Bayesian inference of models and hyper-parameters for robust optic-flow estimation

Patrick Héas, Cédric Herzet, Etienne Mémin

Bayesian inference of models and hyper-parameters for robust optic-flow estimation

Patrick Héas, Cédric Herzet and Etienne Mémin
INRIA Centre de Rennes Bretagne-Atlantique, France,
{Patrick.Heas, Cedric.Herzet, Etienne.Memin}@inria.fr

***Abstract*— Selecting optimal models and hyper-parameters is crucial for accurate optic-flow estimation. This paper provides a solution to the problem in a generic Bayesian framework. The method is based on a conditional model linking the image intensity function, the unknown velocity field, hyper-parameters and the *prior* and *likelihood* motion models. Inference is performed on each of the three-level of this so-defined hierarchical model by maximization of marginalized *a posteriori* probability distribution functions. In particular, the first level is used to achieve motion estimation in a classical *a posteriori* scheme. By marginalizing out the motion variable, the second level enables to infer regularization coefficients and hyper-parameters of non-Gaussian M-estimators commonly used in robust statistics. The last level of the hierarchy is used for selection of the *likelihood* and *prior* motion models conditioned to the image data. The method is evaluated on image sequences of fluid flows and from the "Middlebury" database. Experiments prove that applying the proposed inference strategy yields better results than manually tuning smoothing parameters or discontinuity preserving cost functions of the state-of-the-art methods.**

Choosing appropriate models and fixing hyper-parameters given a couple of images is a tricky and often hidden process in optic-flow estimation. Most of the motion estimators proposed so far provide a large variability of estimates according to different combination of hyper-parameter values. Moreover, this phenomenon can be drastically accentuated if we consider different regularization and data models. Generally the users of such techniques have to rely on successive trials and a empirical strategy for fixing the hyper-parameters values and choosing the adequate model. Besides being computationally expensive, this strategy may produce a catastrophic estimate without any relevant feedback for the end-user, especially when motions are complex deformations or for non-conventional imagery. Imposing hard values on these parameters may yield also poor results when the lighting conditions or if the underlying motions differ from those on which the system has been calibrated. At the extreme, the estimate may be either too smooth or at the opposite it may exhibit inexistant strong motion discontinuities. However, Bayesian analysis has been extensively studied in the past for hyper-parameter estimation and for model selection [23] [33] [34]. In particular, in the context of interpolation of noisy data, a powerful hierarchical Bayesian model has been proposed in the seminal work of [31]. In optic-flow estimation, this framework has been used in [26] for hyper-parameter estimation. However, it does not consider model deviations from Gaussianity. Moreover, very little emphasis has been devoted in motion estimation works to the problem of model selection. In particular, except in a particular case [18], no proper Bayesian formulation has been proposed in the literature for inference of optic-flow data and regularization models.

This work complements state-of-the-art in Bayesian inference techniques for optic-flow with a generic and robust modeling framework. It makes possible the design of non-parametrical estimation methods, able to reliably decide among several data and regularization models with optimal tuning of regularization coefficients and robust model hyper-parameters. The remainder of the paper is organized as follows. In the two next sections, in order to motivate our approach, we make a brief overview on optic-flow estimation techniques and review state-of-the-art inference strategies. Then, in section III, we present our optic-flow hierarchical model together with the associated Bayesian inference levels. Finally, in section IV an exhaustive evaluation of the generic method performed on challenging image sequences of turbulent diffusive flows and computer vision scenes demonstrate the interest of our approach.

I. Bayesian formulation of optic-flow estimation

A. Aperture problem

Optic-flow estimation is well known to be a difficult ill-posed inverse problem in which one aims at estimating the apparent motion of a 3D scene through spatial and temporal image intensity $I(\mathbf{s},t)$ variations. The optic-flow, identified by a 2D velocity fields, $\mathbf{v}(\mathbf{s},t) : \Omega \times \mathbb{R}^+ \to \mathbb{R}^2$, over the image grid Ω is identical to the projection on the image plane of the scene velocity when considering rigid motion and stable lighting conditions. In this situation, motion $\mathbf{v} = (u,v)^T$ respects the standard Optical Flow Constraint (OFC) equation which reads:

$$\frac{dI}{dt} = \partial_t I + \nabla I \cdot \mathbf{v} = 0. \tag{1}$$

For specific image modalities or for observation of particular phenomena, many other brightness evolution models have been proposed in the literature to link the image intensity function to the sought velocity fields [8], [13], [17], [19], [29]. However, all these evolution models remain underconstrained, as they provide only one equation for two unknowns (u,v) at each spatio-temporal location $(\mathbf{s},t)$. To deal with this underconstrained estimation problem, the so called aperture problem, the most common setting consists in enforcing some spatial coherence to the solution. This coherence is imposed

either explicitly through a regularization functional [18] [20] [45] or implicitly, with a parametric (often local and polynomial) spatial representation of the solution [9] [10] [30]. Note that implicit regularization may be seen as a particular case, where no explicit regularization functional is specified. Therefore, we concentrate on explicit regularization schemes in the following.

B. Gaussian models

Explicit regularization schemes in their simplest form define the estimation problem through the minimization of an energy functional composed of two terms balanced by a regularization coefficient γ (> 0):

$$L(I, \mathbf{v}, \alpha, \beta) = f_d(I, \mathbf{v}) + \gamma f_r(\mathbf{v}), \tag{2}$$

where α and β are two positive parameters forming the ratio $\gamma = \alpha/\beta$.

The first term, $f_d(I, \mathbf{v})$ (the "data term"), penalizes discrepancies between observed image temporal differences denoted by $\mathbf{I}_t = \{I_t(\mathbf{s}); \mathbf{s} \in \Omega | I_t(\mathbf{s}) = I(\mathbf{s}, t+1) - I(\mathbf{s}, t)\}$ and the unknowns $\mathbf{v}$ by mean of the observation operator Π_d. For the sake of conciseness, we will limit the presentation to linear operators. Note however that extension to the non-linear case by successive linearization is possible [10]. Based on (1), one can for instance define an observation operator by using the function Π_d:

$$\Pi_d(\mathbf{v}(\mathbf{s}), I) = \nabla I(\mathbf{s}, t+1) \cdot \mathbf{v}(\mathbf{s}), \tag{3}$$

which applies on the unknown motion and which depends on the image spatial gradients. However as mentioned earlier, many others operator could be defined. Discretizing the motion field on the pixel grid Ω with a unit time interval and considering a quadratic penalization results in data terms of the form:

$$f_d(I, \mathbf{v}) = \frac{1}{2} \sum_{\mathbf{s} \in \Omega} \left(I_t(\mathbf{s}) + \Pi_d(\mathbf{v}(\mathbf{s})) \right)^2. \tag{4}$$

Let us now define the data model variable $\mathcal{D}$. This variable is a binary variable defined on the finite set of possible functions Π_d. For functions satisfying $\exp(-f_d) \in \mathcal{L}^1$, an equivalent Gibbs Probability Distribution Function (PDF) *likelihood* can be written as [15]:

$$p(\mathbf{I}_t | \mathbf{v}, \beta, \mathcal{D}) = \frac{\exp\{-\beta f_d(I, \mathbf{v})\}}{Z_{f_d}(\beta)}, \tag{5}$$

where $Z_{f_d}(\beta)$ denotes a normalization constant also called partition function. For quadratic penalization in (4), Z_{f_d} is the partition function of an uncorrelated multidimensional Gaussian variable $\mathbf{I}_t$. Therefore it depends only on the inverse variance β, which is identical for all the components.

The second term, $f_r(\mathbf{v})$ (the "regularization term"), encourages the solution to follow some prior smoothness model formalized with function Π_r:

$$f_r(\mathbf{v}) \quad = \quad \frac{1}{2} \sum_{\mathbf{s} \in \Omega} |\Pi_r(u, \mathbf{s})|^2 + |\Pi_r(v, \mathbf{s})|^2 \tag{6}$$

where $|.|^2$ denotes the $\mathcal{L}^2$ norm. For the sake of conciseness, we will also limit the presentation to linear Π_r. In this quadratic penalization framework, a first-order regularizer enforcing weak spatial gradients of the two components u and v of the velocity field $\mathbf{v}$ is very often used [20]. This regularizer is obtained from (6) defining the following function:

$$\Pi_r(u, \mathbf{s}) = \nabla u(\mathbf{s}), \ \Pi_r(v, \mathbf{s}) = \nabla v(\mathbf{s}). \tag{7}$$

Considering motion variables on the pixel grid requires the approximation of continuous spatial derivatives by discrete functions. For regular pixel grids, this can be done using for instance finite difference schemes or finite elements. The spatial discretization results in the definition of a global energy composed of local potentials $V_r^2(v, \mathbf{s}, \mathbf{t})$ relative to a clique system $\mathcal{C}$ depending on the differential operator considered and on the accuracy (also called the order) of the differentiation scheme. After discretization, the regularization term is rewritten as:

$$f_r(\mathbf{v}) = \frac{1}{2} \sum_{<\mathbf{s},\mathbf{t}> \in \mathcal{C}} V_r^2(u, \mathbf{s}, \mathbf{t}) + V_r^2(v, \mathbf{s}, \mathbf{t}). \tag{8}$$

where $V_r^2(u, \mathbf{s}, \mathbf{t})$ is a quadratic energy defined on mutual neighboring sites pairs (so called clique) $<\mathbf{s}, \mathbf{t}>$ belonging to the set $\mathcal{C}$. For instance, discretizing spatial gradients in (6) using non-centered first-order finite difference schemes results in local potential energies of the form $V_r^2(u, \mathbf{s}, \mathbf{t}) = (u(\mathbf{s}) - u(\mathbf{t}))^2$ on the binary cliques $< \mathbf{s}, \mathbf{t} >$ associated to a first-order neighboorhood system. Let us now define the regularization model variable $\mathcal{R}$. This variable is a binary variable defined on the finite set of possible regularization functions Π_r. As previously for the data-model, regularization terms f_r satisfying $\exp\{-f_r\} \in \mathcal{L}^1$ weighted by a positive parameter α forms the energy of a Gibbs PDF *prior* :

$$p(\mathbf{v} | \alpha, \mathcal{R}) = \frac{\exp\{-\alpha f_r(\mathbf{v})\}}{Z_{f_r}(\alpha)}, \tag{9}$$

where $Z_{f_r}(\alpha)$ denotes the partition function. Note that for quadratic energies such as in (8), the Gibbs *prior* is equivalent to a Gauss-Markov random field and α represents the inverse variance of the regularization model error [15].

The motion *posterior* PDF is formed according to Bayes rule $p(\mathbf{v} | \mathbf{I}_t, \alpha, \beta, \mathcal{D}, \mathcal{R}) \propto p(\mathbf{I}_t | \mathbf{v}, \beta, \mathcal{D}) p(\mathbf{v} | \alpha, \mathcal{R})$ as:

$$p(\mathbf{v} | \mathbf{I}_t, \alpha, \beta, \mathcal{D}, \mathcal{R}) = \frac{\exp\{-\beta L(I, \mathbf{v}, \alpha, \beta)\}}{Z_L(I, \alpha, \beta)}, \tag{10}$$

where $Z_L(I, \alpha, \beta)$ denotes the associated partition function. For quadratic energies, *posterior* PDFs are Gauss-Markov random fields. The energy minimized in (2) for the estimation of motion $\mathbf{v}$ constitutes, up to a multiplicative factor, the energy of the *posterior*. Thus, the maximum of the motion *posterior* probability is reached for the estimate $\hat{\mathbf{v}}$ satisfying:

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}} L(I, \mathbf{v}, \alpha, \beta). \tag{11}$$

The optic-flow Maximum *A Posteriori* (MAP) estimate is obtained by minimizing directly the *posterior* energy given

in (2). An equivalent approach consists in solving the associated Euler-Lagrange equations (derived from the variational formulation of a functional minimization on a continuous image domain). In both cases, efficient deterministic optimization techniques such as the Conjugate Gradient Squared (CGS) algorithm [36] or various multi-grid strategies [5] [32] can be employed to access the minimum. For quadratic penalization and linear models, the *posterior* energies are convex functionals whose global minimum can be reached with such deterministic methods.

C. M-estimators and semi-quadratic models

It is well known that the $\mathcal{L}^2$ norm (*i.e.* Gaussian PDFs for linear models) is limited in the context of strong model deviations. In particular, this norm is inefficient when dealing with observation outliers or motion spatial discontinuities. Such limitations are overcome using so called robust cost functions. Indeed, such penalty functions, also called M-estimators in robust statistics and related to non-Gaussian distributions, are considered in this context [4] [14]. M-estimators are functions defined by the inferior envelop of a family of parabolas. Note that to define an M-estimator, the family of parabolas must have some properties given in appendix A. It follows that M-estimators possess some equivalent semi-quadratic formulations. They therefore constitute interesting non-quadratic functions since by convex duality they locally preserve the convex properties of quadratic functionals [24] [38]. Let us redefine the optic-flow components in this context.

