GENERATIVE MODEL AND CONSISTENT ESTIMATION ALGORITHMS FOR NON-RIGID DEFORMABLE MODELS

S. Allassonnière[1], *E. Kuhn*[1], *A. Trouvé*[2], *Y. Amit*[3]

LAGA- Université Paris 13[1]
99 Av. J-B Clément
93430 Villetaneuse France

CMLA - ENS de Cachan[2]
61, av. du Président Wilson
94325 Cachan, Cedex

University of Chicago[3]
5734 S. University Ave.
Chicago, IL, 60637, USA

ABSTRACT

The link between Bayesian and variational approaches is well known in the image analysis community in particular in the context of deformable models. However, true generative models and consistent estimation procedures are usually not available and the current trend is the computation of statistics mainly based on PCA analysis. We advocate in this paper a careful statistical modeling of deformable structures and we propose an effective and consistent estimation algorithm for the various parameters (geometric and photometric) appearing in the models.

1. INTRODUCTION

One primary difficulty in the context of deformable template models is the initial choice of the template and of various parameters in the energies underlying the registration process. This problem is of utmost importance in the context of medical imaging and computational anatomy where people try to provide statistical models for anatomical and functional variability, but also in many problems of object detection and scene interpretation. Building real generative models, that handle pose variability and yield effective likelihood ratio tests for various discriminative purposes, is a fundamental issue mainly unsolved in the context of non-rigid objects.

In [1], a first step toward a statistical approach for the estimation of templates is proposed. In this paper our goal is to propose a coherent statistical framework for dense deformable templates both in terms of the probability model, and in terms of the effective estimation procedure of the template *and* of the deformation covariance structure. The quality of the learned models is tested on the standard problem of digit classification through simple likelihood ratio tests.

2. THE OBSERVATION MODEL

Let $(y_i)_{1\leq i\leq n}$ be the observed gray level training data. Each y_i is defined on a grid of pixels $\Lambda \hookrightarrow \mathbb{R}^2$ where for each $s \in \Lambda$, x_s is the location of pixel s in a specified domain $D \subset \mathbb{R}^2$. The template is a function from $\mathbb{R}^2$ to $\mathbb{R}$. Working within the small deformation framework ([2]), we assume the existence of an unobserved deformation field $z : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$y(s) = I_0(x_s - z(x_s)) + \sigma\epsilon(s) = zI_0(s) + \sigma\epsilon(s)$$

where $\epsilon(s)$ are i.i.d $\mathcal{N}(0,1)$, independent of all other variables.

2.1. The template and deformation model

The template I_0 and the deformation z are assumed to belong to subspaces of reproducing kernel Hilbert spaces V_p (resp V_g) with kernel K_p (resp K_g). Let $(p_k)_{1\leq k\leq k_p}$ be a set of landmarks which covers the domain D, the template function I_0 is parametrized by coefficients $\alpha \in \mathbb{R}^{k_p}$ through: $I_\alpha = \mathbf{K_p}\alpha,$ where $(\mathbf{K_p}\alpha)(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\alpha(k)$. Taking $(g_k)_{1\leq k\leq k_g} \in D$ to be a different fixed set of landmarks, for $\beta = (\beta^{(1)}, \beta^{(2)}) \in \mathbb{R}^{k_g} \times \mathbb{R}^{k_g}$ define the deformation field as

$$z_\beta(x) = (\mathbf{K_g}\beta)(x) = \sum_{k=1}^{k_g} K_g(x, g_k)(\beta^{(1)}(k), \beta^{(2)}(k)).$$

Assuming that the underlying deformation field is Gaussian a Gaussian distribution is induced on β. We denote the covariance matrix of this distribution by Γ_g.

