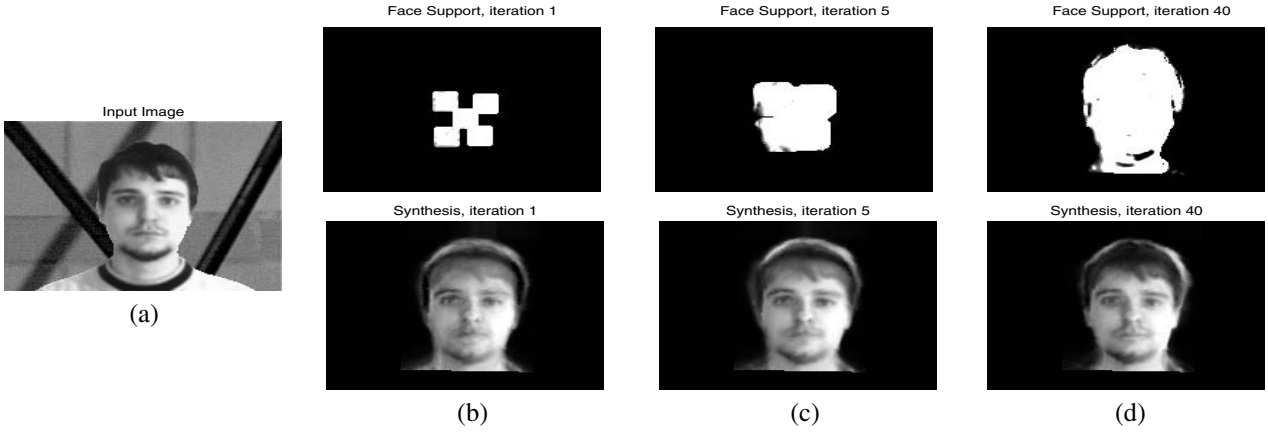

Figure 3. Input image (a) and EM iterations (b) - (d). Top row: evolution of the support of the face hypothesis on a pixel-wise basis (E-step). Bottom row: synthesis results, using the above supports (M-step).

Using Eqn. (3) the expectation of the expression in Eqn. (2) becomes:

$$E_Z(\log(P(X,Z|\theta))|\theta_k^*) = \sum_{n,k} R_{n,k} \log\left(\pi_k P_k(X_n|\theta_k)\right) \quad (4)$$

This last expression can be directly optimized with respect to θ_k since the summation is outside the logarithm. The EM algorithm amounts to repeatedly estimating the values of $R_{n,k}$ using θ^* (E-step) and then maximizing Eqn. (4) w.r.t. θ (M-step); this process consistently increases the log-likelihood [10] and converges to a local maximum of Eqn. (1).

2.2. Application to Synergy

We can apply the EM framework to the problem we address by treating segmentation as the E-step, where the parts of the image which belong to each object are assigned to it and object analysis as the M-step, where the model parameters are fit to the data that it has occupied.

In order to clarify the main idea, we present how it works using only one hypothesis H_0 for the background and another, H_1 for the object. As in the EM algorithm, we introduce two non-overlapping fields of hidden binary random variables, Z_0, Z_1, corresponding to these two hypotheses. For a set of parameters of the two models, we have a posterior distribution R on Z; we will be referring to the support of hypothesis k as the set $S_k = \{n : R_{n,k} = \max_j R_{n,j}\}$.