Let ρ_{τ_d} be an M-estimator and τ_d its free parameter. The associated family of parabolas, denoted by $\varphi_{\tau_d}(z_d)$ is continuously indexed by z_d, which is here defined as a normalized auxiliary variables ($z_d \in]0,1]$) and have an inferior envelop defining ρ_{τ_d}. For the data term, we thus define a family of *likelihood* energies by associating to each pixel a parabola family indexed by $z_d(\mathbf{s})$:

$$f_d(I,\mathbf{v},\mathbf{z}_d,\tau_d)=\frac{1}{2\tau_d}\sum_{\mathbf{s}}\varphi_{\tau_d}(V_d(I,\mathbf{v},\mathbf{s}),z_d(\mathbf{s})), \quad (12)$$

where $\mathbf{z}_d=\{z_d(\mathbf{s})\}$ and V_d^2 represents a quadratic energy encoding the deviation to the observations:

$$V_d^2(I,\mathbf{v},\mathbf{s})=(I_t(\mathbf{s})+\Pi_d(\mathbf{v}(\mathbf{s})))^2. \quad (13)$$

A parabola is a semi-quadratic functional of the form:

$$\varphi_{\tau_d}(V_d(I,\mathbf{v},\mathbf{s}),z_d(\mathbf{s}))=\tau_d z_d(\mathbf{s})V_d^2(I,\mathbf{v},\mathbf{s})+\psi(z_d(\mathbf{s})), \quad (14)$$

where the form of the function ψ is defined in appendix A. To make the energy (14) equivalent to the penalization by the M-estimator $\rho_{z_d}(V_d(I,\mathbf{v},\mathbf{s}))$, the indices $\hat{\mathbf{z}}_d=\{\hat{z}_d\}$ of the parabolas must satisfy:

$$\hat{z}_d(\mathbf{s}) \quad = \quad \frac{\rho'_{\tau_d}(V_d(I,\mathbf{v},\mathbf{s}))}{2\tau_d V_d(I,\mathbf{v},\mathbf{s})}, \quad (15)$$

given some quadratic energies $\{V_d(I,\mathbf{v},\mathbf{s})\}$. The *likelihood* PDF can be redefined in the robust context as

$$p(\mathbf{I}_t|\mathbf{z}_d,\mathbf{v},\beta,\tau_d,\mathcal{D})=\frac{\exp^{-\beta f_d(I,\mathbf{v},\mathbf{z}_d,\tau_d)}}{Z_{f_d}(\beta,\mathbf{z}_d,\tau_d)}. \quad (16)$$

For the regularization term, deviations from the prior motion smoothness are analogously defined through some parabola family φ_{τ_r} depending on index variable z_r. The inferior envelope of the parabola family define some M-estimator ρ_{τ_r} parametrized by τ_r. Regularizers given in (6) are thus redefined using a family of *prior* energies [4]: $f_r(\mathbf{v},\mathbf{z}_r,\tau_r)=$

$$\frac{1}{2\tau_r}\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}}\varphi_{\tau_r}(V_r(u,\mathbf{s},\mathbf{t}),z_r(\mathbf{s},\mathbf{t}))+\varphi_{\tau_r}(V_r(v,\mathbf{s},\mathbf{t}),z_r(\mathbf{s},\mathbf{t})), \quad (17)$$

where the parabolas indices are $\mathbf{z}_r=\{z_r(\mathbf{s},\mathbf{t})\}$ and where each parabola is a semi-quadratic energy of the form:

$$\varphi_{\tau_r}(V_r(u,\mathbf{s},\mathbf{t}),z_r(\mathbf{s},\mathbf{t}))=\tau_r z_r(\mathbf{s},\mathbf{t})V_r^2(u,\mathbf{s},\mathbf{t})+\psi(z_r(\mathbf{s},\mathbf{t})). \quad (18)$$

To make the energy (17) equivalent to the penalization resulting from the associated M-estimator that is: $\rho_{z_r}(V_r(u,\mathbf{s},\mathbf{t}))+\rho_{z_r}(V_r(v,\mathbf{s},\mathbf{t}))$, the indices $\hat{\mathbf{z}}_r=\{\hat{z}_r\}$ of the parabolas must satisfy:

$$2\hat{z}_r(\mathbf{s},\mathbf{t}) \quad = \quad \frac{\rho'_{\tau_r}(V_r(u,\mathbf{s},\mathbf{t}))}{2\tau_r V_r(u,\mathbf{s},\mathbf{t})}+\frac{\rho'_{\tau_r}(V_r(v,\mathbf{s},\mathbf{t}))}{2\tau_r V_r(v,\mathbf{s},\mathbf{t})}, \quad (19)$$

given some quadratic energies $\{V_r(u,\mathbf{s},\mathbf{t}),V_r(v,\mathbf{s},\mathbf{t})\}$. The *prior* can then be redefined in the robust context as:

$$p(\mathbf{v}|\mathbf{z}_r,\alpha,\tau_r,\mathcal{R})=\frac{\exp^{-\alpha f_r(\mathbf{v},\mathbf{z}_r,\tau_r)}}{Z_{f_r}(\alpha,\mathbf{z}_r,\tau_r)}. \quad (20)$$

Let us denote the hyper-parameter vector by $\boldsymbol{\theta}=(\alpha,\beta,\tau_d,\tau_r)^T$. Using Bayes' relation, the motion *posterior* PDF is redefined in the robust context as:

$$\begin{aligned} p(\mathbf{v}|\mathbf{z}_d,\mathbf{z}_r,\mathbf{I}_t,\boldsymbol{\theta},\mathcal{D},\mathcal{R}) &\propto p(\mathbf{I}_t|\mathbf{z}_d,\mathbf{v},\beta,\tau_d,\mathcal{D})p(\mathbf{v}|\mathbf{z}_r,\alpha,\tau_r,\mathcal{R}) \\ &= \frac{\exp^{-\beta L(I,\mathbf{v},\mathbf{z}_d,\mathbf{z}_r,\boldsymbol{\theta})}}{Z_L(I,\mathbf{z}_d,\mathbf{z}_r,\boldsymbol{\theta})}, \end{aligned} \quad (21)$$

with the energy:

$$L(I,\mathbf{v},\mathbf{z}_d,\mathbf{z}_r,\boldsymbol{\theta})=f_d(I,\mathbf{v},\mathbf{z}_d,\tau_d)+\gamma f_r(\mathbf{v},\mathbf{z}_r,\tau_r). \quad (22)$$

Based on the minimization of the energy (22) w.r.t. motion, we obtain the MAP estimate $\hat{\mathbf{v}}$. Furthermore, indices $\hat{\mathbf{z}}=\{\hat{\mathbf{z}}_d,\hat{\mathbf{z}}_r\}$ satisfying the equivalence with M-estimators constitute the minimizer of (22) and can also be interpreted as MAP estimates. As a matter of fact, the motion-indices joint *posterior* PDF may be defined as:

$$p(\mathbf{v},\mathbf{z}_d,\mathbf{z}_r|\mathbf{I}_t,\boldsymbol{\theta},\mathcal{D},\mathcal{R})\propto\exp^{-\beta L(I,\mathbf{v},\mathbf{z}_d,\mathbf{z}_r,\boldsymbol{\theta})}. \quad (23)$$

This joint *posterior* probability is built using the relation:

$$\begin{aligned} &p(\mathbf{v},\mathbf{z}_d,\mathbf{z}_r|\mathbf{I}_t,\boldsymbol{\theta},\mathcal{D},\mathcal{R}) \\ &\propto p(\mathbf{v}|\mathbf{z}_d,\mathbf{z}_r,\mathbf{I}_t,\boldsymbol{\theta},\mathcal{D},\mathcal{R})p(\mathbf{z}_d|\beta,\mathcal{D})p(\mathbf{z}_r|\alpha,\mathcal{R}), \end{aligned} \quad (24)$$

where the *prior* PDFs $p(\mathbf{z}_d|\beta,\mathcal{D})$ and $p(\mathbf{z}_r|\alpha,\mathcal{R})$ on indices $\mathbf{z}_d$ and $\mathbf{z}_r$ are chosen in order to induce the energy defined in (22). Moreover, following appendix A, $\hat{\mathbf{z}}$ is the minimizer of the *posterior* energy:

$$\hat{\mathbf{z}}=\arg\min_{\mathbf{z}_d,\mathbf{z}_r}L(I,\mathbf{v},\mathbf{z}_d,\mathbf{z}_r,\boldsymbol{\theta}). \quad (25)$$

Therefore, one finally obtains that indices satisfying the equivalence with M-estimators given by (15) and (19) are estimates maximizing the joint *posterior*:

$$\{\hat{\mathbf{v}}, \hat{\mathbf{z}}\} = \arg \max_{\mathbf{v},\mathbf{z}_d,\mathbf{z}_r} p(\mathbf{v}, \mathbf{z}_d, \mathbf{z}_r | \mathbf{I}_t, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R}). \quad (26)$$

The tuning of the hyper-parameters $\boldsymbol{\theta}$ and the selection of the optimal models $(\mathcal{D}, \mathcal{R})$ for optic-flow is a crucial problem which can yield drastically different results.

II. Related work on hyper-parameter estimation and model selection

The goal we are pursuing in this work aims at complementing optical flow estimation techniques with a Bayesian method inferring the different parameters involved in the models described previously. To the best of our knowledge, only very little attention has been devoted to this non trivial problem in the literature. The following sections give an overview on Bayesian and deterministic state-of-the-art methods in this field.

A. Bayesian inference

Bayesian analysis has been intensively studied in the past for estimation of hyper-parameters [33] [34] or degrees of freedom [39], and model selection in different contexts than motion estimation [23]. In particular, for interpolation of noisy data a Bayesian *evidence* framework relying on successive marginalization of the joint posterior probability has been proposed in the seminal work of [31]. In the case of Gaussian models, this framework has been recently applied to the context of optic-flow estimation. In particular, estimation of motion regularization parameters [25], the weighting of concurrent *likelihood* models [26], or the selection of a *prior* motion scaling model [18] constitute noticeable achievements.

However, state-of-the-art inference techniques remain limited in motion estimation issues since they do not consider *likelihood* and *prior* model deviating from Gaussianity. Such non-Gaussian models are nevertheless very common in computer vision, where we have to cope with motion discontinuities and observation outliers due to noise or varying lighting conditions. The common non-Gaussian robust statistics used to manage such problems raise the additional difficulty of the choice of the robust norm and of its hyper-parameters, since in general it is a parametrical model. This choice is crucial and different tuning of these parameters can lead to motion estimates that differ drastically. Moreover, although being crucial for accurate motion measurement, surprisingly only very little emphasis has been devoted in the computer vision literature to the problem of model selection for optic-flow estimation. In particular, except in the particular case of [18], no proper Bayesian formulation considering the marginalization of the joint posterior distribution with respect to the motion field and the hyper-parameter vector has been proposed in the literature for the selection of optimal *prior* and *likelihood* optic-flow models.

Before exploring more in details such a formulation, in the following, we briefly describe deterministic techniques for hyper-parameter inference, point out some similarities and emphasize differences with the Bayesian point of view.

B. Deterministic inference

Most of deterministic methods for hyper-parameter inference rely on the minimization of a mean square error estimate between the true and the estimated deterministic hyper-parameter, or equivalently, on the minimization of the conditional variance and bias of the hyper-parameter estimator [28] [41] [43]. In this context, Stein unbiased risk estimate has been successfully applied to infer optic-flow hyper-parameters of a *Horn & Schunck* model [35], or of a local *Lucas & Kanade* model [40]. Generalized cross validation is another approach to minimize a mean square error estimate, which has been applied to the optic-flow context [11]. It is based on the idea that the best hyper-parameter for the observations is the one that best predicts each observation as a function of the others. There exist also heuristic approaches in optic-flow literature for choosing regularization hyper-parameters. L-Curve and U-Curve methods are based on simultaneous minimization of the norm of the regularized residual and the norm of the regularized solution [27]. Hyper-parameter and model selection based on learning strategies should also be mentioned, although they are essentially different since they rely on training data [22] [42].

It is interesting to notice that some of the previous approaches share in their analysis common quantities with Bayesian methods: the trace of the inverse matrix of the functional Hessian in approaches based on the mean square error estimate, or the balance between the regularized residual and the regularized solution in heuristic methods. Nevertheless, the deterministic and the Bayesian approaches are generally different for several reasons [31] [37]. In particular, because deterministic inference is derived from a different perspective: the selection of the most probable deterministic parameter rather than a measure of plausibility of the entire posterior ensemble.