2.2. Parameters and likelihood

We present a general model which includes mixtures of deformable templates. The model parameters of each component are denoted by $(\alpha_\tau, \sigma_\tau, \Gamma_g^\tau)_{1\leq\tau\leq T}$, where T denotes the number of model components, and the weight of the different mixtures is given by $(\rho(\tau))_{1\leq\tau\leq T}$. We introduce the following notation:

$\eta = (\theta, \rho)$ with $\theta = (\theta^\tau)_{1\leq\tau\leq T}$ and $\rho = (\rho(\tau))_{1\leq\tau\leq T}$, where θ^τ is composed of a geometric part $\theta_g^\tau = \Gamma_g^\tau$ and a photometric part $\theta_p^\tau = (\alpha_\tau, \sigma_\tau^2)$. We assume that $\theta = (\theta_g^\tau, \theta_p^\tau)_{1\leq\tau\leq T}$ belongs to the parameter space Θ defined as the open set $\Theta = \{\, \theta = (\alpha_\tau, \sigma_\tau^2, \Gamma_g^\tau)_{1\leq\tau\leq T} |\, \forall\tau \in \{1, \ldots, T\}$

$\alpha_\tau \in \mathbb{R}^{k_p},\ \sigma_\tau^2 > 0,\ \Gamma_g^\tau \in \Sigma^+_{2k_g,*}(\mathbb{R})\ \}$. Here $\Sigma^+_{2k_g,*}(\mathbb{R})$ is the set of strictly positive symmetric matrices.

In this general model, the unobserved variables corresponding to an observation y_i are the pair $\xi_i = (\beta_i, \tau_i)$. The likelihood of the observed data can be expressed as an integral over the unobserved variables:

$$q(y|\theta,\rho) = \sum_{\tau=1}^{T} \int q(y|\beta_\tau,\theta_p,\rho) q(\beta_\tau|\theta_g,\rho)\rho(\tau) d\beta_\tau \ ,$$

where the density functions are given by a Bayesian model.

2.3. The Bayesian model

Even though the parameters are finite dimensional it is unreasonable to compute a maximum-likelihood estimator when the training sample is small. Our goal is to demonstrate that with the introduction of apriori distributions on the parameters, estimation with small samples is still possible even within the rather complex framework described here, yielding good results in some concrete examples. The prior laws used are the following. Let μ_p and Γ_p be a fixed mean and covariance matrix, then:

$$\begin{cases} \rho \sim \nu_\rho \\ \theta = (\theta_g^\tau, \theta_p^\tau)_{1\le\tau\le T} \sim \otimes_{\tau=1}^T (\nu_g \otimes \nu_p) \mid \rho \\ \tau_1^n \sim \otimes_{i=1}^n \rho \mid \eta = (\theta,\rho) \\ \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g^{\tau_i}) \mid \eta,\ \tau_1^n \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(z_{\beta_i} I_{\alpha_i}, \sigma^2_{\tau_i} Id_\Lambda) \mid \beta_1^n,\ \eta,\ \tau_1^n \end{cases}$$

with

$$\begin{aligned} \nu_g(d\Gamma_g) &\propto \left(\exp(-\langle \Gamma_g^{-1}, \Gamma_g^0 \rangle/2) \tfrac{1}{\sqrt{|\Gamma_g|}} \right)^{a_g} d\Gamma_g, \\ &\quad (a_g > 2k_g + 1), \\ \nu_p(d\sigma^2, d\alpha) &\propto \left(\exp\left(-\tfrac{\sigma_0^2}{2\sigma^2} \right) \tfrac{1}{\sqrt{\sigma^2}} \right)^{a_p} \cdot \\ &\quad \exp\left((\alpha-\mu_p)^t\ (\Gamma_p)^{-1} (\alpha-\mu_p) \right) d\sigma^2 d\alpha, \\ \nu_\rho(\rho) &\propto \left(\prod_{\tau=1}^T \rho(\tau) \right)^{a_\rho} \end{aligned}$$

All priors are assumed independent. A natural choice for the apriori covariance matrices Γ_p and Γ_g^0 is to consider the matrices induced by the metric of the spaces V_p and V_g. Define the square matrices

$$\begin{aligned} M_p(k,k') &= K_p(p_k, p_{k'})\ \forall 1 \le k,k' \le k_p \\ M_g(k,k') &= K_g(g_k, g_{k'})\ \forall 1 \le k,k' \le k_g, \end{aligned} \tag{1}$$

and then set $\Gamma_p = M_p^{-1}$ and $\Gamma_g^0 = M_g^{-1}$, which are typical prior matrices used in many matching algorithms.

