Shape Priors and Online Appearance Learning for Variational Segmentation and Object Recognition in Static Scenes

Martin Bergtholdt and Christoph Schnörr

Computer Vision, Graphics, and Pattern Recognition Group
Department of Mathematics and Computer Science
University of Mannheim, 68131 Mannheim, Germany
{bergtholdt,schnoerr}@uni-mannheim.de

Abstract. We present an integrated two-level approach to computationally analyzing image sequences of static scenes by variational segmentation. At the top level, estimated models of object appearance and background are probabilistically fused to obtain an a-posteriori probability for the occupancy of each pixel. The data-association strategy handles object occlusions explicitly.
At the lower level, object models are inferred by variational segmentation based on image data and statistical shape priors. The use of shape priors allows to distinguish between recognition of known objects and segmentation of unknown objects. The object models are sufficiently flexible to enable the integration of general cues like advanced shape distances. At the same time, they are highly constrained from the optimization viewpoint: the globally optimal parameters can be computed at each time instant by dynamic programming.
The novelty of our approach is the integration of state-of-the-art variational segmentation into a probabilistic framework for static scene analysis that combines both on-line learning and prior knowledge of various aspects of object appearance.

1 Introduction

Since the seminal work of Mumford and Shah on variational image segmentation [13], research has focused on generalizations of the Mumford-Shah functional along several directions.

A first direction concerns algorithmic schemes and contour representation by level sets for efficiently computing a good local minimum [3, 18, 19]. A second line of research investigates probabilistic models of image classes for variational segmentation that are richer than the piecewise-smooth image model underlying the original Mumford-Shah functional [22, 14, 12, 11]. Thirdly, statistical shape priors have been considered recently to complement data-driven variational approaches with a model-driven component [8, 7, 16, 5, 4, 15, 21].

This work contributes to the latter two directions in a twofold novel way. Firstly, probabilistic representations for *spatially structured* intensity distributions – as opposed to homogeneous textures - like the appearance of clothes,

are learnt on-line. Secondly, statistical shape priors based on a *psychophysically relevant* shape distance and prototypical views are learnt off-line from examples through structure-preserving Euclidean embedding and clustering. Both shape distance and embedding, as well as the incorporation of this knowledge into the overall variational approach as a statistical prior, distinguish our work from related hierarchical shape representations introduced by Gavrila for pedestrian detection in traffic scenes through template matching ([9, 10]).

Organization. In section 2, we describe the overall probabilistic model for image intensity, conditioned on estimated models of object appearance. The components of this model and the parameters that constitute object appearance in terms of intensity and shape, are explained in section 3. Section 4 describes the computation of optimal contours and data association through variational segmentation. We conclude with a discussion of experimental results and pointing out further work.

2 Probabilistic Image Model

At every time instant t, image intensity $I(\mathbf{x}, t)$ depends at each location $\mathbf{x} \in \Omega$ on the presence of N objects $\mathcal{O} = \{O_0; O_1, \ldots, O_N\}$, where O_0 denotes the background. Each object $O_k, k = 1, \ldots, N$, is specified by parameters, $O_k = \{\Theta_k, \mathbf{c}_k(s)\} \in \mathcal{O}$ which have the following meaning:

- Θ_k parametrizes a distribution p_{Θ_k} which models object appearance in terms of intensity. These distributions form the components of the overall image intensity distribution in eqn. (1). They are described in section 3.2.
- $\mathbf{c}_k(s)$ denotes the boundary contour of image region Ω_k occupied by object O_k, $\mathbf{c}_k(s) := \partial\Omega_k \quad \big(\Omega_k$ image region of object $O_k\big)$

With each object region Ω_k, we associate its characteristic function: $\chi_k(\mathbf{x}) = 1$ if $\mathbf{x} \in \Omega_k$ and 0 otherwise.

Given the parameters of all objects $\mathcal{O}$, the probabilistic image model reads:

$$p\big(I(\mathbf{x}) \,|\, \mathcal{O}\big) = \sum_{k=0}^{N} \pi_k(\mathbf{x})\; p_{\Theta_k}\big(I(\mathbf{x}) \,|\, \mathbf{c}_k\big)\;, \quad \pi_k(\mathbf{x}) = \frac{\chi_k(\mathbf{x})}{\sum_{j=0}^{N} \chi_j(\mathbf{x})}\;, \quad \forall \mathbf{x} \in \Omega \quad (1)$$

Parameters π_k are deterministically obtained from the object regions, the factors $p_{\Theta_k}\big(I(\mathbf{x}) \,|\, \mathbf{c}_k\big)$ are detailed in section 3.2. This "mixture of objects" model is less restrictive than partitioning models that divide an image into mutually exclusive regions, since in our case several objects are allowed to occupy the same image location (occlusions).