III. Bayesian hierarchical modeling of optic-flow

Let us describe the optic-flow estimation problem in the light of the *evidence* framework. The problem is formalized by a hierarchical model factorizing the joint probability of the following variables: the image temporal difference $\mathbf{I}_t$, motion $\mathbf{v}$ and the parabolas indices $\{\mathbf{z}_d, \mathbf{z}_r\}$, hyper-parameters $\boldsymbol{\theta}$, and the model $\mathcal{D}$ and $\mathcal{R}$. Model $\mathcal{D}$ (resp. $\mathcal{R}$) is redefined as the binary variable on the finite set $\{(\Pi_d, \rho_{\tau_d})\}$ (resp. $\{(\Pi_r, \rho_{\tau_r})\}$). This ensemble is the Cartesian products between the finite set of possible observation function Π_d (resp. regularization function Π_r) and the finite set of possible penalization norms given by the M-estimators ρ_{τ_d} (resp. ρ_{τ_r}). The joint probability

factorization reads:

$$p(\mathbf{I}_t, \mathbf{v}, \mathbf{z}_d, \mathbf{z}_r, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R}) = \underbrace{p(\mathbf{I}_t|\mathbf{v}, \mathbf{z}_d, \boldsymbol{\theta}, \mathcal{D})p(\mathbf{z}_d|\boldsymbol{\theta}, \mathcal{D})}_{\text{robust likelihood}} \quad (27)$$
$$\underbrace{p(\mathbf{v}|\mathbf{z}_r, \boldsymbol{\theta}, \mathcal{R})p(\mathbf{z}_r|\boldsymbol{\theta}, \mathcal{R})}_{\text{robust motion prior}} \quad \underbrace{p(\boldsymbol{\theta})p(\mathcal{D}, \mathcal{R})}_{\text{hyper-par. \& model prior}}$$

Using hierarchical model (27), we obtain three levels of inference for the estimation of $(\hat{\mathbf{v}}, \hat{\mathbf{z}}_d, \hat{\mathbf{z}}_r), \hat{\boldsymbol{\theta}}$ and $(\hat{\mathcal{D}}, \hat{\mathcal{R}})$, solution of the following set of problems:

$$(\hat{\mathbf{v}}, \hat{\mathbf{z}}_d, \hat{\mathbf{z}}_r) = \arg\max_{(\mathbf{v}, \mathbf{z}_d, \mathbf{z}_r)} \left\{ p(\mathbf{v}, \mathbf{z}_d, \mathbf{z}_r|\mathbf{I}_t, \hat{\boldsymbol{\theta}}, \hat{\mathcal{D}}, \hat{\mathcal{R}}) \right\}, \quad (28)$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\boldsymbol{\theta}|\mathbf{I}_t, \mathbf{z}^\star(\boldsymbol{\theta}, \hat{\mathbf{v}}), \hat{\mathcal{D}}, \hat{\mathcal{R}}) \right\}, \quad (29)$$

$$(\hat{\mathcal{D}}, \hat{\mathcal{R}}) = \arg\max_{(\mathcal{D}, \mathcal{R})} \left\{ p(\mathcal{D}, \mathcal{R}|\mathbf{I}_t, \mathbf{z}^\star(\hat{\boldsymbol{\theta}}, \hat{\mathbf{v}})) \right\}, \quad (30)$$

where $\mathbf{z}^\star(\boldsymbol{\theta}, \mathbf{v}) = \{\{z_d^\star(\boldsymbol{\theta}, \mathbf{v}(\mathbf{s}))\}, \{z_r^\star(\boldsymbol{\theta}, \mathbf{v}(\mathbf{s}, \mathbf{t}))\}\}$ with

$$z_d^\star(\boldsymbol{\theta}, \mathbf{v}(\mathbf{s})) = \frac{\rho'_{\tau_d}(V_d(I, \mathbf{v}, \mathbf{s}))}{2\tau_d V_d(I, \mathbf{v}, \mathbf{s})}, \quad (31)$$

$$z_r^\star(\boldsymbol{\theta}, \mathbf{v}(\mathbf{s}, \mathbf{t})) = \frac{1}{2}\Big(\frac{\rho'_{\tau_r}(V_r(u, \mathbf{s}, \mathbf{t}))}{2\tau_r V_r(u, \mathbf{s}, \mathbf{t})} + \frac{\rho'_{\tau_r}(V_r(v, \mathbf{s}, \mathbf{t}))}{2\tau_r V_r(v, \mathbf{s}, \mathbf{t})} \Big). \quad (32)$$

The system (28)-(30) is inspired from the Bayesian evidence framework proposed in [31]. It defines three levels of inference. In the first level (28), $\mathbf{v}$ and $(\mathbf{z}_d, \mathbf{z}_r)$ are estimated by relying on hyper-parameter and model estimates $(\hat{\boldsymbol{\theta}}, \hat{\mathcal{D}}, \hat{\mathcal{R}})$. In the second level (29), the dependence on $\mathbf{v}$ is marginalized out and $\boldsymbol{\theta}$ is inferred by assuming $(\mathcal{D}, \mathcal{R}) = (\hat{\mathcal{D}}, \hat{\mathcal{R}})$. Finally, in the last level (30), $(\hat{\mathcal{D}}, \hat{\mathcal{R}})$ is computed by maximizing a likelihood function in which both the dependence on $\mathbf{v}$ and $\boldsymbol{\theta}$ has been removed.

Note that the set of problems defined in (28)-(30) is slightly different from the one presented in [31] since the dependence on $(\mathbf{z}_d, \mathbf{z}_r)$ is not removed in (29)-(30). Instead, we constraint $(\mathbf{z}_d, \mathbf{z}_r)$ to have a particular structure, namely (31)-(32). As we will see in the remainder of this section, this digression from the original Bayesian evidence framework allows a tractable implementation of the inference algorithm. On the other hand, it also forces an interconnection between all level of inference: $\hat{\boldsymbol{\theta}}$ depends on $\hat{\mathbf{v}}$ through $\mathbf{z}^\star(\boldsymbol{\theta}, \hat{\mathbf{v}})$ whereas $\hat{\mathbf{v}}$ is the maximum of a function depending on $\hat{\boldsymbol{\theta}}$, etc.

In the following, we detail successively the analytical solution of the three inference problems (28)-(30).

A. Estimation of motion field $\mathbf{v}$

In this first level of inference (28), we consider the dependence of the two first levels of the hierarchy: $\mathbf{I}_t$ and $\{\mathbf{v}, \mathbf{z}_d, \mathbf{z}_r\}$. The problem is thus conditioned by some hyper-parameters $\boldsymbol{\theta} = (\alpha, \beta, \tau_d, \tau_r)^T$ and the models $(\mathcal{D}, \mathcal{R})$. As explained in section I, Bayes relation is used in order to define the motion *posterior* PDF:

$$\begin{aligned} p(\mathbf{v}|\mathbf{I}_t, \mathbf{z}_d, \mathbf{z}_r, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R}) &= \frac{p(\mathbf{I}_t|\mathbf{v}, \mathbf{z}_d, \boldsymbol{\theta}, \mathcal{D})p(\mathbf{v}|\mathbf{z}_r, \boldsymbol{\theta}, \mathcal{R})}{p(\mathbf{I}_t|\mathbf{z}_d, \mathbf{z}_r, \boldsymbol{\theta}, \mathcal{R}, \mathcal{D})} \\ &= \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}, \quad (33) \\ &\propto p(\mathbf{I}_t|\mathbf{v}, \mathbf{z}_d, \boldsymbol{\theta}, \mathcal{D})p(\mathbf{v}|\mathbf{z}_r, \boldsymbol{\theta}, \mathcal{R}), \end{aligned}$$

where the *likelihood*, the *prior* and the *posterior* are respectively defined in (16), (20) and in (21). As already mentioned in section I, this *posterior* probability is maximized in order to obtain a velocity field estimate $\hat{\mathbf{v}}$ according to a MAP criterion. Maximizing the joint indices-motion *posterior* probability $p(\mathbf{v}, \mathbf{z}_d, \mathbf{z}_r|\mathbf{I}_t, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R})$ given by (24), MAP indices $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_d, \hat{\mathbf{z}}_r\}$ are jointly inferred for a certain M-estimator. This first MAP inference level is widely used in computer vision for robust motion estimation [4].

B. Estimation of hyper-parameters $\boldsymbol{\theta}$

The second level of inference (29) considers the dependence of the image, $\mathbf{I}_t$, to the parameter vector $\boldsymbol{\theta} = (\alpha, \beta, \tau_d, \tau_r)^T$ whose components are the regularization parameters α and β and the robust parameters τ_d and τ_r. Inference of vector $\hat{\boldsymbol{\theta}}$ is conditioned to some models $(\mathcal{D}, \mathcal{R})$ and to the MAP estimate $\mathbf{z}^\star$ of the first inference level. It implies first marginalizing out the motion variable $\mathbf{v}$ and then, considering these dependences in Bayes relation, one can write:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{I}_t, \mathbf{z}^\star, \mathcal{D}, \mathcal{R}) &= \frac{p(\mathbf{I}_t|\mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R})p(\boldsymbol{\theta}|\mathbf{z}^\star, \mathcal{D}, \mathcal{R})}{p(\mathbf{I}_t|\mathbf{z}^\star, \mathcal{D}, \mathcal{R})} \\ &\propto p(\mathbf{I}_t|\mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R})p(\boldsymbol{\theta}|\mathbf{z}^\star, \mathcal{D}, \mathcal{R}) \quad (34) \end{aligned}$$

For flat *priors* $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{R})$, the MAP $\hat{\boldsymbol{\theta}}$ in (34) is simply the Maximum *Likelihood* (ML) estimate also known as the maximum of the *evidence*. Since no information is generally available on the hyper-parameter distribution, we will consider in the following a flat prior. The *evidence* probability is defined by marginalization w.r.t. the velocity field[1]:

$$p(\mathbf{I}_t|\mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R}) = \int_{\mathbb{R}^n} p(\mathbf{I}_t|\mathbf{v}, \mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{D})p(\mathbf{v}|\mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{R})\mathrm{d}\mathbf{v}. \quad (35)$$

1) Evaluation of hyper-parameter's evidence :

The *evidence* integral in (35) may be written as a partition function ratio:

$$\begin{aligned} p(\mathbf{I}_t|\mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R}) &= \int_{\mathbb{R}^n} \frac{\exp^{-\beta f_d(I, \mathbf{v}, \mathbf{z}_d^\star, \tau_d) - \alpha f_r(\mathbf{v}, \mathbf{z}_r^\star, \tau_r)}}{Z_{f_d}(\beta, \mathbf{z}_d^\star, \tau_d) Z_{f_r}(\alpha, \mathbf{z}_r^\star, \tau_r)} \mathrm{d}\mathbf{v} \\ &= \frac{Z_L(I, \alpha, \beta, \mathbf{z}^\star, \tau_d, \tau_r)}{Z_{f_d}(\beta, \mathbf{z}_d^\star, \tau_d) Z_{f_r}(\alpha, \mathbf{z}_r^\star, \tau_r)}. \quad (36) \end{aligned}$$

First, for the robust model defined in (12), the *likelihood* partition function reads[1]:

$$\begin{aligned} Z_{f_d}(\beta, \mathbf{z}_d^\star, \tau_d) &= \int_{\mathbb{R}^m} \exp\{-\beta f_d(I, \mathbf{v}, \mathbf{z}_d^\star, \tau_d)\} \mathrm{d}\mathbf{I}_t \quad (37) \\ &= \int_{\mathbb{R}^m} \exp^{-\frac{\beta}{2}\sum_{\mathbf{s}} z_d^\star(\mathbf{s}) V_d^2(I, \mathbf{v}, \mathbf{s}) + \frac{\psi(z_d^\star(\mathbf{s}))}{\tau_d}} \mathrm{d}\mathbf{I}_t. \end{aligned}$$

$\mathbf{I}_t = \{I_t(s)\}$ is a m-dimensional Gaussian vector with uncorrelated components where m is the number of point

[1] With a slight abuse of notation which is usually clear from the context, the variable is confused with its realization. Therefore, integrals are defined over the variable realizations and not over the random variable itself (*i.e.* integrals are not stochastic).

composing the image grid. The observation operator defined in (13) implies that the integration of the *likelihood* energy w.r.t. the image temporal derivative or the potential V_d are equivalent up to a sign: $\mathrm{d}I_t(\mathbf{s}) = \pm \mathrm{d}V_d(I, \mathbf{v}, \mathbf{s})$. Therefore, the partition function in (37) is the integral of the exponential of the energy of a m dimensional Gaussian with uncorrelated components of variance $(\beta z_d^\star(\mathbf{s}))^{-1}$ added to the energy of m constants of the form $\frac{\beta\psi(z_d^\star(\mathbf{s}))}{2\tau_d}$. In consequence, the *likelihood* partition function writes:

$$Z_{f_d}(\beta, \mathbf{z}_d^\star, \tau_d) = \prod_{\mathbf{s}\in\Omega} \left(\frac{2\pi}{\beta z_d^\star(\mathbf{s})}\right)^{\frac{1}{2}} \exp^{\frac{-\beta\psi(z_d^\star(\mathbf{s}))}{2\tau_d}}, \tag{38}$$

resulting in: $-\log Z_{f_d}(\beta, \mathbf{z}_d^\star, \tau_d) =$

$$\sum_{\mathbf{s}\in\Omega} -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\beta z_d^\star(\mathbf{s}) + \frac{\beta\psi(z_d^\star(\mathbf{s}))}{2\tau_d}. \tag{39}$$