For equal mixing weights the expectation of the log-likelihood of the image I can then be written as:

$$E_Z(\log P(I,Z|\theta)|\theta^*) = \sum_n \sum_{k=1,2} R_{n,k} \log\left(\frac{1}{2} P_k(I_n|\theta_k)\right)$$

where n indexes the image pixels. This leads to the following EM scheme:

$$E: \quad R_{n,k} = \frac{P_k(I_n|\theta_k^*)}{\sum_{j=1,2} P_j(I_n|\theta_j^*)}$$

$$M: \quad \theta_k^* = \text{argmax}_\theta \sum_n R_{n,k} P_k(I_n|\theta_k)$$

We can see the result of alternatively applying these two steps in Fig. 3: a location in the image proposed by a front-end detection system is used to initialize the object's support. Subsequently an initial synthesis of the object is estimated using this small area, which leads subsequently to the assignment of a larger area to the object; in the E-step information about neighboring labels is used to avoid the emergence of wiggly boundaries. The synthesis and segmentation gradually improve, converging to a solution that adequately models a region of the image in terms of an object.

2.3. EM for segments

The discussion up to now has dealt with the problem pixel-wise, treating each pixel as a separate observation, while the whole image is explained in terms of only two models, object and background. This does not apply to most practical situations where background regions are of highly variable structure, so simple models rarely suffice for the whole background. Another problem is that when performing this hypothesis-competition process on a pixel-wise basis many resources can be spent until an obviously homogeneous image patch is passed from one hypothesis to the other, while it is easier to get stuck in local minima.

In our approach we perform the competition process over segments instead of pixels, as shown in Fig.2(a). This results on the one hand in robustness, since better background models can be built over small image regions and on the

other hand in increased efficiency since the computational burden is lightened and local minima are avoided. The previous discussion about the EM algorithm can be repeated as is, with the modification that the observations are not single pixels but the whole content of a segment. This means that the $R_{n,k}$ that were previously estimated on a per-pixel basis are estimated at the level of segments, and are inherited to the underlying pixels. Details about the how these quantities can be estimated will be given in section 4.

3. Synergy System Components

Our system accepts as its bottom-up input an oversegmentation of the image and a set of locations proposed by an object detection system. The components we have used for bottom up segmentation and detection combine simplicity and efficiency; we do not elaborate on these since other models could be used as well, as long as we get an oversegmentation of the image and few misses from the detector. Details will be given in a larger version of the paper.

3.1 Low Level Segmentation

The goal of this stage is to partition the image into regions of pixels that should come together, whatever higher-level object they may belong to. Accomplishing this step is not trivial, but we can get acceptable results for our purpose if we oversegment the input image.

Given that low-level segmentation on its own is not the objective of this paper, we have experimented with efficient segmentation algorithms and decided to use the morphological watershed algorithm [3]. The boundaries of the image are found using the Brightness-Gradient edge detector of [19] and the regional minima of edge strength are used as markers. Even though more elaborated image segmentation schemes could be used they would add unnecessary complexity to the overall system. We note that we have experimented with different segmentation algorithms but found no significant difference in overall performance.

3.2. Object Detection System

We experimented initially with the parts-based detection system of [11] which relies on a Markov Random Field (MRF) formulation for object detection and estimates the object locations by efficiently performing message passing operations. After filtering the image with a multiscale derivative-of-Gaussians filterbank, a set of probabilistic detectors is used to estimate the probability $\Phi(l_n) = P(\mathbf{F}|l_n)$ of the filterbank outputs $\mathbf{F}$ being due to the existence of object part n at location l_n. Likely sets of locations of object parts $L = \{l_1, \ldots, l_n\}$ are found at maxima of the quantity

$$P(\mathbf{F}|I) = \prod_n \Phi_n(l_n) \prod_{n,m \in V} \Psi_{n,m}(l_n, l_m) \qquad (5)$$

where V denotes the set of vertices of the MRF, and the clique potentials $\Psi_{n,m}$ encode information about the relative positions of the object keypoints.

A drawback of this model is that it requires training with hand-labelled keypoint data which are difficult to determine for arbitrary object categories. We therefore experimented with a simplification of the bottom-up part of the model in [17], which uses an interest point operator to automatically determine salient object points during training. A detailed presentation of this method can be found in [17], so below we only sketch how it works. During the training phase an interest operator is used to pick up salient object parts and a codebook of joint local appearance and figure-ground labels is built along the lines of [1] by clustering the observations around the detected keypoints. When presented with a new image the regions picked up by the interest operator are matched to the codebook entries and vote for potential object locations. For each hypothesized object a segmentation map is built using the figure/ground masks of the codebook entries that led to its detection.