Basically, eqn. (1) models object appearance in terms of both *intensity and shape*. For each object O_k (including background), a parameterized intensity model is learnt and updated from frame to frame. This intensity information is combined with statistical prior information about possible object shapes, which has been learnt off-line. Through optimizing the contours $\mathbf{c}_k$ (see section 4), object regions Ω_k compete with each other in order to provide the "best explanation" of given image data.

3 Object Appearance

3.1 Statistical Shape Priors

We assume that, for each object class, a database is given containing shapes of different object views. These data are represented by a small subset of representative, prototypical views, obtained by pairwise dissimilarity clustering.

In order to cope with non-rigid objects like human shapes, we adopted from [1] the shape distance:

$$d_E(\mathbf{c}_1, \mathbf{c}_2) = \min_g E(g;\, \mathbf{c}_1, \mathbf{c}_2)\ , \qquad (2)$$

which is computed by minimizing the matching functional:

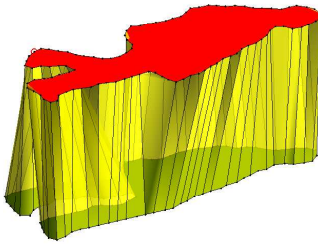

Fig. 1. Matching by minimizing (3) leads to an accurate correspondence of *parts* of non-rigid objects.

$$E(g;\, \mathbf{c}_1, \mathbf{c}_2) = \oint_0^1 \left\{ \frac{[\kappa_2(s) - \kappa_1(g(s))g'(s)]^2}{|\kappa_2(s)| + |\kappa_1(g(s))g'(s)|} + \lambda \frac{|g'(s) - 1|^2}{|g'(s)| + 1} \right\} ds \qquad (3)$$

over all smooth reparametrizations $g : [0,1] \longrightarrow [0,1]$. Here, κ_1, κ_2 denote the curvature functions of the contours $\mathbf{c}_1, \mathbf{c}_2$. Functional (3) involves bending (change of curvature) and stretching $g'(s)$ of contours, which allows to group contours that are *perceptually close* to each other, despite transformed parts (cf. [1]). The minimization in (2) is carried out by dynamic programming over all piecewise-linear and strictly monotonously increasing functions g. Figure 1 illustrates the result for two human shapes.

To determine representative shapes by clustering, we compute an Euclidean embedding $\{\mathbf{p}_k\}_{k=1,2,\ldots}$ of the given shape examples such that $\|\mathbf{p}_i - \mathbf{p}_j\| \approx d_E(\mathbf{c}_i, \mathbf{c}_j)\ ,\ \forall i, j$ [6], followed by k-means clustering [2]. As a result of this procedure, we obtain for each object class a small set of representative shapes, henceforth called *templates* $\mathcal{T} = \{\mathbf{c}_1^t, \ldots, \mathbf{c}_T^t\}$, see figure 2. We assume equal prior probabilities for the templates.

3.2 Object Intensity

We model the background O_0 by Gaussian mixture distributions for each pixel with - in our implementation - three components [17]. This simple approach proved to be effective and runs in real-time on current PCs.

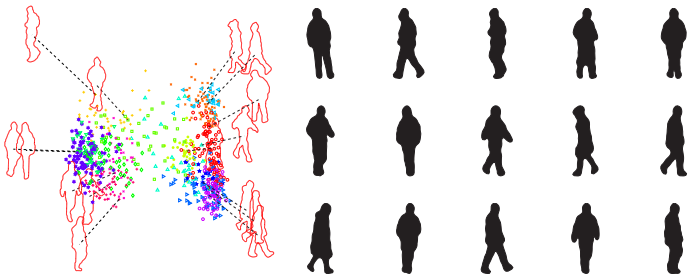

Fig. 2. Left, clustering of the views of human shapes, projected to the first two principal components. The clusters are indicated by prototypical shapes (cluster centers) dominating a range of corresponding views. Right, the templates corresponding to the cluster centers used in our experiments.