3. ESTIMATION: THEORETICAL RESULTS

For the theoretical results we focus here on the particular case of a single component (i.e. $T = 1$). The parameter estimates are obtained by maximizing the posterior density on θ conditional on y_1^n: $\hat\theta_n = \mathrm{argmax}_\theta\, q(\theta|y_1^n)$. Denoting by P the distribution governing the observations (which may lie outside the prescribed family of models) below are several results regarding the MAP estimator in terms of the set $\Theta_* = \{\ \theta_* \in \Theta \mid E_P(\log q(y|\theta_*)) = \sup_{\theta\in\Theta} E_P(\log q(y|\theta))\}$.

Theorem 1 (Existence of the MAP estimator) *([3]) For any sample y_1^n, there exists $\hat\theta_n \in \Theta$ such that*

$$q(\hat\theta_n|y_1^n) = \sup_{\theta\in\Theta} q(\theta|y_1^n)\,.$$

Theorem 2 (Consistency) *([3]) Assume that Θ_* is non empty. Then, for any compact set $K \subset \Theta$,*

$$\lim_{n\to+\infty} P(\ \delta(\hat\theta_n, \Theta_*) \ge \epsilon\ \wedge \hat\theta_n \in K\ ) = 0\,,$$

(δ is any metric compatible with the usual topology on Θ).

Moreover, if we introduce a baseline image $I_b : \mathbb{R}^2 \to \mathbb{R}$, set the template as $I_\alpha = \mathbf{K_p}\alpha + I_b$, and denote for any $R > 0$:

$$\begin{cases} \Theta^R = \{\ \theta \in \Theta \mid |\alpha| \le R,\ \},\ v(R) = \sup_{\theta\in\Theta^R} E_P(\log q(y|\theta)) \\ \Theta_*^R = \{\ \theta \in \Theta^R \mid E_P(\log q(y|\theta)) = v(R)\ \} \end{cases} \tag{2}$$

then the following result holds for the corresponding MAP estimator $\hat\theta_n^R$.

Theorem 3 (Consistency on bounded prototypes) *([3]) Assume that $2k_g < |\Lambda|$, that $P(dy) = p(y)dy$ where the density p is bounded with exponentially decaying tails and that the observations y_1^n are i.i.d under P. Assume also that the baseline I_b satisfies $|I_b(x)| > a|x| + b$ for some positive constant a. Then $\Theta_*^R \neq \emptyset$ and for any $\epsilon > 0$*

$$\lim_{n\to\infty} P(\delta(\hat\theta_n^R, \Theta_*^R) \ge \epsilon) = 0\,,$$

where δ is any metric compatible with the topology on Θ^R.

The condition $2k_g < |\Lambda|$ is quite weak and easily fulfilled in our applications. The condition on the baseline image is somewhat less natural but is necessary to guarantee that the estimate remains in Θ^R. In practice $I_b \equiv 0$ works fine.

4. ESTIMATION WITH THE EM ALGORITHM

Since the deformation coefficients β_i^τ and the mixture coefficients $\rho(\tau)$ are unobserved the natural approach is to use iterative algorithms such as EM ([4]) to maximize the posterior given the observations $\hat\eta = \mathrm{argmax}_\eta\, q(\eta|y_1^n)$, This can be rewritten as:

$$\max_{\eta,\nu} \left[\int \log q(y,u|\eta)\nu(u)\mu(du) - \int \nu(u)\log\nu(u)\mu(du) \right]. \tag{3}$$

The EM algorithm consists of iterating these two maximization steps. Given a current value η_c of η, the maximization with respect to the density ν is seen to yield $\nu_c(u) = q(u|\eta_c, y)$, or with multiple independent observations, $\nu_c(u_1^n) = \prod_{i=1}^n q(u_i|\eta_c, y_i)$. This is often called the posterior density. Once ν_c is determined the second maximization - updating the parameters - involves only the first term in equation (3).