Secondly, according to (17), the partition function $Z_{f_r}(\alpha, \mathbf{z}_r^\star, \tau_r)$ of the *prior* reads[1]:

$$\begin{aligned} Z_{f_r}(\alpha, \mathbf{z}_r^\star, \tau_r) &= \int_{\mathbb{R}^n} \exp^{-\alpha f_r(\mathbf{v}, \mathbf{z}_r^\star, \tau_r)}\, \mathrm{d}\mathbf{v} \qquad (40)\\ &= \int_{\mathbb{R}^n} \exp^{-\frac{\alpha}{2}\sum\limits_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} z_r^\star(\mathbf{s},\mathbf{t})V_r^2(\mathbf{v},\mathbf{s},\mathbf{t}) + \frac{\psi(z_r^\star(\mathbf{s},\mathbf{t}))}{\tau_r}}\, \mathrm{d}\mathbf{v} \end{aligned}$$

where $n = 2m$ denotes the number of unknown velocity variables. Potential are linear functions and this integral represents the partition function of a *n*-dimensional Gaussian random field. This integral is thus entirely dependent on the inverse covariance, that is to say the Hessian matrix. The analytical calculation thus requires the determination of the Hessian matrix of the *prior* energy:

$$\partial^2_{\mathbf{v}}(f_r(\mathbf{v}, \mathbf{z}_r^\star, \tau_r)) = \mathbf{\Pi}_r^T \Lambda_{\mathbf{z}_r^\star} \mathbf{\Pi}_r, \tag{41}$$

where $\Lambda_{\mathbf{z}_r^\star}$ is a diagonal matrix whose diagonal is the vector $((\mathbf{z}_r^\star)^T, (\mathbf{z}_r^\star)^T)^T$ and where $\mathbf{\Pi}_r$ is the matrix associated to function Π_r. For instance, in the case of the gradient smoothing model given in (7), each line of this matrix represents minus a discretized Laplacian operator weighted by $\mathbf{z}_r^\star$. This result can be obtained using Euler-Lagrange equations [44]. Noting that the energy of the *prior* is the energy of a centered *n*-dimensional Gaussian of covariance matrix $(\mathbf{\Pi}_r^T \Lambda_{\mathbf{z}_r^\star} \mathbf{\Pi}_r)^{-1}$ and of n additional constants of the form $\frac{\alpha\psi(z_r^\star(\mathbf{s},\mathbf{t}))}{2\tau_r}$, we obtain: $Z_{f_r}(\alpha, \tau_r) =$

$$\left(\frac{(2\pi)^n}{\det(\alpha \mathbf{\Pi}_r^T \Lambda_{\mathbf{z}_r^\star} \mathbf{\Pi}_r)}\right)^{\frac{1}{2}} \prod_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} \exp\{\frac{-\alpha\psi(z_r^\star(\mathbf{s},\mathbf{t}))}{2\tau_r}\}. \tag{42}$$

Since the determinant of a matrix product of square matrices equals the product of their determinants, one gets that:

$$\log\det(\alpha \mathbf{\Pi}_r^T \Lambda_{\mathbf{z}_r^\star} \mathbf{\Pi}_r) \propto n\log(\alpha) + 2\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} \log z_r^\star(\mathbf{s},\mathbf{t}), \tag{43}$$

resulting in the expression:

$$\begin{aligned} -\log Z_{f_r}(\alpha, \tau_r) \quad &\propto \quad -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\alpha) \qquad (44)\\ &+ \sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} \log z_r^\star(\mathbf{s},\mathbf{t}) + \frac{\alpha\psi(z_r^\star(\mathbf{s},\mathbf{t}))}{\tau_r}. \end{aligned}$$

Finally, the partition function $Z_L(I, \alpha, \beta, \mathbf{z}^\star, \tau_d, \tau_r)$ is an integral which possesses also an analogous analytical expression. Indeed, this partition function is[1]:

$$\begin{aligned} Z_L(I, \alpha, \beta, \mathbf{z}^\star, \tau_d, \tau_r) &= \int_{\mathbb{R}^n} \exp^{-\beta L(I, \mathbf{v}, \mathbf{z}^\star, \boldsymbol{\theta})} \mathrm{d}\mathbf{v} \qquad (45)\\ &= \int_{\mathbb{R}^n} \exp^{-\frac{\alpha}{2}\sum\limits_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} z_r^\star(\mathbf{s},\mathbf{t})V_r^2(\mathbf{v},\mathbf{s},\mathbf{t}) + \frac{\psi(z_r^\star(\mathbf{s},\mathbf{t}))}{\tau_r}} \\ &\exp^{-\frac{\beta}{2}\sum\limits_{\mathbf{s}} z_d^\star(\mathbf{s})V_d^2(I,\mathbf{v},\mathbf{s}) + \frac{\psi(z_d^\star(\mathbf{s})}{\tau_d}}\, \mathrm{d}\mathbf{v}. \end{aligned}$$

Potentials V_d and V_r are linear functions and this integral represents again the partition function of a *n*-dimensional Gaussian random field which is dependent on the Hessian matrix of the *posterior* energy. The Hessian matrix of the *posterior* is the sum of the *likelihood* and *prior* energy Hessian matrices. Let us calculate the Hessian of the *likelihood* energy:

$$\partial^2_{\mathbf{v}}(f_d(I, \mathbf{v}, \mathbf{z}_d^\star, \tau_d)) = \mathbf{\Pi}_d^T \Lambda_{\mathbf{z}_d^\star} \mathbf{\Pi}_d, \tag{46}$$

where $\Lambda_{\mathbf{z}_d^\star}$ is a diagonal matrix whose diagonal is the vector $((\mathbf{z}_d^\star)^T, (\mathbf{z}_d^\star)^T)^T$, and where $\mathbf{\Pi}_d$ is the matrix associated to function Π_d. For notational convenience, let us adopt the following notation for the posterior covariance inverse, *i.e.*, the Hessian of its energy:

$$\mathbf{\Gamma}^{-1}_{\mathbf{z}^\star,\gamma} = \mathbf{\Pi}_d^T \Lambda_{\mathbf{z}_d^\star} \mathbf{\Pi}_d + \gamma \mathbf{\Pi}_r^T \Lambda_{\mathbf{z}_r^\star} \mathbf{\Pi}_r.$$

Therefore, the *posterior* partition function integral reads:

$$Z_L(I, \mathbf{z}^\star, \boldsymbol{\theta}) = \exp^{-\beta L(I, \hat{\mathbf{v}}, \mathbf{z}^\star, \boldsymbol{\theta})}\, 2\pi^{\frac{n}{2}}\, \det(\beta \mathbf{\Gamma}^{-1}_{\mathbf{z}^\star,\gamma})^{-\frac{1}{2}} \tag{47}$$

Hence:

$$\begin{aligned} -\log Z_L(I, \mathbf{z}^\star, \boldsymbol{\theta}) = \beta L(I, \hat{\mathbf{v}}, \boldsymbol{\theta}) &- \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\beta) \\ &+ \frac{1}{2}\log\det \mathbf{\Gamma}^{-1}_{\mathbf{z}^\star,\gamma} \qquad (48) \end{aligned}$$

Replacing (39), (44), (48) in (36), one obtains an analytical expression for minus the *log evidence*:

$$\begin{aligned} -\log p(\mathbf{I}_t | \mathbf{z}^\star, \boldsymbol{\theta}, \mathcal{D}, \mathcal{R}) \propto\ & \frac{\beta}{2}\sum_{\mathbf{s}} z_d^\star(\mathbf{s}) V_d^2(I, \hat{\mathbf{v}}, \mathbf{s}) \qquad (49)\\ &+ \frac{\alpha}{2}\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} z_r^\star(\mathbf{s},\mathbf{t}) V_r^2(\hat{\mathbf{v}}, \mathbf{s}, \mathbf{t}) \\ &+ \frac{1}{2}\log\det \mathbf{\Gamma}^{-1}_{\mathbf{z}^\star,\gamma} - \sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} \log z_r^\star(\mathbf{s},\mathbf{t}). \\ &- m\log(\gamma) - \frac{1}{2}\sum_{\mathbf{s}} \log(\beta z_d^\star(\mathbf{s})) \end{aligned}$$

From (49), one can notice that the evidence is the multiplication of the *likelihood* by the so-called Occam factor. The *likelihood* evaluates the deviations from the observations while the Occam factor can be interpreted as a penalization of model complexity. In (49), the quadratic *likelihood* energy is $\frac{\beta}{2}\sum_{\mathbf{s}} z_d^\star(\mathbf{s}) V_d^2(I, \hat{\mathbf{v}}, \mathbf{s})$ and the log Occam factor is constituted by all the other terms appearing in the right hand side. Let us

remark that this equation reduces for quadratic cost functions to:

$$-\log p(\mathbf{I}_t|\boldsymbol{\theta},\mathcal{D},\mathcal{R}) \propto L(I,\hat{\mathbf{v}},\boldsymbol{\theta}) - m\log(\alpha) - \frac{m}{2}\log(\beta) + \frac{1}{2}\log\det\{\beta\boldsymbol{\Pi}_d^T\boldsymbol{\Pi}_d + \alpha\boldsymbol{\Pi}_r^T\boldsymbol{\Pi}_r\}. \quad (50)$$

This energy differs slightly by sign differences from the functional proposed in [26].

2) Maximization of hyper-parameter's evidence :
To finally achieve the estimation of the regularization and robust parameters, we now want to minimize the energy of the *evidence* PDF w.r.t to $\boldsymbol{\theta}$. Noting the following equality:

$$\frac{\partial}{\partial\eta}\log\det C = \text{trace}(C^{-1}\frac{\partial C}{\partial\eta}) \quad (51)$$

the partial derivatives of the *log evidence* $\log p(I|\boldsymbol{\theta},\mathbf{z}^\star,\mathcal{D},\mathcal{R})$ in (49) read:

$$\begin{aligned}
\frac{\partial\log p}{\partial\alpha} &= \frac{1}{2\beta}\left(\frac{n}{\gamma} - \text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_r^T\Lambda_{\mathbf{z}_r^\star}\boldsymbol{\Pi}_r)\right) \\
&- \frac{1}{2}\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} z_r^\star(\mathbf{s},\mathbf{t})V_r^2(\hat{\mathbf{v}},\mathbf{s},\mathbf{t}) \quad (52)\\
\frac{\partial\log p}{\partial\beta} &= \frac{m}{2\beta} - \frac{1}{2\beta}\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_d^T\Lambda_{\mathbf{z}_d^\star}\boldsymbol{\Pi}_d) \\
&- \frac{1}{2}\sum_{\mathbf{s}} z_d^\star(\mathbf{s})V_d^2(I,\hat{\mathbf{v}},\mathbf{s}) \quad (53)\\
\frac{\partial\log p}{\partial\tau_d} &= -\frac{1}{2}\sum_{\mathbf{s}}\frac{\partial z_d^\star(\mathbf{s})}{\partial\tau_d}\{\beta V_d^2(I,\hat{\mathbf{v}},\mathbf{s}) - \frac{1}{z_d^\star(\mathbf{s})}\} \\
&- \frac{1}{2}\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_d^T\frac{\partial\Lambda_{\mathbf{z}_d^\star}}{\partial\tau_d}\boldsymbol{\Pi}_d) \quad (54)\\
\frac{\partial\log p}{\partial\tau_r} &= \sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}}\frac{\partial z_r^\star(\mathbf{s},\mathbf{t})}{\partial\tau_r}\{\frac{1}{z_r^\star(\mathbf{s},\mathbf{t})} - \alpha V_r^2(\hat{\mathbf{v}},\mathbf{s},\mathbf{t})\} \\
&- \frac{\gamma}{2}\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_r^T\frac{\partial\Lambda_{\mathbf{z}_r^\star}}{\partial\tau_r}\boldsymbol{\Pi}_r), \quad (55)
\end{aligned}$$

A necessary condition for the *log evidence* to be maximum is that the partial derivatives cancel. The non-linear system composed of (52) and (53) can therefore be solved by iterating until convergence the following rules. In order to cancel the partial derivative, 1) the maximum *likelihood* estimate of the variance ratio $\hat{\gamma} = \frac{\hat{\alpha}}{\hat{\beta}}$ is updated at iteration k with:

$$\hat{\gamma}^{[k]} = \frac{n - \hat{\gamma}^{[k-1]}\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}^{[k-1]}}\boldsymbol{\Pi}_r^T\Lambda_{\mathbf{z}_r^\star}\boldsymbol{\Pi}_r)}{\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} z_r^\star(\mathbf{s},\mathbf{t})V_r^2(\hat{\mathbf{v}},\mathbf{s},\mathbf{t})\hat{\beta}^{[k-1]}}, \quad (56)$$

and 2) the maximum *likelihood* estimate of the observation model inverse variance with:

$$\hat{\beta}^{[k]} = \frac{m - \text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}^{[k-1]}}\boldsymbol{\Pi}_d^T\Lambda_{\mathbf{z}_d^\star}\boldsymbol{\Pi}_d)}{\sum_{\mathbf{s}} z_d^\star(\mathbf{s})V_d^2(I,\hat{\mathbf{v}},\mathbf{s})}. \quad (57)$$

For the estimation of the robust parameters, solution of the non-linear system composed of (54) and (55) can be efficiently obtained using a gradient descent method. Since an analytical expressions of the 2-nd order partial derivatives of the log *evidence* in (49) may be obtained with some approximations (see appendix B for the case of the Leclerc's function), we calculate the Newton directions [36] to accelerate the convergence of the algorithm. As these Newton directions are inexact, since they have been found only up to some approximations, we also check at each iteration of the gradient descent that a decrease of the *evidence* energy is observed. It results in the maximum *likelihood* estimates of hyper-parameters $\hat{\tau}_d$ and $\hat{\tau}_r$. Note that to perform inference in the second-level, the gradient descent procedure needs at each step the MAP estimate (motion and indices) of the first level.

Let us remark that the system used for estimating the hyper-parameters gathers non-linear equations. Since this system is solved using iterative rules or gradient descent methods (more precisely a Newton method), we can only guarantee convergence towards a local solution depending on the initial condition.