For each foreground object $O_1, \ldots, O_N$, we combine this approach with the statistical shape information described in the previous section as follows.

Each template $\mathbf{c}_j^t$ in $\mathcal{T}$ is represented in *normalized* coordinates $\mathbf{x}'$. For the current contour $\mathbf{c}_k$ of object O_k in the *image*, we compute in parallel the optimal matchings to all templates $\mathbf{c}_j^t$ in $\mathcal{T}$ through (2), and corresponding registrations of the enclosed regions by thin-plate splines [20] using the correspondences on the boundary. This establishes a one-to-one correspondence between image and template coordinates $\mathbf{x}, \mathbf{x}'$, see figure 3.

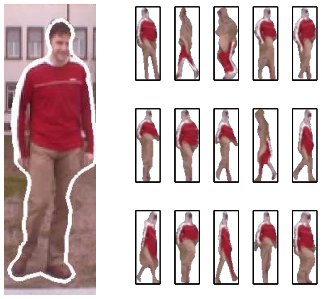

Fig. 3. A human shape and images in normalized coordinates. Note the plausible distortions of intensity. Also note that some matchings for perceptually dissimilar templates are wrong, but these integrate out over time.

The distribution $p_{\Theta_k}\big(I(\mathbf{x}) \,|\, \mathbf{c}_k\big)$ in (1) modeling the intensity of object O_k is then given by marginalizing out the shape templates:

$$p_{\Theta_k}\big(I(\mathbf{x}) \,|\, \mathbf{c}_k\big) = \sum_{j=1}^{T} p\big(I(\mathbf{x}) \,|\, \mathbf{c}_j^t\big)\; p(\mathbf{c}_j^t \,|\, \mathbf{c}_k)\;,$$

where $p\big(I(\mathbf{x}) \,|\, \mathbf{c}_j^t\big)$ records – analogously to the background – for each template $\mathbf{c}_j^t$ a pixel-wise Gaussian mixture model in normalized coordinates $\mathbf{x}'$:

$$p(I(\mathbf{x}) \,|\, \mathbf{c}_j^t) = \sum_{i=1}^{3} \pi_i^j(\mathbf{x}') \mathcal{N}\big(I(\mathbf{x}); \mu_i^j(\mathbf{x}'), \Sigma_i^j(\mathbf{x}')\big)\;, \tag{4}$$

and where the probability that template $\mathbf{c}_j^t$ is representative for the current object contour $\mathbf{c}_k$ in the image, is given by:

$$p(\mathbf{c}_j^t \,|\, \mathbf{c}_k) := \frac{\exp\big(-d_E(\mathbf{c}_j^t, \mathbf{c}_k)\big)}{\sum_{l=1}^{T} \exp\big(-d_E(\mathbf{c}_l^t, \mathbf{c}_k)\big)}$$

Hence, given the current image contour $\mathbf{c}_k$, parameters Θ_k comprise the mixture parameters for all templates, $\Theta_k = \big\{\pi_i^j(\mathbf{x}'), \mu_i^j(\mathbf{x}'), \Sigma_i^j(\mathbf{x}')\big\}_{i=1,2,3;\, j=1,\ldots,T}$.

4 Variational Inference

Having estimated parameters Θ_k given the image contour $\mathbf{c}_k$ for object O_k, we wish to update $\mathbf{c}_k$. This is accomplished by computing in parallel for all templates $\mathbf{c}_j^t$ the contour $\mathbf{c}$ minimizing the functional:

$$J(\mathbf{c}; \mathbf{c}_j^t, \mathcal{O}) = J_d(\mathbf{c}; \mathbf{c}_j^t, \mathcal{O}) + \alpha\, J_p(\mathbf{c}; \mathbf{c}_j^t) \quad (5)$$

As usual in variational segmentation (cf. section 1), this functional comprises a data term and a prior. The prior is simply the matching functional (2):

$$J_p(\mathbf{c}; \mathbf{c}_j^t) = d_E(\mathbf{c}, \mathbf{c}_j^t)\ , \quad (6)$$

whereas the data term has the form:

$$J_d(\mathbf{c}; \mathbf{c}_j^t, \mathcal{O}) = -\oint \Big\{ \log p\big(I(\mathbf{c}) \,|\, \mathbf{c}_j^t\big) + \log p(I(\mathbf{c}) \,|\, \mathcal{O} \setminus O_k) + \log p(\nabla I(\mathbf{c})) \Big\} \quad (7)$$