In the present context we initialize the algorithm with the prior model η_0 and we iterate the following two steps:

E Step: Compute the posterior law on $(\beta_i, \tau_i), i = 1, \ldots, n$ as a product of the following distributions which have a density in β for each τ and are discrete in τ for each β:

$$\nu_{l,i}(\beta, \tau) = \frac{q(y_i|\beta, \alpha_{\tau,l}) q(\beta|\Gamma_{g,l}^{\tau}) \rho_l(\tau)}{\sum_{\tau'} \int q(y_i|\beta', \alpha_{\tau',l}) q(\beta'|\Gamma_{g,l}^{\tau'}) \rho_l(\tau') d\beta'}$$

M Step:

$$\eta_{l+1} = \underset{\eta}{\operatorname{argmax}} E_{\nu_l(d\xi_1^n)}(\log q(\eta, \beta_1^n, \tau_1^n | y_1^n)).$$

4.1. Fast approximation with modes

The expressions in the M step require the computation of expectations with respect to $\nu_{i,l}(\beta, \tau)$ which has no simple form. To overcome this obstacle this distribution is approximated by the Dirac law $\nu_{i,l}^*(d\beta_{i,\tau}, \tau) = \delta_{\beta_{i,\tau}^*}$ where for each component τ, $\beta_{i,\tau}^*$ maximize the conditional distribution on β.

$$\beta_{i,\tau}^* = \arg\max_{\beta} \log q(\beta|\alpha_{\tau,l}, \sigma_{\tau,l}, \Gamma_{g,l}^{\tau}, y_i) = \arg\min_{\beta} \left\{ \frac{1}{2} \beta^t R_{g,l}^{\tau} \beta + \frac{1}{2\sigma_{l,\tau}^2} |y_i - K_p^{\beta} \alpha_{\tau,l}|^2 \right\},$$

where $R_{g,l}^{\tau} = (\Gamma_{g,l}^{\tau})^{-1}$. We then approximate the joint posterior on (β_i, τ_i) as a discrete distribution concentrated at the T points $\beta_{i,\tau}^*$ with weights given by

$$w_l(\tau) = \frac{q(y_i|\beta_{i,\tau}^*, \alpha_{\tau,l}) q(\beta_{i,\tau}^*|\Gamma_{g,l}^{\tau}) \rho_l(\tau)}{\sum_{\tau'} q(y_i|\beta_{i,\tau'}^*, \alpha_{\tau',l}) q(\beta_{i,\tau'}^*|\Gamma_{g,l}^{\tau'}) \rho_l(\tau')}. \quad (4)$$

4.2. Using a stochastic version of the EM algorithm

An alternative to the computation of the E-step in a complex nonlinear context, is to use the stochastic approximation of the EM algorithm (SAEM) coupled with an MCMC procedure ([5]). Each iteration of this algorithm consists of three steps: (i) the missing data, here the deformation coefficients, are drawn using a transition probability of a convergent Markov Chain having the posterior distribution as stationary distribution, (ii) a stochastic approximation is done on the complete likelihood using the simulated values of the missing data, (iii) the parameters are updated in the M-step.

- Simulation step : $\beta^{l+1} \sim \Pi_{\theta_l}(\beta^l, \cdot)$
- Stochastic approximation :

  $Q_{l+1}(\theta) = Q_l(\theta) + \Delta_l[\log q(y, \beta^{l+1}|\theta) - Q_l(\theta)]$ where (Δ_l) is a decreasing sequence of positive step-sizes.
- Maximization step : $\theta_{l+1} = \operatorname{argmax} Q_{l+1}(\theta)$

The almost sure convergence of this algorithm toward a (local) maximum of the observed likelihood was studied and proved under general regularity assumptions in ([5]).

5. EXPERIMENTS

We illustrate this theoretical framework with the US Postal handwritten digit data base. In this context, it is possible to compare various model settings in terms of classification rates, although our goal is not to obtain optimal results.

The classification is performed by computing the maximum posterior on class given the image. The likelihood terms involve an integral over the hidden variables which is replaced again by the mode. Some classification results as a function of number of components per class and size of training set are given in table 1. We optimized the hyper-parameters involved in the model (a_g, σ_g, σ_p) by looking at the classification rates.

Nb. of components	1	2	3	5	10
20 per class	6.58	6.13	5.28	9.57	9.72
100 per class	9.42	4.98	4.58	5.13	**4.136**

Table 1. Classification results as a function of number of components per image and size of training set.