C. Selection of models $(\mathcal{D},\mathcal{R})$

The third level of inference (30) can be performed considering directly the dependence of the image variable I to the *likelihood* $\mathcal{D}$ and *prior* $\mathcal{R}$ models. Model inference is performed for the parabola indices $\mathbf{z}^\star$ estimated in the first MAP inference level given some M-estimators. Let us recall that the *likelihood* model $\mathcal{D}$ and the *prior* model $\mathcal{R}$ are chosen within a finite collection of hypotheses. More precisely $\mathcal{D}$ is a binary variable defined on the finite set $\{(\Pi_d,\rho_{\tau_d})\}$, which is a Cartesian product of a set of observation function Π_d (*e.g.* brightness conservation, passive scalar advection-diffusion, etc.) by a set of M-estimators ρ_{τ_d} (*e.g.* Gaussian , Laplacian or Leclerc's penalty function). $\mathcal{R}$ is a binary variable defined on the finite set $\{(\Pi_r,\rho_{\tau_r})\}$, which is the Cartesian product of a set of regularization function Π_r (e.g. weak spatial gradients or weak divergence/curl, etc.) by a set of M-estimators ρ_{τ_r}. Bayes relation for this dependence reads:

$$\begin{aligned}
p(\mathcal{D},\mathcal{R}|\mathbf{I}_t,\mathbf{z}^\star) &= \frac{p(\mathbf{I}_t|\mathbf{z}^\star,\mathcal{D},\mathcal{R})p(\mathcal{D},\mathcal{R}|\mathbf{z}^\star)}{p(\mathbf{I}_t|\mathbf{z}^\star)} \\
&\propto p(\mathbf{I}_t|\mathbf{z}^\star,\mathcal{D},\mathcal{R})p(\mathcal{D},\mathcal{R}|\mathbf{z}^\star), \quad (58)
\end{aligned}$$

The evaluation of the data-dependent term, the *evidence* $p(\mathbf{I}_t,\mathbf{z}^\star|\mathcal{D},\mathcal{R})$ for $(\mathcal{D},\mathcal{R})$, implies marginalizing out variable $\boldsymbol{\theta}$. Integrating the *evidence* w.r.t these variables yields:

$$p(\mathbf{I}_t|\mathbf{z}^\star,\mathcal{D},\mathcal{R}) = \int_{\mathbb{R}^4} p(\mathbf{I}_t|\mathbf{z}^\star,\boldsymbol{\theta},\mathcal{D},\mathcal{R})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \quad (59)$$

As $\boldsymbol{\theta}$ varies, an *evidence* maximum (at least a local) is obtained at $\hat{\boldsymbol{\theta}}$. The *evidence* maximum is usually well approximated by an uncorrelated Gaussian, where the diagonal elements of the covariance matrix are σ_α^2, σ_β^2, $\sigma_{\tau_d}^2$ and $\sigma_{\tau_r}^2$. Therefore,

using the Laplace's approximation [2] of integral (59), the model *evidence* reads:

$$p(\mathbf{I}_t|\mathbf{z}^\star,\mathcal{D},\mathcal{R}) \simeq p(\mathbf{I}_t|\mathbf{z}^\star,\hat{\boldsymbol{\theta}},\mathcal{D},\mathcal{R})p(\hat{\boldsymbol{\theta}})(2\pi)^2\sigma_\alpha\sigma_\beta\sigma_{\tau_d}\sigma_{\tau_r},\sigma_{\tau_r}, \tag{61}$$

where the variance inverses are defined as second order derivatives of the *evidence* energy w.r.t. α, β, τ_d and τ_r. Differentiating twice (49) and employing estimates defined in (56) and (57), we obtain an approximation for variances:

$$\begin{aligned}\sigma_\beta^2 &\simeq \frac{2\hat{\beta}^2}{m-\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}\boldsymbol{\Pi}_d^T\Lambda_{\mathbf{z}_d^\star}\boldsymbol{\Pi}_d)}\\ \sigma_\alpha^2 &\simeq \frac{2\hat{\alpha}^2}{n}\end{aligned} \tag{62}$$

Approximated variances $\sigma_{\tau_d}^2$ and $\sigma_{\tau_r}^2$ can be obtained similarly by using the Newton directions at point $\hat{\tau_d}$ and $\hat{\tau_r}$, that is to say the 2-nd order partial derivatives of (49). Finally, considering a flat *prior* $p(\alpha,\beta)$ we obtain the energy of the *evidence* for models $(\mathcal{D},\mathcal{R})$: $-\log p(\mathbf{I}_t|\mathbf{z}^\star,\mathcal{D},\mathcal{R})$

$$\begin{aligned}\propto\ & \beta\sum_{\mathbf{s}} z_d^\star(\mathbf{s})V_d^2(I,\hat{\mathbf{v}},\mathbf{s}) + \alpha\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}} z_r^\star(\mathbf{s},\mathbf{t})V_r^2(\hat{\mathbf{v}},\mathbf{s},\mathbf{t})\\ -\ & n\log(\hat{\gamma}) + \log\det\{\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}^{-1}\}\\ -\ & 2\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}}\log(z_r^\star(\mathbf{s},\mathbf{t},\hat{\tau_r})) - \sum_{\mathbf{s}}\log(\beta z_d^\star(\mathbf{s},\hat{\tau_d}))\\ -\ & \log\sigma_\alpha^2 - \log\sigma_\beta^2 - \log\sigma_{\tau_d}^2 - \log\sigma_{\tau_r}^2.\end{aligned} \tag{63}$$

The maximum *likelihood* estimate $(\hat{\mathcal{R}},\hat{\mathcal{D}})$ is the minimizer of (63) within a collection of model candidates $\{(\mathcal{R},\mathcal{D})\}$ and represents the selected *prior* and data models.

D. Complexity

In terms of complexity, one of the two bottlenecks of the proposed Bayesian inference technique corresponds to the computation of the high-dimensional posterior covariance trace in (56), (57), (62), (77) and (78). In order to overcome this problem (without neglecting the spatial interaction of motion variables as done in [26]), we use a trace randomization technique [16] [21], which is efficiently achieved using the CGS algorithm. More precisely, the drawing of L random samples of a n-dimensional normalized and centered Gaussian distribution $r_j \sim \mathcal{N}(0,\mathbf{1}^n)$ leads to the following approximation. For any square n-dimensional matrix $\boldsymbol{B}$, we have:

$$\begin{aligned}\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}\boldsymbol{B}) &= \text{trace}\mathbb{E}[\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}r_j r_j^t\boldsymbol{B}] = \mathbb{E}[(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}r_j)^t(r_j\boldsymbol{B})]\\ &\simeq \frac{1}{L}\sum_{j=1}^{L}[\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}r_j]^t\boldsymbol{B}r_j\end{aligned} \tag{64}$$

[2]Laplace's method approximates the integral of a function by fitting a Gaussian at its maximum and computing the volume under the Gaussian. For a k dimensional variable $\mathbf{x}$ and a function $f(\mathbf{x})$, the Laplaces' approximation reads:

$$\int f(\mathbf{x})d\mathbf{x} \simeq f(\hat{\mathbf{x}})(2\pi)^{k/2}[-\det\{\partial_{\mathbf{x}}^2\log f(\hat{\mathbf{x}})\}]^{-1/2}, \tag{60}$$

where $\partial_{\mathbf{x}}^2\log f(\hat{\mathbf{x}})$ represents the Hessian of the logarithm of the function and where $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} f(\mathbf{x})$.

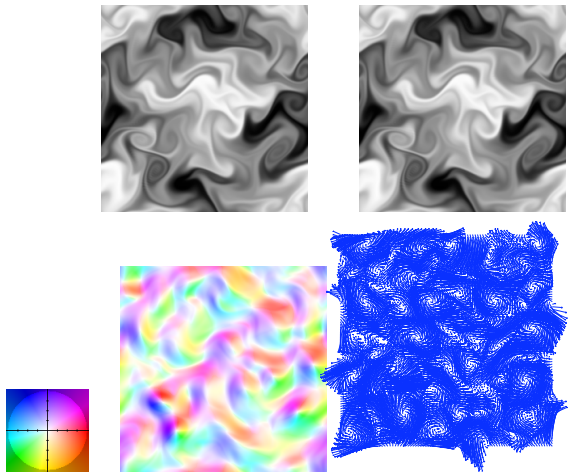


Fig. 1. **Simulated scalar advection-diffusion by two-dimensional turbulence.** Generated scalar images $I(t)$ and $I(t+1)$, and ground truth motion field visualization in a vectorial and in a color representation [1]. In the latter visualizations, color and intensity code vector orientations and magnitudes.

where vector $\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}r_j$ is the solution provided by the CGS algorithm of the inverse problem $\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}^{-1}\mathcal{X} = r_j$, and $\mathcal{X}$ is the unknown. This empirical average converges towards the trace when L tends to infinity. Note that we can reuse the same vector $\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}r_j$ to approach the different traces in (56), (57), (62), (77) and (78). The complexity of these approximations is $\mathcal{O}(n)$.

The second bottleneck is the computation of the determinant of the Hessian of the posterior energy in (63). The determinant of this large and sparse positive-definite matrix is efficiently calculated via a Lower Upper (LU) decomposition[3], without any independence assumption in the structure of the Hessian matrix. Indeed, for the matrix decomposition $\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}^{-1} = LU$, the determinant is equal to the product of the diagonal elements u_{ii} of the upper triangular matrix U:

$$\det(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\hat{\gamma}}^{-1}) = \prod_i u_{ii}. \tag{65}$$

The determinant is obtained with a complexity of $\mathcal{O}(n)$, yielding a linear overall algorithm complexity.

Other Bayesian methods [26] or deterministic methods [35] [16] require in principle the same amount of computation since they rely on some analogous trace computation. Obviously, the smaller the dimension n, the lower the complexity, making local inference approaches attractive [12], such as in [40]. Finally, L-Curve or U-Curve strategies [27] possess a negligible complexity in comparison.

IV. Experiments

In the following experiments, hyper-parameters and motion have been estimated at the different level of a standard multi-resolution representation [3] whereas inference on *prior* and *likelihood* models has been performed only on the finest resolution level.

[3]The LU decomposition is approached by an algorithm known as ILU(0) of complexity $\mathcal{O}(n)$.

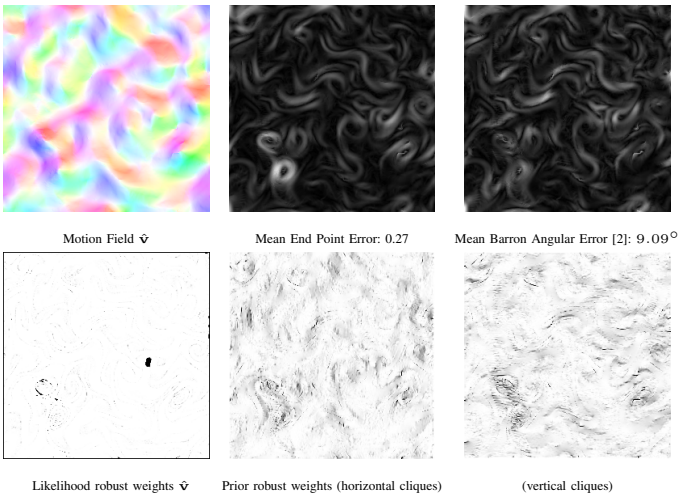


Fig. 2. **Evaluation of most likely motion $\hat{\mathbf{v}}$ estimate.** Color visualization [1] and errors corresponding to an estimation with the robust data model f_d of (66) and a robust 1-st order regularizer f_r defined in (7). Hyper-parameter vector $\hat{\boldsymbol{\theta}} = (\hat{\gamma}, \hat{\tau}_d, \hat{\tau}_r) = (0.59, 2.12, 20.65)$ and the data model diffusion coefficient $\hat{\nu} = 0.4$ were selected by Bayesian inference. The bottom line displays spatial maps of robust weights $\hat{z}_d$ and $\hat{z}_r$ corresponding to the Leclerc's norm parameters $\hat{\tau}_d$ and $\hat{\tau}_r$.

A. *Fluid motion image sequence*

To evaluate the performance of the method, a synthetic scalar image sequence of 256×256 pixels was generated by transporting a passive and diffusive scalar with two-dimensional turbulence [6]. The dynamical process was obtained by direct numerical simulation (DNS) of Navier-Stokes equations coupled with the advection-diffusion of a passive scalar equation:

$$\partial_t I + \mathbf{v} \cdot \nabla I = \nu \Delta I, \tag{66}$$

where ν represents a unknown diffusion coefficient and Δ denotes the Laplacian operator. Fig. 1 presents two scalar images of the sequence together with the vectorial and the color representation of the true underlying velocity field [1]. Fig. 2 displays the *posterior* motion estimate $\hat{\mathbf{v}}$ obtained applying the proposed Bayesian inference framework on this difficult image sequence. In the following we show that the use of Bayesian inference with a simple robust 1-st order regularization outperforms the most accurate state-of-the-art results [6], [8], [45].