A simplification we introduced was to cluster the keypoints based on both edge and intensity information using a k-means procedure similar to [29], thereby obtaining a compact codebook. The codebook used in [17] has more than 10^3 entries while ours uses less than 50. Only a small proportion of the object is covered using such a codebook, but this is compensated by the global generative models we use subsequently.

3.3 Morphable Models

A core idea of our approach is that generative models can quantify how well a part of the observed image can be explained by an object hypothesis; this leads to the assignment of that part of the image to the object's support. The generative models for objects that we use are based on Active Appearance/Morphable Models [16, 7, 20] which have been successful in high level tasks like object recognition, pose estimation etc. Our focus is on their probabilistic interpretation which we review below.

Morphable models are based on separating the spatial deformations and appearance variations of the images belonging to an object category, thereby reducing the ghost effects of typical PCA analysis techniques [21]. Matching an observed image I to a morphable model is phrased as the minimization of

$$E(S,T) = \sum_{i \in P_T} (I(S(i)) - T_i)^2 \qquad (6)$$

where S (Shape) is a deformation that brings I into registration with the prototypical object, T (Texture) is the shape-free prediction of the image appearance according to the model and P_T is the set of template pixels. S and T are expressed in terms of the expansion coefficients $\mathbf{S} = (s_1, \ldots, s_{N_S})$, $\mathbf{T} = (t_1, \ldots, t_{N_T})$ on a low dimensional set of eigenvectors, $\mathcal{S}$ and $\mathcal{T}$ respectively:

$$S = \mathcal{S}_0 + \sum_{i=1}^{N_S} s_i \mathcal{S}_i, \quad T = \mathcal{T}_0 + \sum_{i=1}^{N_T} t_i \mathcal{T}_i, \qquad (7)$$

$\mathcal{S}_0$ and $\mathcal{T}_0$ being the mean shape and texture vectors. These eigenvectors are found during a training phase using either labelled data [7] or bootstrapping [16].

During the matching phase, the deformation S and synthesis T are iteratively updated to minimize the above criterion by performing gradient descent on $\mathbf{S}, \mathbf{T}$. This matching process can be seen [8] as a maximum-likelihood estimation of the model parameters, by writing $P(I) \propto \exp(-E(S,T))$, so that the matching process converges to a mode of $P(I)$. By adding a penalty term on the expansion coefficients the matching process provides a maximum a posteriori estimate of the model parameters, since the penalty acts like a prior term.

The likelihood of the data inside the object's support can be expressed as:

$$P(I|O) = \int_{\mathbf{S},\mathbf{T}} P(I|\mathbf{S},\mathbf{T}) P(\mathbf{S},\mathbf{T}|O) d\mathbf{S} d\mathbf{T} \qquad (8)$$

Using a common series of simplifications, we assume initially that the integral is dominated by a small area around the maximum location $\mathbf{S}^*, \mathbf{T}^*$ of the integrated function. We also assume that the prior on the model parameters is separable into two independent distributions over $\mathbf{S}$ and $\mathbf{T}$, and the noise process is considered independently distributed. The above expression is then proportional to

$$P(I|O) \propto \prod_{i \in P_T} P(I_i|\mathbf{S}^*, \mathbf{T}^*) P(\mathbf{S}^*|O) P(\mathbf{T}^*|O) \qquad (9)$$

Assuming the modelling error is an independent Gaussian process and the distribution of the model parameters is a Gaussian with diagonal covariance matrix, the logarithm $L = \log P(I|O)$ of the above expression becomes:

$$L = \sum_{i \in P_T} \frac{(I(S(i)) - T_i)^2}{2\sigma_i^2} + \sum_{i=1}^{N_S} \frac{s_i^2}{2{\lambda_{S_i}}^2} + \sum_{i=1}^{N_T} \frac{t_i^2}{2{\lambda_{T_i}}^2} + c \qquad (10)$$

where $\lambda_{S_i}, \lambda_{T_i}$ are the variances of the model parameters and c is a constant.