The first term of the integrand maximizes the probability that the intensity observed at $\mathbf{c}(s)$ matches the model associated with template $\mathbf{c}_j^t$ – see (4). The second term in (7), on the other hand, maximizes the probability that $I(\mathbf{c}(s))$ matches the model of another object, or the background – see (1). As a result, both terms together invoke segmentation through "region competition". Finally, the third term in (7) attracts $\mathbf{c}$ to edges, as is common in geodesic snake approaches and accounts for our prior belief that boundary contours are more probable at image edges. It has the form $-\log p(\nabla I(\mathbf{c})) = \frac{1}{1+|\nabla I(\mathbf{c})|}$. In our implementation, functional (5) is *globally optimized* over a region centered around $\mathbf{c}_k$, as indicated in Figure 4.

Fig. 4. Optimal updates of image contours are given by a deformation along the normal direction. We minimize the segmentation functional (5) through dynamic programming over a set of discrete putative contour locations (top). Intensity information is approximated locally by samples along inner normal directions of the current boundary (bottom).

5 Experiments and Discussion

In order to evaluate our novel combination of on-line object appearance models and statistical shape priors within the framework of variational segmentation,

we processed an image sequence of 1400 frames containing four different persons entering and leaving the scene. This was deliberately done without any additional knowledge about object dynamics or elaborate scheme for tracking, with the exception of simple first-order Kalman predicting the bounding boxes of current object contours. Candidate regions were generated for regions which yielded local minima in the posterior probability of (1). New objects are automatically instantiated for these regions or existing objects re-instantiated if they have been absent. In subsequent frames we refine the contours of the objects using variational segmentation as described in section 4 followed by updates of the object and background intensity model as described in section 3.2 and [17].

We point out that not any tuning parameters are involved in our approach. Data-association is entirely accomplished by (1), and decisions are based on MAP-estimates. The algorithm automatically tracked the four humans in the sequence and due to the learned appearance information, recovers easily after occlusions or if an object leaves and reenters the scene. Several images of the sequence are shown in figure 5.

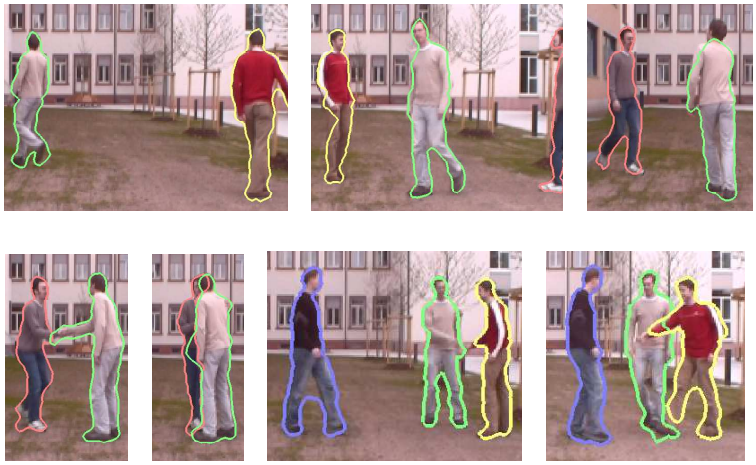

Fig. 5. Segmentation results. Shown are clipped frames 349, 519, 823, 842, 874, 1076, and 1091.

After processing the entire sequence we show in figure 6 images sampled from the learned intensity model for the four objects. The samples are generated by choosing a template and sampling from the Gaussian mixture model at each pixel location. Overall we can see that the intensity information was correctly learned for the objects. A closer look also reveals that the multivariate nature of the intensity information due to clothes and viewpoint is captured by the mixture model: e.g. the white stripes on the shoulders of the person in the lower right appear in the sampled template images. Moreover probable locations for the hands can also be identified as the flesh-colored areas inside the regions, see e.g. template 2 of the person in the upper left.

Fig. 6. Samples from the appearance model for four persons. Each triplet is: image from the sequence, sampled template 2 and 7.