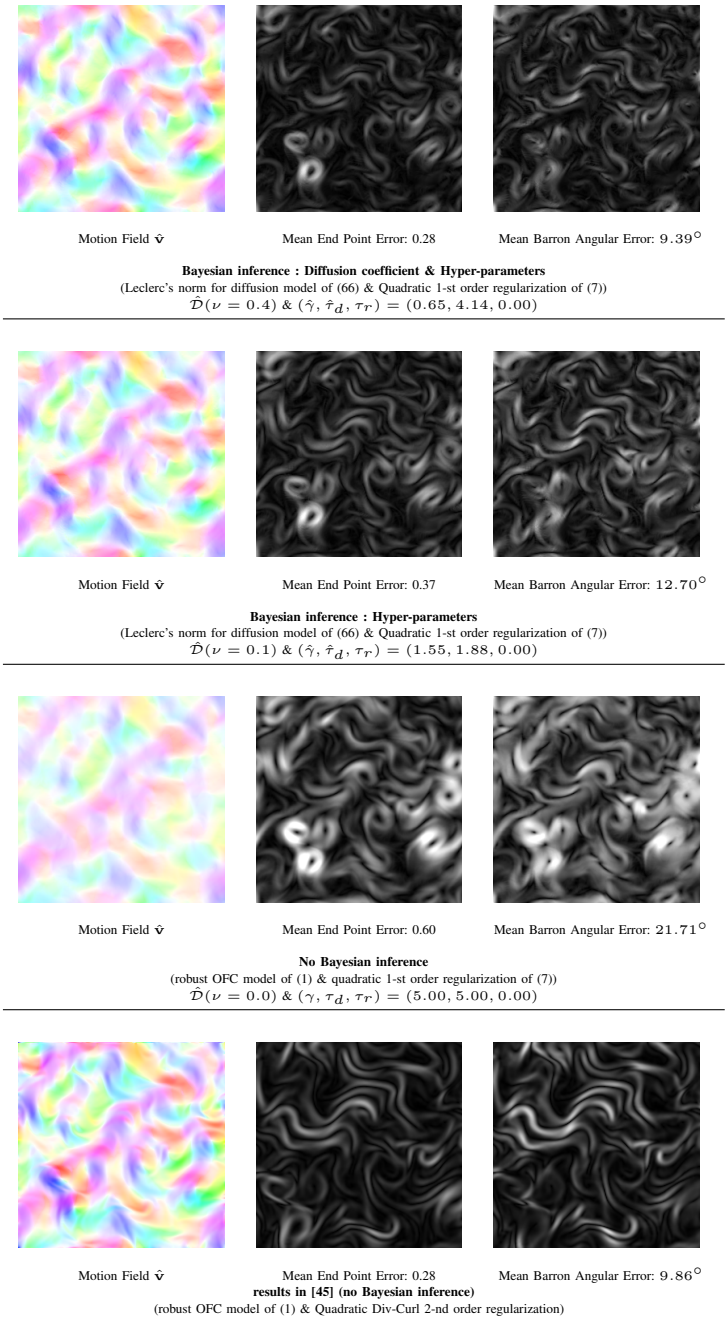


Fig. 3. **Evaluation of motion estimates.** Importance of hyper-parameters estimation and model selection & comparison to state-of-the-art results

Model Selection. Inference of $(\hat{\mathcal{R}}, \hat{\mathcal{D}})$ was performed on a collection of model combinations $\{(\mathcal{D}, \mathcal{R})\}$. The set of *likelihood* models $\{\mathcal{D}\}$ was defined by combining a finite set of scalar advection-diffusion observation model Π_d defined for different values of diffusion coefficient ν in (66), with two different penaltization ($\mathcal{L}^2$ norm or the Leclerc's function). The set of *prior* models $\{\mathcal{R}\}$ was defined by combining a 1-st order regularizer Π_r defined in (7) with the same two previous penalty functions. Using the *evidence* framework, the diffusion coefficient $\hat{\nu} = 0.4$ and the Leclerc's functions were selected accordingly to the inferred hyper-parameter vector $\hat{\boldsymbol{\theta}}$. The choice of the diffusion coefficient appears to be crucial when comparing the results displayed in the two first rows of Fig. 3. Indeed, performing model selection (first row) instead of choosing an arbitrary diffusion coefficient (second row) induces a relevant error decrease. The left plot of Fig. 4 shows that the diffusion coefficient minimizing the model *evidence* energy $-\log p(I|\hat{\mathbf{z}}, \mathcal{D}, \mathcal{R})$ is very close to the diffusion coefficient minimizing the average end point error. It is interesting to note in the right plot of the same figure that the selected value $\hat{\nu}$ also corresponds respectively to a minimum and a maximum of the inferred regularization coefficient $\hat{\gamma}$ and Leclerc's parameter $\hat{\tau}_d$. Leclerc's robust

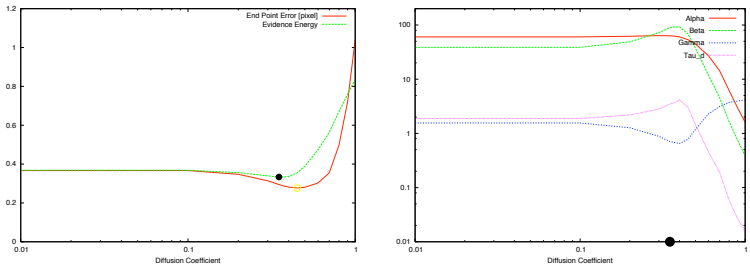


Fig. 4. **Evaluation of the scalar diffusion model selection.** Left: the energy curve of model *evidence* w.r.t. the diffusion coefficient ν (continuous red line) shares the same minimum (full circle) in $\hat{\nu}$ up to a slight bias as the end-point error curve (green dashes) between ground truth and the motion estimate. Right: Evolution of regularization coefficient estimates $\hat{\alpha}$ (red line), $\hat{\beta}$ (green dashes) and $\hat{\gamma}$ (blue points) and robust data model parameter estimate $\hat{\tau}_d$ (pink dashes) w.r.t. the diffusion coefficient. The circle indicates the diffusion coefficient selected value.

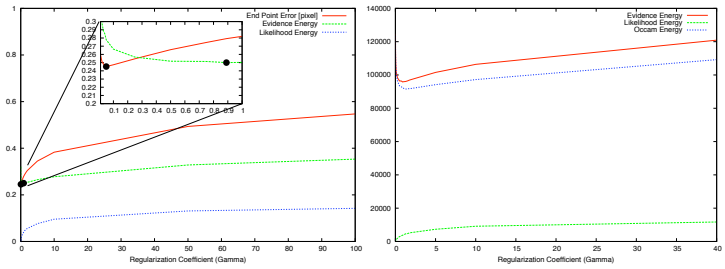


Fig. 5. **Evaluation of regularization coefficient estimation. Left**: energy curve of the regularization coefficient *evidence* (green dashes) and *likelihood* (blue points). In the zoom, the estimate $\hat{\gamma}$ yielding the minimum (bolded dot) of the *evidence* energy appears to be very close to the regularization coefficient yielding the minimum (other bolded dot) of the end point error (continuous red line) between ground truth and the estimated motion. **Right**: evidence energy (red line) composed of the addition of log of the Occam factor (blue points) and of likelihood energy (green dashes) w.r.t regularization coefficient. The evidence energy is largely influenced by the Occam factor while the likelihood energy appears to have a minor influence and obviously does not constitute a good criterion for regularization coefficient estimation.

function was also selected based on model *evidence* for the 1-st order regularizer. But, fluid motion fields which we are dealing with are spatially continuous. Nevertheless, the choice of a Leclerc's cost on the 1-st order regularizer is not really incoherent since a quadratic first order *prior* may not be able to model appropriately the flow regularity. Indeed, comparing errors of the reconstructed motion field with errors displayed in the first row of Fig. 3 shows that using robust Leclerc's M-estimators yielded slightly better results than using quadratic norms for *priors* . Note that considering 2-nd order regularizers proposed in [8] or [45] as candidates in the model selection stage could have resulted in a significant enhancement of results. However, for sake of simplicity, we limited the *prior* model selection stage to the choice of the M-estimator for penalization.

Hyper-parameter Estimation. The vector estimate $\hat{\boldsymbol{\theta}}$ is constituted by regularization coefficients $\hat{\alpha}$ and $\hat{\beta}$ and by parameters $\hat{\tau}_r$ and $\hat{\tau}_d$. The ratio of the two former parameters constitute the estimated regularization coefficient $\hat{\gamma} = \hat{\alpha}/\hat{\beta}$. Fig. 5 shows that the estimated regularization coefficient appears to be very closed fom the value minimizing the average end point error between the true and the estimated motion. This figure also highlights the fact that minimizing

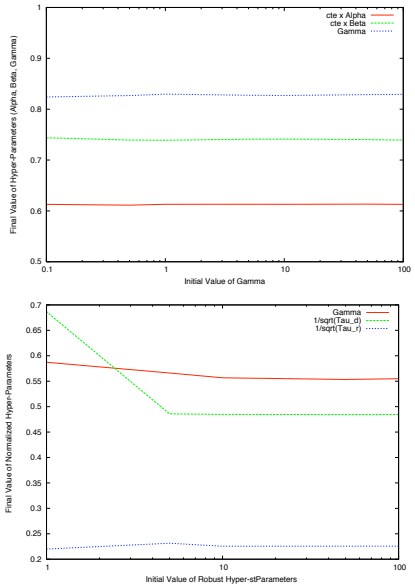


Fig. 6. **Evaluation of sensitivity to initial conditions.** Above: final versus initial values of regularization parameter estimates $\hat{\alpha}$ (continuous red line), $\hat{\beta}$ (green dashes) and $\hat{\gamma}$ (blue points). Below: final regularization coefficient estimate $\hat{\gamma}$ (red line) and robust data parameter $(\hat{\tau}_d)^{-1/2}$ (green dashes) and robust regularization parameter $(\hat{\tau}_r)^{-1/2}$ (blue points) parameters w.r.t initial values of $\tau_d = \tau_r$.

the *likelihood* energy is not a reliable criterion for selecting the regularization parameter. The parameters $\hat{\tau}_d$ and $\hat{\tau}_r$ are related to the two different Leclerc's M-estimators, associated respectively to a robust penalization of the scalar diffusion model and to a robust 1-st order regularization *prior* energy. Let us remark that in previous work, it has been shown that a 1-st order regularizer is generally unsuited for fluid flows whereas 2-nd order divergence and curl regularizers models more accurately the flow [7]. This behavior can also be observed in our experiments by comparing the last two rows of Fig. 3: results of the two regularizers employed with an arbitrary set of hyper-parmeters and the OFC data model illustrate that the error can be reduced by half using fluid-dedicated regularization model. Nevertheless, it is important to notice that such results are entirely dependent on hyper-parmeter tuning. Indeed, comparing the first and last row of Fig. 3 shows that fitting an inappropriate regularizers (in other words estimating hyper-parmeters) while selecting data models by Bayesian inference yields better results than fine regularizers adjusted by manually tuning hyper-parameters.

Sensitivity to initial conditions. Some experiments have then been conducted to evaluate the sensitivity of the hyper-parameter estimation method with respect to the initial condition. In this perspective, several initial conditions have been considered for the estimation of the regularization coefficients and the robust parameters. As shown in the left plots of Fig. 6, the initial values of the regularization coefficient ranging from 0.1 to 100 led to an identical value for the final estimate $\hat{\gamma}$. Based on this heuristic result, one can argue that in this case, the energy of the hyper-parameter *evidence* defined in (49) behaves as a convex functional.

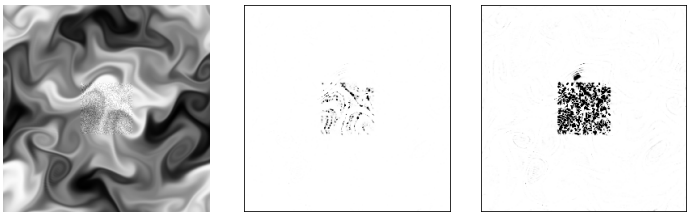
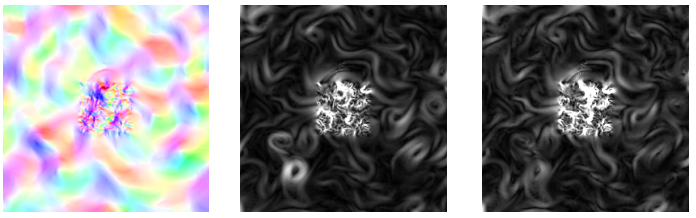


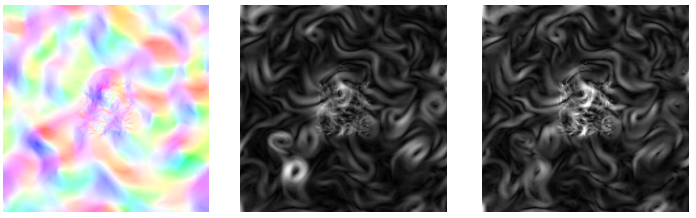
Image with noise: $\mathcal{N}(0, 20^2)$ Robust weights : Fixed $(\tau_d)^{-1/2} = 5$. Estimated $(\hat{\tau}_d)^{-1/2} = 2.21$

Fig. 7. **Evaluation of robustness to noisy observations.** Left: Visualization of additional independent and identically distributed Gaussian noise $\mathcal{N}(0, 20^2)$ at the center of original image. Middle: Map of robust weights associated to the data term for a parameter value of $(\tau_d)^{-1/2} = 5.00$ chosen arbitrarily. Right: Map of robust weights associated to the data term for the estimated parameter value of $(\hat{\tau}_d)^{-1/2} = 2.21$.



Motion Field | Mean End Point Error: 0.35 | Mean Barron Angular Error: 11.83°

Fixed hyper-parameters: $\tau = 0.04, \gamma = 1$.



Motion Field | Mean End Point Error: 0.34 | Mean Barron Angular Error: 11.58°

Estimated hyper-parameters: $\hat{\tau} = 0.20, \hat{\gamma} = 2.79$

Fig. 8. **Motion field reconstruction from incomplete observations.** Motion estimates, end point and Barron angular error for an arbitrary robust parameter value $(\tau_d)^{-1/2} = 5.00$ (row above) and for the most likely estimate $(\hat{\tau}_d)^{-1/2} = 2.21$ (row below).