A noteworthy point is that in Eqn. (10) the reconstruction error variance σ_i is usually considered to equal a constant; this is a strong assumption and does not account for the varying difficulty in modelling different object areas. For example, cars have both large uniform areas like doors which can be modelled well using as few as 2 or 3 eigenvectors and more difficult areas like wheels where a larger modelling error is expected. We estimated the variance of the reconstruction error by backward wrapping the modelling error for each of the training images onto the model template end estimating for each template point the error's mean square value across the training set. As can be seen in Fig. 4 the variance of the error is far from homogeneous across the template support. We observed experimentally that using this spatially varying map gives better results during the E-step.

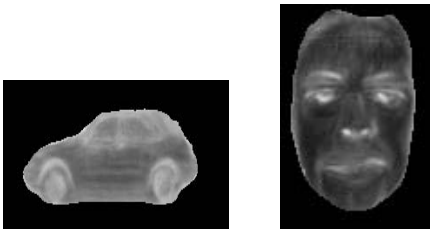

Figure 4. The variance of the reconstruction error for a morphable model for cars and faces is larger at areas of high complexity, like wheels and eyes, respectively.

4 E and M steps

Our system works by iteratively labelling image patches as belonging to either objects or background, according to how well each hypothesis explains their content. The background hypothesis assumes that the intensity distribution can be modelled using a Gaussian distribution; this arguably simple distribution can usually model adequately the observations within each segment. We describe below how the EM-based approach to synergy applies to our system.

4.1 E-step: Object-Based Segmentation

In the E-step the object and background hypotheses compete for the occupancy of image regions. The content I_R of region R is modelled by generative models, which means that we have a term $P(I_R|H_k)$ for each hypothesis H_k; these terms can be used to assign segments to hypotheses in a soft manner, based on Bayes' formula. It is natural however that a global object model cannot describe the observations within each segment as accurately as the locally determined background model. It is therefore necessary to modify the straightforward application of the E-step so as to make up for this imbalance. We therefore introduce a penalty term on the increased number of parameters used

by the background models in order to explain the same image area with a global object model. The modified log-likelihood of I_R under hypothesis H_k can be written

$$L'(I_R|H_k) = \log P_k(I_R|F_k^*) + \log P_k(F_k^*) - C_k \quad (11)$$

In the above expression F^* are the generative model parameters, which are the mean and deviation for the background model and the shape and texture expansion coefficients for the morphable model. $P_k(F_k^*)$ is a prior on the model parameters and C_k is the coding cost for the model parameters. The assignment of region I_R to hypothesis k can then be expressed according to the formula

$$P(H_k|I_R) = \frac{\exp(L'(I_R|H_k))}{\sum_j \exp(L'(I_R|H_j))} \quad (12)$$

Despite its simplicity this approach faces practical problems; for example the pixel observations are typically correlated and numerous, so the white Gaussian noise model results in almost binary decisions leading easily to local minima. Weighting terms should therefore be introduced in Eqn. (11) to avoid overly basing our decision on pixel intensity information.

After the introduction of weighting terms in the criterion above, the posterior probability of I_R belonging to hypothesis 1 can be written as the output of a sigmoidal function:

$$P_1 = \frac{1}{1+\exp(\sum_{j=1,2} \alpha_j \log P_j(I_R|F_j^*) + \beta_j \log P_j(F_j^*) + \gamma_j C_j)}$$

The weights $\alpha_j, \beta_j, \gamma_j$ of this expression have been estimated during the training phase using gradient descent, maximizing the likelihood of the figure/ground labels for the segments in the training set.