6 Conclusion

We have presented a novel framework for scene analysis by the combination of offline and online object learning together with variational image segmentation. As we match shape and intensity information along the boundary we can efficiently solve for the global minimum of the correspondence problem using dynamic programming. This is a big advantage of 1D matchings that can not be transferred to the 2D case. We have found that for the case of human objects, the 2D problem is sufficiently well approximated by the 1D matching on the boundary and subsequent transformation to 2D using the correspondences.

In the future we want to augment our shape database to more objects. We will also investigate how shape information may be learned online.

References

1. R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. *Vision Res.*, 38:2365–2385, 1998.
2. M. Bergtholdt, D. Cremers, and C. Schnörr. Variational segmentation with shape priors. In N. Paragios, Y. Chen, and O. Faugeras, editors, *Mathematical Models in Computer Vision: The Handbook*. Springer, 2005.
3. T. Chan and L. Vese. Active contours without edges. *IEEE Trans. Image Processing*, 10(2):266–277, 2001.
4. T. Chan and W. Zhu. Level set based shape prior segmentation. Technical Report 03-66, Computational Applied Mathematics, UCLA, Los Angeles, 2003.
5. Y. Chen, H.D. Tagare, S. Thiruvenkadam, F. Huang, D. Wilson, K.S. Gopinath, R.W. Briggs, and E.A. Geiser. Using prior shapes in geometric active contours in a variational framework. *Intl. J. of Computer Vision*, 50(3):315–328, 2002.

6. T. F. Cox and M. A. A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 2001.
7. D. Cremers, N. Sochen, and C. Schnörr. Towards recognition-based variational segmentation using shape priors and dynamic labeling. In L. Griffith, editor, *Int. Conf. on Scale Space Theories in Computer Vision*, volume 2695 of *LNCS*, pages 388–400, Isle of Skye, 2003. Springer.
8. D. Cremers, F. Tischhäuser, J. Weickert, and C. Schnörr. Diffusion Snakes: Introducing statistical shape knowledge into the Mumford–Shah functional. *Intl. J. of Computer Vision*, 50(3):295–313, 2002.
9. D. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *Proc. of IEEE International Conference on Pattern Recognition*, pages 439–444. Brisbane, Australia, 1998.
10. D. Gavrila. Sensor-based pedestrian protection. In *IEEE Intelligent Systems*, volume 16, pages 77–81. 2001.
11. Matthias Heiler and Christoph Schnörr. Natural image statistics for natural image segmentation. *Intl. J. of Computer Vision*, 63(1):5–19, 2005.
12. S. Jehan-Besson, M. Barlaud, and G. Aubert. Dream^{2}s: Deformable regions driven by an eularian accurate minimization method for image and video segmentation. *Int. J. Computer Vision*, 53(1):45–70, 2003.
13. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.
14. N. Paragios and R. Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *Intl. J. of Computer Vision*, 46(3):223–247, 2002.
15. T. Riklin-Raviv, N. Kiryati, and N. Sochen. Unlevel sets: Geometry and prior-based segmentation. In T. Pajdla and V. Hlavac, editors, *European Conf. on Computer Vision*, volume 3024 of *LNCS*, pages 50–61, Prague, 2004. Springer.
16. M. Rousson and N. Paragios. Shape priors for level set representations. In A. Heyden et al., editors, *Proc. of the Europ. Conf. on Comp. Vis.*, volume 2351 of *LNCS*, pages 78–92, Copenhagen, May 2002. Springer, Berlin.
17. C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Computer Science Conference on Computer Vision and Pattern Recognition (CVPR-99)*, pages 246–252, Los Alamitos, June 1999. IEEE.
18. A. Tsai, A. J. Yezzi, and A. S. Willsky. Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Trans. on Image Processing*, 10(8):1169–1186, 2001.
19. L.A. Vese and T.F. Chan. A multiphase level set framework for image segmentation using the mumford and shah model. *Intl. J. of Computer Vision*, 50(3):271–293, 2002.
20. G. Wahba. *Spline models for observational data.* SIAM, Philadelphia, 1990.
21. J. Yang and J.S. Duncan. 3d image segmentation of deformable objects with joint shape-intensity prior models using level sets. *Medical Image Analysis*, 8(3):285–294, 2004.
22. S. C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE PAMI*, 18(9):884–900, 1996.