Concerning the sensitivity of the robust Leclerc's parameters associated to the *prior* and the *likelihood*, we considered initial couples (τ_d, τ_r) with identical values $\tau_d = \tau_r$ ranging from 0.1 to 1. Results displayed in the right plots of Fig. 6 show convergence towards identical hyper-parameter estimates, even if some instabilities can be noticed on the final estimate $\hat{\tau}_d$ for too large initial values. This seems a quite natural behavior since non-linearities are likely to increase when τ_d tends to infinity.

Sensitivity to noise. Uncorrelated Gaussian noise of zero mean and of standard deviation equal to about 10% of the image intensity dynamic was added to one of the two images of the sequence, in a localized spatial area at the image grid center. This noise inclusion allows us to evaluate the method ability to discard unreliable observations. This noise induces obviously also a change in the statistics of the *likelihood* and the *prior* models *i.e.* in the value of $\hat{\gamma}$. As shown in Fig. 7, the Bayesian inference scheme was able to estimate the most likely parameter $\hat{\tau}_d$ and to produce a relevant outlier map. To illustrate the non-efficiency of a manual tuning, a comparison with a fixed robust parameter is provided in this figure. The reconstructed motion fields from the incomplete observations are shown in Fig. 8. A comparison of the corresponding error maps is also provided in this figure.

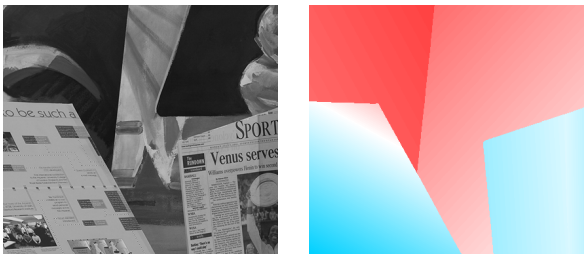


Fig. 9. **"Venus" image sequence.** Frame 10 of the "Venus" sequence and motion ground truth in color visualization.

Prior f_r	$\hat{\gamma}$	$\hat{\tau}_d$	$\hat{\tau}_r$	$-\log p(I\|\mathcal{D},\mathcal{R})$	Mean End Point Error	Mean Barron Error
$\mathcal{L}^2$	11.65	0.01	0	567808	0.485	8.423
Leclerc	11.36	0.01	0.34	565274	0.485	8.419
$\mathcal{L}^1$	10.50	0.01	1.25	**556641**	**0.482**	**8.348**

Fig. 10. **Prior model $\hat{f}_r = \mathcal{L}^1$ selection and hyper-parameters $(\hat{\gamma}, \hat{\tau}_d, \hat{\tau}_r)$ estimation**. Table showing hyper-parameter estimates together with the score obtained for different regularization norms in terms of *evidence* energy, end point and Barron angular error.

B. Computer vision scenes

The Bayesian inference method has then been assessed using scalar image sequences of the Middleburry database. In this database, challenging computer vision scenes are provided for optic-flow estimation evaluation [1].

"Venus" sequence. We first concentrate on the "Venus" sequence. Frame 10 and 11 were chosen for evaluation. The first frame and the associated ground truth are displayed in Fig. 9. A collection of models $\mathcal{R}$ was constituted by combining a *prior* 1-st order regularizer Π_r with three different penalization ($\mathcal{L}^2$, $\mathcal{L}^1$ or Lecler's cost function[4]). The *likelihood* model $\mathcal{D}$ was chosen to be the OFC equation penalized by Leclerc's function. The table displayed in Fig. 10 shows that the the energy of the *prior* model *evidence* reaches its minimum for the $\mathcal{L}^1$ norm and that this selected model $\hat{\mathcal{R}}$ also induces the weakest error between the true and the estimated motion. The estimated motion field $\hat{\mathbf{v}}$ and error maps obtained with selected model $\hat{\mathcal{R}}$ and hyper-parameters $\hat{\boldsymbol{\theta}}$ are shown in Fig. 11. For visual evaluation, the most likely maps of data outliers and motion spatial discontinuities related respectively to estimate $\hat{\tau}_d$ and $\hat{\tau}_r$ are displayed in the same figure. As shown in the table of Fig. 15, errors obtained with Bayesian inference on the simplest *likelihood* and *prior* models are comparable to error of manually tuned hyper-parameters of affine regularization model or specialized data term dedicated to such scenes composed of rigid objects [4], [5].

[4]The robust penalty function defined in (71) is here abusively called $\mathcal{L}^1$.

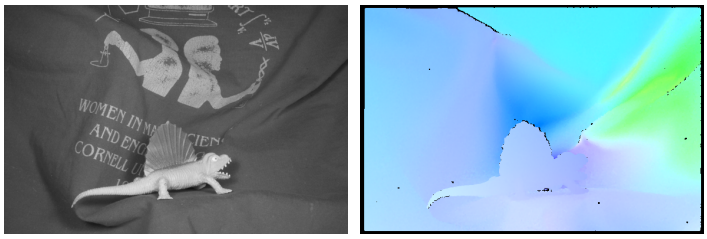


Fig. 12. **"Dimetrodon" image sequence.** Frame 10 of the "Dimetrodon" sequence and motion ground truth in color visualization.

	"Venus"	"Dimetrodon"
Bayesian inference	**8.348**	**3.303**
Bruhn&al [5].	8.732	10.993
Black&Anandan [4].	7.641	9.261
Pyramid Lucas-Kanade [30]	14.614	10.272
Zitnick & al [46].	11.423	30.105

Fig. 15. **Comparison with state-of-the-art.** Mean Barron angular error comparison with other methods available in Middleburry's database

"Dimetrodon" sequence. We then assessed the inference method on the "Dimetrodon" sequence. Frame 10 and 11 were chosen for evaluation. The first frame and the associated ground truth are displayed in Fig. 12. The collection of candidate models $(\mathcal{D}, \mathcal{R})$ was constituted by associating a *prior* 1-st order regularizers Π_r with an OFC based observation operator Π_d, considering all possible combination of penalization for the *likelihood* and the *prior* models among $\mathcal{L}^2$, $\mathcal{L}^1$ and the Lecler's M-estimator. Table in Fig. 13 shows that a combination of two $\mathcal{L}^1$ norm performs the best in terms of angular and end point error. Note that this association $(\hat{\mathcal{D}}, \hat{\mathcal{R}})$ corresponds also to the models which have been selected by maximizing the *evidence*. The estimated motion field $\hat{\mathbf{v}}$ and error maps obtained with the selected models $(\hat{\mathcal{D}}, \hat{\mathcal{R}})$ and related hyper-parameters $\hat{\boldsymbol{\theta}}$ are displayed in Fig. 14. Table in Fig. 15 shows that a Bayesian inference scheme enables to outperform significantly more refined method with manually tuned parameters. For an additional visual evaluation, the most likely maps of data outliers and motion spatial discontinuities related respectively to estimate $\hat{\tau_d}$ and $\hat{\tau_r}$ are displayed in Fig. 14. Not surprisingly, comparing the outlier maps of the "Venus" and the "Dimetrodon" scenes reveals that, conversely to the Leclerc's cost which discard only strong occluded areas, a $\mathcal{L}^1$ norm favors zero values for robust weights in region with weak gradients, occlusions, ambiguities or areas carrying little information. This is somehow in agreement with the concept of sparsity encouraged by the $\mathcal{L}^1$ norm.

V. Conclusion

Inspired by the Bayesian *evidence* framework proposed in [31] to solve noisy interpolation problems, we have derived a generic and reliable hierarchical model for robust optic-flow estimation where motion fields, models together with their hyper-parameters are treated as random variables. Marginalizing out intermediate variables of the hierarchy enables us to express probabilistically direct conditional dependences between the variable we want to estimate and the image data. Optic-flow, regularization coefficients, M-estimator parameters, *prior* and *likelihood* motion models are in this way simultaneously inferred by maximizing the *posterior* or the *likelihood* of the marginalized distributions. The method has been assessed on several image sequences. In particular, the proposed Bayesian inference scheme succeeds to select the appropriate underlying physical data model driving the diffusion of a scalar by a turbulent flow. Moreover, the Bayesian choice of robust penalty functions for the data and the regularization models appeared to maximize optic-flow estimation accuracy. Furthermore, the hyper-parameter estimation process proved to be robust to noise and not sensitive to the initial conditions. Comparison with the state-of-the-art estimators in fluid flow estimation and computer vision assessed the efficiency of the method and highlighted the importance of such inference schemes.

Appendix

A - M-estimators and semi-quadratic models

We recall in this appendix the properties of M-estimators. Let ρ_τ be a real-valued continuously differentiable even function such that:

1) ρ_τ is increasing on $\mathbb{R}^+$
2) $\phi(V^2) \triangleq \rho_\tau(V)$ is strictly convex on $\mathbb{R}^+$
3) $\lim_{V^2 \to +\infty} \phi'(V^2) = 0$
4) $\tau \triangleq \lim_{V^2 \to 0^+} \phi'(V^2) < +\infty$.

There exists a function ψ continuously differentiable on $]0,1]$, such that:

$$\forall V \in \mathbb{R}^+,\ \rho_\tau(V) = \min_{z \in]0,1]} \varphi_\tau(V, z) \tag{67}$$

where parabola φ_τ is by definition a semi-quadratic parametric function:

$$\varphi_\tau(V, z) \triangleq \tau z V^2 + \tau \psi(z), \tag{68}$$

with function ψ defined as:

$$\psi(z)\tau \triangleq \phi \circ (\phi')^{-1}(\tau z) - \tau z (\phi')^{-1}(\tau z). \tag{69}$$

This means that the graph of ρ is the inferior envelope of a family of parabolas φ_τ continuously indexed by $z \in]0,1]$. The minimum in (67) is given in closed form by:

$$\arg\min_{z \in]0,1]} \varphi_\tau(V, z) = \frac{\rho'_\tau(V)}{2\tau V}. \tag{70}$$

M-estimators correspond to some well-known functions. For instance, in the limit of $\tau_d \to \infty$, using in (12) the following function:

$$\rho_{\tau_d}(V_d(I, \mathbf{v}, \mathbf{s})) \quad = \quad \sqrt{\frac{1}{4\tau_d^2} + V_d^2(I, \mathbf{v}, \mathbf{s})} \tag{71}$$

or equivalently in (14) the associated normalized indices:

$$\hat{z}_d(V_d(I, \mathbf{v}, \mathbf{s}), \tau_d) = \{1 + (2\tau_d V_d(I, \mathbf{v}, \mathbf{s}))^2\}^{-1/2}, \tag{72}$$

we obtain a $\mathcal{L}^1$ norm or equivalently the energy of a Laplace's distribution. Another robust norm is the Leclerc's function. It reads:

$$\rho_{\tau_d}(V_d(I, \mathbf{v}, \mathbf{s})) = 1 - \exp\{-\tau_d V_d^2(I, \mathbf{v}, \mathbf{s}))\} \tag{73}$$

and is associated to the normalized exponential weights:

$$\hat{z}_d(V_d(I, \mathbf{v}, \mathbf{s}), \tau_d) = \exp\{-\tau_d V_d^2(I, \mathbf{v}, \mathbf{s})\}. \tag{74}$$

B - Newton's method for estimation of M-estimator hyper-parameters

Let us denote by $f_{\partial\tau_d}$ and $f_{\partial\tau_r}$ minus two times the partial derivatives given in (54) and (55) of the log *evidence* $\log p(I|\boldsymbol{\theta},\mathcal{D},\mathcal{R})$ w.r.t. respectively to the hyper-parameters τ_d and τ_r :

$$\begin{aligned} f_{\partial\tau_d}(\tau_d) &= \text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_d^T\frac{\partial\Lambda_{\mathbf{z}_d^\star}}{\partial\tau_d}\boldsymbol{\Pi}_d) \\ &+ \sum_{\mathbf{s}}\frac{\partial z_d^\star(\mathbf{s})}{\partial\tau_d}\{\beta V_d^2(I,\hat{\mathbf{v}},\mathbf{s})-\frac{1}{z_d^\star(\mathbf{s})}\} \\ f_{\partial\tau_r}(\tau_r) &= \gamma\text{trace}(\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_r^T\frac{\partial\Lambda_{\mathbf{z}_r^\star}}{\partial\tau_r}\boldsymbol{\Pi}_r) \\ &+ 2\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}}\frac{\partial z_r^\star(\mathbf{s},\mathbf{t})}{\partial\tau_r}\{\alpha V_r^2(\hat{\mathbf{v}},\mathbf{s},\mathbf{t})-\frac{1}{z_r^\star(\mathbf{s},\mathbf{t})}\} \end{aligned} \quad (75)$$

The value $\hat{\tau}_d$ and $\hat{\tau}_r$ where these partial derivatives vanish correspond to the maximum of the *evidence* probability. To access this minimum, we use a Newton's method [36] independently on each component using a fixed step λ in the gradient algorithm:

$$\left\{\begin{array}{l} \tau_d^{[k+1]} = \tau_d^{[k]} - \lambda\frac{f_{\partial\tau_d}}{f_{\partial^2\tau_d}}(\tau_d^{[k]}) \\ \tau_r^{[k+1]} = \tau_r^{[k]} - \lambda\frac{f_{\partial\tau_r}}{f_{\partial^2\tau_r}}(\tau_r^{[k]}) \end{array}\right. , \quad (76)$$

with $0<\lambda\leq 1$ and where the limit case of $\lambda=1$ corresponds to the Newton-Raphson algorithm. In the previous equations $f_{\partial^2\tau_d}$ and $f_{\partial^2\tau_r}$ represent minus two times the second order partial derivatives of the log *evidence* with respect to τ_d and τ_r.