4.2 M-step - Parameter Estimation

In the M-step the parameters of the morphable models are updated in order to model the areas of the image assigned to them during the E-step. This can be phrased straightforwardly based on the probabilistic analysis of morphable models of section 3.3 by modifying the data likelihood term in Eqn. (10) according to Eqn. (4):

$$L = \sum_{i \in P_T} R(S(i)) \frac{(I(S(i)) - T_i)^2}{2\sigma_i^2} + \sum_{i=1}^{N_S} \frac{s_i^2}{2{\lambda_{S_i}}^2} + \sum_{i=1}^{N_T} \frac{t_i^2}{2{\lambda_{T_i}}^2} \quad (13)$$

In the above expression $S(i)$ is the image pixel registered to the template pixel i by S, and $R(S(i))$ equals the expected value of pixel $S(i)$ being assigned to the object hypothesis. This is what has been estimated during the E-step.

This expression can be minimized using a simple modification of the original matching equations [16], that weighs by $R(S(i))$ the error terms involved in the parameter update rules. Efficient algorithms like those in [20] can also be applied with minor modifications.

An interesting point is how to incorporate other sorts of information, like knowledge about the object location derived during the detection process; this can constrain the top-down matching process by exploiting the estimate provided by the bottom-up detection process [8]. These constraints can be incorporated in the parameter update rules for both the original and the EM-based matching functional.

For example, the figure/ground label $P(F|H_k)$ provided by the codebook-based detection system of [17] can be considered as an observation that has to be explained by the object's shape. This can be expressed by modifying the original likelihood function $P(I|\mathbf{S}, \mathbf{T})$ as

$$P(I, F|\mathbf{S}, \mathbf{T}) = P(I|\mathbf{S}, \mathbf{T}) P(F|\mathbf{S}) \quad (14)$$

where the image intensity and figure/ground labelling are considered independent. The last term forces the deformations to satisfy the constraints imposed by the detection process, namely the part of the image labelled figure by the detection system should fall within the object's support. This term can be expressed as:

$$P(F|\mathbf{S}) = \prod_P P(F_P|\mathbf{S}) \quad (15)$$

where F_P is the figure-ground labelling at pixel P. The product is over all pixels for which some observation is available concerning their figure-ground labelling and $P(O_P|\mathbf{S})$ is a generative model of the object's support obtained by deforming the template's support using $\mathbf{S}$. For convenience we have used a Gaussian distribution, though different models could probably be more appropriate.

When using the MRF-based detection system, the work in [8] can be directly applied to force the detected keypoint locations to be registered with the template keypoint locations.

5 Experimental Results

We have evaluated our approach on two object categories, namely faces and cars. For faces the MRF-based detection system and the morphable model were trained using 100 hand-labelled face images from the database in [15], and the system was evaluated on the data set of [29]. For cars we used 40 manually segmented car images from the database of [17] to learn the codebook-based detection model; since no hand-labelled keypoints were available, we used the bootstrapping algorithm in [16] with 35 of these images. The system was evaluated on the data set of [1], according to the criteria described therein. The prior distributions on the model parameters as well as the reconstruction error variances shown in Fig. 4 were estimated using

Figure 6. Top-down final segmentations of car and face images. By thresholding the E-step results at a fixed value, the parts of the image belonging to the objects hypothesis are assigned to it. Please see in color on screen.