5.1. The estimated template and noise variance

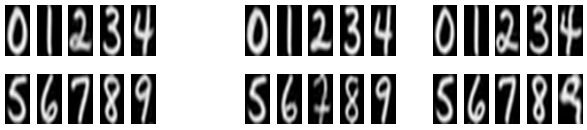

Fig. 1. Left: one component prototype. Right: 2 components prototypes

In figure 1, we show the templates of the 10 classes estimated with the mode approximation with 20 (resp 40) images per class. For the prior distribution of the template we choose a mean equal to the background value $\mu_p = 0$. In this setting the first iteration of the EM algorithm yields to $\beta_i^* = 0$ so that the resulting estimated template is the simple mean of the training images. They are blurred because of the geometrical variability within each class. As the iterations proceed, the prototypes present higher contrast thanks to the nonrigid registration.

The variance σ also evolves through the iterations. It starts to increase in the first step because of both the geometric and photometric variability. As the algorithm proceeds it decreases since the photometric variations due to geometric variability are reduced. With one component per class, some classes such as "2" or "4" have a higher noise variance than others, which seems to imply that they require more than one template. Increasing the number of components (2 are enough) this phenomenon disappears.

5.2. The estimated geometric distribution

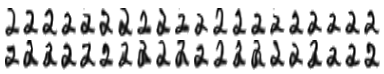

Fig. 2. Top: Synthesized 2's with template from second component of figure 1 and proper covariance. Bottom: Same template using covariance matrix of other 2 component.

To illustrate the form of the estimated geometric distribution, we show (figure 2, 1st row) 20 synthesized examples of class 2 (with loop) using its estimated photometric prototype and geometric covariance matrix. By contrast, in the second row, we show simulations using the same estimated prototype with the geometric distribution of the other class 2 cluster. The deformed 2's are still realistic but not as natural looking as the previous ones, implying a non trivial estimated geometric covariance which differs significantly from one component to another. This is also observed from one class to another as can be viewed in figure 5.2 where the 3's are synthesized using a 3 photometric prototype but the geometric covariance of digit 2.

5.3. In presence of noise

In figure 4 are shown the single component results when the training set is noisy, comparing the mode approximation to the stochastic EM algorithm. With the mode approximation, the final prototypes are less contrasted and realistic than those given with the stochastic EM algorithm. It seems that because of the presence of multiple local minima the gradient descent used in the mode approximation converges toward a "wrong" minimum. This is not as severe with stochastic simulation which allows a larger exploration of the geometric parameter space.

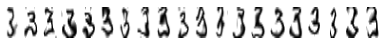

Fig. 3. 20 synthesized examples of each class

Fig. 4. Left: prototypes in noisy framework learned with the mode approximation. Right: prototypes in noisy framework learned with the stochastic EM algorithm

6. CONCLUSION

We have proposed a coherent statistical framework for dense template models and described two methods to compute the maximum posterior estimation. The results on likelihood based classification of handwritten digits, with very small training sets, demonstrates that this formulation is of practical use and gives acceptable classification as shown in table 1. The deformation model employed here does not necessarily produce diffeomorphisms, leading to certain difficulties such as behavior near the boundaries of the domain. Using diffeomorphisms of the domain onto itself, as proposed in [6], would eliminate these problems perhaps yielding a more stable algorithm.

7. REFERENCES

[1] C. A. Glasbey and K. V. Mardia, "A penalised likelihood approach to image warping," *Journal of the Royal Statistical Society, Series B*, vol. 63, pp. 465–492, 2001.

[2] Y. Amit, U. Grenander, and M. Piccioni, "Structural image restoration through deformable template," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 376–387, 1991.

[3] S. Allassonnière, Y. Amit, and A. Trouv´e, "Toward a coherent statistical framework for dense deformable template estimation," .

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 1, pp. 1–22, 1977.

[5] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of em with a mcmc procedure," *ESAIM P& S, Vol. 8*, pp. 115–131, 2004.

[6] M. I. Miller, A. Trouv´e, and L. Younes, "On the metrics and euler-lagrange equations of computational anatomy," *Annual Review of biomedical Engineering*, vol. 4, pp. 375–405, 2002.