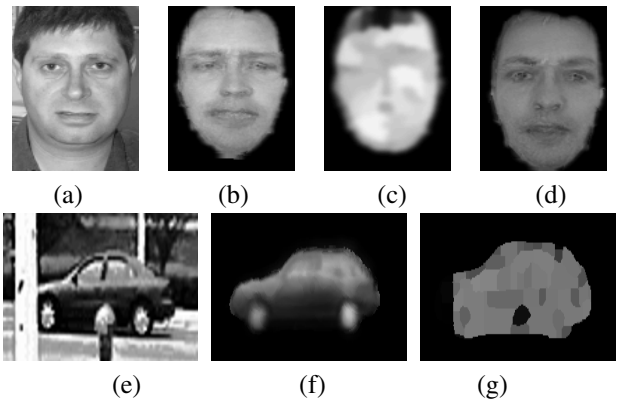

Figure 5. Use of segmentation information for of object analysis: Top row: (a) Input Image (b) face synthesis, using EM (c) E-step results, indicating the hair as occluding part of the face (d) face synthesis with the same algorithm, but using no E-step-based weighting term. Bottom row: (e)-(g): same as (a)-(c) above: the regions occluding the object are correctly classified as belonging to the background.

200 images from the test sets, that were not included in the evaluation set.

In Fig. 5 we see the importance of using segmentation information during the model fitting process: for the face on the top row our model *automatically* determines that the hair on the forehead occludes the face region and accurately synthesizes the object's face. Ignoring the segmentation information the synthesis results deteriorate, since the object's forehead would have to be heavily deformed in order to occupy a thin bright strip and the matching algorithm therefore converges to an alternative suboptimal solution. Along the same lines, in the bottom row we can see that our system successfully separates the part of the image belonging to the object from the occluding regions. Its appearance is synthesized based only on the parts of the image that are assigned to it.

In Fig. 6 we view some top-down segmentation results for faces and cars which 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 detection result. We observed that applying the EM algorithm to object hypotheses corresponding to false positives typically results in lower E-step values, indicating that the generative model cannot explain well the corresponding image area.

In order to systematically validate the claim that the joint treatment of the two tasks actually helps detection and not only segmentation, we have performed a face detection task on the database of [29] and a car detection task on the database provided by [1]. We used our system's bottom-up detector outputs and applied the EM-algorithm on the strongest proposed hypotheses. The score for each object was set proportional to the evidence received from the image in favor of its hypothesis, estimated in terms of the sum over the object's support of the corresponding E-step results. The scores of the bottom-up detection process and the top-down support were normalized to lie in $[0, 1]$ and their average was taken to fuse their results. We show in Fig. 7 the Receiver Operating Characteristic/Precision-Recall curves of these three detection systems, along with the typical performance measures used for these tasks in related publications [1, 17, 29, 12]. Specifically, for the fused system the F-measure in the car detection task equals .91 while the Equal Error Rate for face detection is equal to 89%; the individual systems have lower performance in both cases. Even though we have not focused on the object detection part of our work, these results are only slightly inferior to the current state-of-the-art (e.g. [17],[12]). We consider it however more important that the use of top-down information consistently improved the performance of a baseline bottom-up detection system, demonstrating that these two streams of information can complement each other. Similar conclusions have been drawn from a car detection task in [17].

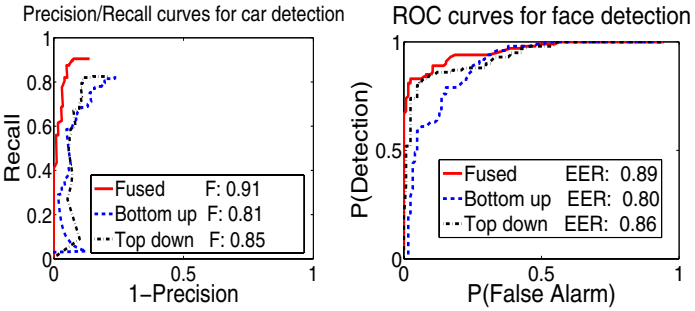

Figure 7. Precision-Recall Curves for car detection, ROC curves for face detection and associated F-measure and Equal Error Rate values. Please see text for details.

6 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. This offers a sound mathematical basis for a recently opened problem and helps clarify some of its aspects. Based on the EM algorithm we have built a system that has the potential to segment in a top-down manner images of objects belonging to highly variable categories. Efficiency and simplicity are two major advantages of our approach, which we have shown to be able to cooperate in a modular manner with the bottom-up processes of detection and segmentation.

In future work we wish to extend our approach to articulated objects by combining graphical models for object representation with the EM algorithm. It is also interesting to examine how low-level affinity information can be combined with top-down information in the E-step, as has been accomplished in the non-probabilistic framework of [30].

References

[1] Agrawal and Roth. Learning a Sparse Representation for Object Detection. In *ECCV*, 2002.

[2] A. Barbu and S. Zhu. Graph partition by Swendsen-Wang cuts. In *ICCV*, 2003.

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

[4] C. Bishop. Latent variable models. In M. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998.

[5] E. Borenstein, E. Sharon, and S. Ullman. Combining Top Down and Bottom-Up Segmentation. In *CVPR*, 2004.

[6] E. Borenstein and S. Ullman. Class-Specific, Top-Down Segmentation. In *ECCV*, 2002.

[7] T. Cootes, G. J. Edwards, and C. Taylor. Active Appearance Models. *IEEE Trans. PAMI*, 23(6):681–685, 2001.

[8] T. Cootes and C. Taylor. Constrained Active Appearance Models. In *ICCV*, 2001.

[9] D. Cremers, N. Sochen, and C. Schnorr. Multiphase Dynamic Labelling for Variational Recognition-Driven Image Segmentation. In *ECCV*, 2004.

[10] 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.

[11] P. Felzenszwalb and D. Huttenlocher. Efficient Matching of Pictorial Structures. In *CVPR*, 2000.

[12] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *CVPR*, 2003.

[13] V. Ferrari, T. Tuytelaars, and L. Gool. Simultaneous Object Recognition and Segmentation by Image Exploration. In *ECCV*, 2004.

[14] B. Frey and N. Jojic. Estimating Mixture Models of Images and Inferring Spatial Transformations Using the EM Algorithm. In *CVPR*, 1999.

[15] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust Face Detection Using the Hausdorff Distance. In *AVBPA*, 2001.

[16] M. Jones and T. Poggio. Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes. *Int.l. J. of Comp. Vision*, 29(2):107–131, 1998.

[17] B. Leibe, A. Leonardis, and B. Schiele. Combined Object Categorization and Segmentation with an Implicit Shape Model. In *ECCV*, 2004. SLCV workshop.

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

[19] D. Martin, C. Fowlkes, and J. Malik. Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues. *IEEE Trans. PAMI*, 26(5):530–549, 2004.

[20] I. Matthews and S. Baker. Active Appearance Models Revisited. *Int.l. J. of Comp. Vision*, 60(2):135–164, 2004.

[21] B. Moghaddam and A. Pentland. Probabilistic Visual Learning for Object Representation. *IEEE Trans. PAMI*, 19(7):693–710, 1997.

[22] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering Human Body Configurations: Combining Segmentation and Recogniton. In *CVPR*, 2004.

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

[24] D. Mumford. Pattern Theory: A Unifying Approach. In *Perception as Bayesian Inference*. 1996.

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

[26] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image Parsing: Unifying Segmentation, Detection, and Recognition. In *ICCV*, 2003.

[27] S. Ullman. Sequence Seeking and Counterstreams. In *Large Scale Theories of the Cortex*. 1994.

[28] Y. Weiss and E. Adelson. Perceptually organized EM: a framework for motion estimation that combines information about form an motion. In *ICCV*, 1995.

[29] M. Welling, M. Weber, and P. Perona. Unsupervised Learning of Models for Recognition. In *ECCV*, 2000.

[30] S. Xu and J. Shi. Object Specific Figure-Ground Segregation. In *CVPR*, 2003.

[31] S. Zhu and A. Yuille. Region Competition: Unifying Snakes, Region Growing and Bayes/MDL for Multiband Image Segmentation. *IEEE Trans. PAMI*, 18(9):884–900, 1996.