For instance, in the case of Leclerc's M-estimator, noting that $\frac{\partial z_d^\star(\mathbf{s})}{\partial\tau_d} = -V_d^2(\mathbf{s})z_d^\star(\mathbf{s})$ and $\frac{\partial z_r^\star(\mathbf{s},\mathbf{t})}{\partial\tau_r} = -V_r^2(\mathbf{s},\mathbf{t})z_r^\star(\mathbf{s},\mathbf{t})$, the first order partial derivatives of (75) can be written as:

$$\begin{aligned} f_{\partial\tau_d}(\tau_d) &= -\text{trace}[\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_d^T\Lambda_{\mathbf{z}_d^\star}\Lambda_{V_d^2}\boldsymbol{\Pi}_d] \\ &+ \sum_{\mathbf{s}}\{V_d^2(I,\hat{\mathbf{v}},\mathbf{s})-\beta V_d^4(I,\hat{\mathbf{v}},\mathbf{s})z_d^\star(\mathbf{s})\} \\ f_{\partial\tau_r}(\tau_r) &= -\gamma\text{trace}[\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_r^T\Lambda_{\mathbf{z}_r^\star}\Lambda_{V_r^2}\boldsymbol{\Pi}_r] \\ &+ 2\sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}}\{V_r^2(\hat{\mathbf{v}},\mathbf{s},\mathbf{t})-\alpha z_r^\star(\mathbf{s},\mathbf{t})V_r^4(\hat{\mathbf{v}},\mathbf{s},\mathbf{t})\}, \end{aligned} \quad (77)$$

where $\Lambda_{V_d^p}$ and $\Lambda_{V_r^p}$ are diagonal matrices which are composed of the vector elements $(V_d(I,\hat{\mathbf{v}},\mathbf{s}))^p$ and $(V_r(\hat{\mathbf{v}},\mathbf{s}))^p$.

Neglecting the inverse hessian derivative in comparison to other terms in the trace operator, one obtains by differentiating two times the log *evidence* an approximated analytical expression of the second order partial derivatives:

$$\begin{aligned} f_{\partial^2\tau_d}(\tau_d) &\simeq \text{trace}[\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_d^T\Lambda_{\mathbf{z}_d^\star}\Lambda_{V_d^4}\boldsymbol{\Pi}_d] \\ &+ \sum_{\mathbf{s}}\beta z_d^\star(\mathbf{s})V_d^6(I,\hat{\mathbf{v}},\mathbf{s}) \\ f_{\partial^2\tau_r}(\tau_r) &\simeq \gamma\text{trace}[\boldsymbol{\Gamma}_{\mathbf{z}^\star,\gamma}\boldsymbol{\Pi}_r^T\Lambda_{\mathbf{z}_r^\star}\Lambda_{V_r^4}\boldsymbol{\Pi}_r] \\ &+2 \sum_{<\mathbf{s},\mathbf{t}>\in\mathcal{C}}\alpha z_r^\star(\mathbf{s},\mathbf{t})V_r^6(\hat{\mathbf{v}},\mathbf{s},\mathbf{t}). \end{aligned} \quad (78)$$

Maximum *likelihood* estimates $\hat{\tau}_d$ and $\hat{\tau}_r$ are thus obtained for Leclerc's M-estimator using the Newton hyper-parameter update rule (76) with derivates given by (77) and (78).

References

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Int. Conf. on Comp. Vis., ICCV 2007*, 2007.

[2] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *Int. J. Computer Vision*, 12(1):43–77, 1994.

[3] J. Bergen, P. Burt, R. Hingorani, and S. Peleg. A 3-frame algorithm for estimating two-component image motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(9):886–895, Sept. 1992.

[4] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[5] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnorr. A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. *International Journal of Computer Vision*, 70:257–277, 2006.

[6] J. Carlier and B. Wieneke. Report 1 on production and diffusion of fluid mechanics images and data. *Fluid project deliverable 1.2. http://www.fluid.irisa.fr*, 2005.

[7] T. Corpetti, D. Heitz, G. Arroyo, E. Memin, and A. S. Cruz. Fluid experimental flow estimation based on an optical flow scheme. *Experiments in Fluids*, 40:80–97, 2006.

[8] T. Corpetti, E. Mémin, and P. Pérez. Dense estimation of fluid flows. *Pattern Anal Mach Intel*, 24(3):365–380, 2002.

[9] A. Cuzol and E. Memin. A low dimensional fluid motion estimator. *Int. J. Computer Vision*, 75(3):329–349, 2007.

[10] P. Derian, P. Heas, C. Herzet, and E. Memin. Wavelet-based fluid motion estimation. In *Scale Space Methods and Variational Methods (SSVM) in Computer Vision*, Israel, June 2011.

[11] V. Estrela and N. Galatsanos. Spatially-adaptive regularized pel-recursive motion estimation based on cross-validation. In *International Conference on Image Processing*, volume 2, pages 200 –203 vol.2, oct 1998.

[12] J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman & Hall, 1996.

[13] J. Fitzpatrick. The existence of geometrical density-image transformations corresponding to object motion. *Comput. Vision, Graphics, Image Proc.*, 44(2):155–174, Nov. 1988.

[14] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(3):367–383, 1992.

[15] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984.

[16] A. Girard. A fast monte-carlo cross-validation procedure for large least squares problems with noisy data. *Numerische Mathematik*, 56:1–23, 1989. 10.1007/BF01395775.

[17] P. Heas and E. Memin. Three-dimensional motion estimation of atmospheric layers from image sequences. *IEEE trans. on Geo. and Rem. Sensing*, 46(8):2385–2396, 2008.

[18] P. Heas, E. Memin, D. Heitz, and P. Mininni. Bayesian selection of scaling laws for motion modeling in images. In *International Conference on Computer Vision (ICCV'09)*, Kyoto, Japan, October 2009.

[19] P. Heas, E. Memin, N. Papadakis, and A. Szantai. Layered estimation of atmospheric mesoscale dynamics from satellite imagery. *IEEE trans. on Geo. and Rem. Sensing*, 45(12):4087–4104, 2007.

[20] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[21] M. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Commun. Stat. Simula.*, 18:1059–1076, 1989.

[22] S.-S. Ieng, J.-P. Tarel, and P. Charbonnier. Estimation robuste pour la détection et le suivi par caméra. *Traitement du Signal*, 21(3):205–226, 2004. http://perso.lcpc.fr/tarel.jean-philippe/publis/ts04.html.

[23] E. T. Jaynes. Bayesian methods: General background, 1986.

[24] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[25] K. Krajsek and R. Mester. A maximum likelihood estimator for choosing the regularization parameters in global optic flow methods. In *ICIP*, pages 1081–1084, 2006.

[26] K. Krajsek and R. Mester. Bayesian model selection for optical flow estimation. In *DAGM-Symposium*, pages 142–151, 2007.

[27] D. Krawczyk-Stando and M. Rudnicki. Regularization parameter selection in discrete ill-posed problems - the use of the u-curve. *Applied Mathematics and Computer Science*, 17(2):157–164, 2007.

[28] H. Linhart and W. Zucchini. *Model selection*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[29] T. Liu and L. Shen. Fluid flow and optical flow. *Journal of Fluid Mechanics*, 614:253, Oct. 2008.

[30] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereovision. In *Int. Joint Conf. on Artificial Intel. (IJCAI)*, pages 674–679, 1981.

[31] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[32] E. Mémin and P. Pérez. Dense estimation and object-based segmentation of the optical flow with robust techniques. *IEEE Trans. Im. Processing*, 7(5):703–719, 1998.

[33] A. Mohammad-Djafari. On the estimation of hyperparameters in bayesian approach of solving inverse problems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 495–498 vol.5, Apr 1993.

[34] R. Molina, A. K. Katsaggelos, and J. Mateos. Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Trans. Image Processing*, 8:231–246, 1999.

[35] L. Ng and V. Solo. A data-driven method for choosing smoothing parameters in optical flow problems. In *Proceedings., International Conference on Image Processing.*, volume 3, pages 360–363 vol.3, Oct 1997.

[36] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, New York, NY, 1999.

[37] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics)*. Springer Verlag, New York, 2nd edition, June 2007.

[38] R. T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, N.J., 1970.

[39] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[40] M. Shi and V. Solo. Empirical choice of smoothing parameters in optical flow with correlated errors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages III – 161–4 vol.3, april 2003.

[41] B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London ; New York :, 1986.

[42] D. Sun, S. Roth, J. P. Lewis, and M. Black. Learning optical flow. In *Proceedings of the 10th European Conference on Computer Vision*, pages 83–97, 2008.

[43] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[44] J. Weickert and C. Schnörr. A theoretical framework for convex regularizers in pde-based computation of image motion. *Int. J. Computer Vision*, pages 245–264, 2004.

[45] J. Yuan, C. Schnoerr, and E. Memin. Discrete orthogonal decomposition and variational fluid flow estimation. *Journ. of Math. Imaging & Vison*, 28:67–80, 2007.

[46] C. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. In *Tenth IEEE International Conference on Computer Vision (ICCV' 05)*, volume 2, pages 1308–1315 Vol. 2, Oct. 2005.

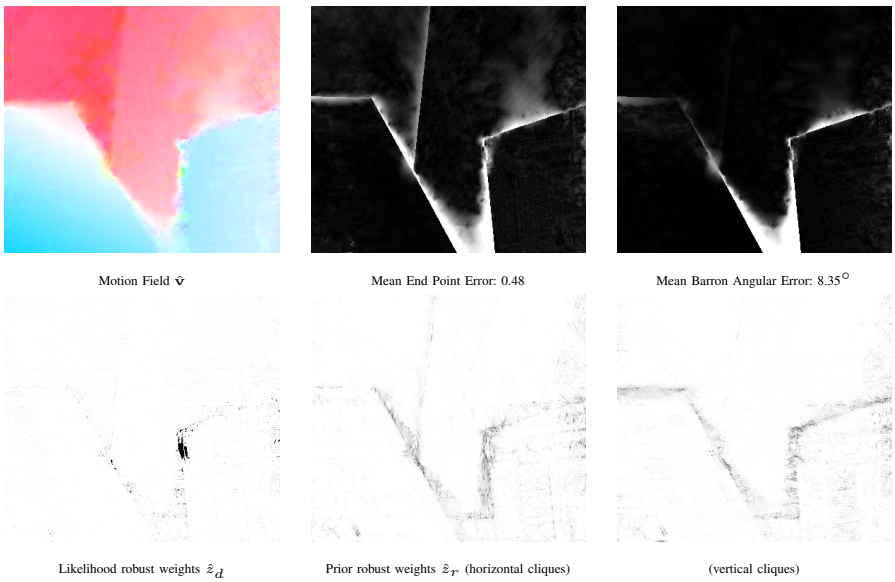


Fig. 11. **Evaluation of most likely motion $\hat{\mathbf{v}}$ estimate.** Color visualization, end point and angular error maps of the motion estimate obtained with the selected *likelihood* model ($\hat{\mathcal{D}}$ = OFC model with Leclerc's norm) and *prior* model ($\hat{\mathcal{R}}$ = first order regularization with $\mathcal{L}^1$ norm) and with the most likely hyper-parameters $(\hat{\gamma}, \hat{\tau}_d, \hat{\tau}_r) = (10.50, 0.01, 1.25)$. The bottom row displays spatial maps of robust weights $\hat{z}_d$ and $\hat{z}_r$ corresponding to the Leclerc's norm parameters $\hat{\tau}_d$ and $\hat{\tau}_r$

Likelihood model $\mathcal{D}$	Prior model $\mathcal{R}$	$\hat{\gamma}$	$\hat{\tau}_d$	$\hat{\tau}_r$	$-\log p(I\|\mathcal{D},\mathcal{R})$	Mean End Point Error	Mean Barron Error
$\mathcal{L}^2$	$\mathcal{L}^2$	12.48	0	0	443013	0.201	3.656
Leclerc	$\mathcal{L}^2$	14.97	0.32	0	418662	0.199	3.542
$\mathcal{L}^1$	$\mathcal{L}^2$	1.85	20.0	0	337990	0.191	3.309
$\mathcal{L}^2$	Leclerc	9.58	0	2.00	437326	0.206	3.760
Leclerc	Leclerc	15.01	0.32	1.34	418602	0.199	3.542
$\mathcal{L}^1$	Leclerc	1.83	20.0	0.39	338097	0.191	3.308
$\mathcal{L}^2$	$\mathcal{L}^1$	3.24	0	10.0	434656	0.258	4.883
Leclerc	$\mathcal{L}^1$	12.28	0.34	10.0	417340	0.204	3.657
$\hat{\mathcal{D}} = \mathcal{L}^1$	$\hat{\mathcal{R}} = \mathcal{L}^1$	1.70	20.0	10.0	**335564**	**0.190**	**3.303**

Fig. 13. **Selection of the most likely norms for the data and the regularization terms.** Hyper-parameters estimates and score of the different norm associations in terms of *evidence* energy, end point and angular error. The *evidence* energy is minimized for the *likelihood* $\hat{\mathcal{D}}$= (OFC model with $\mathcal{L}^1$) and the *prior* $\hat{\mathcal{R}}$= (1-st order regularizer with $\mathcal{L}^1$ norm). The selected models also minimize the end point and the angular average errors.

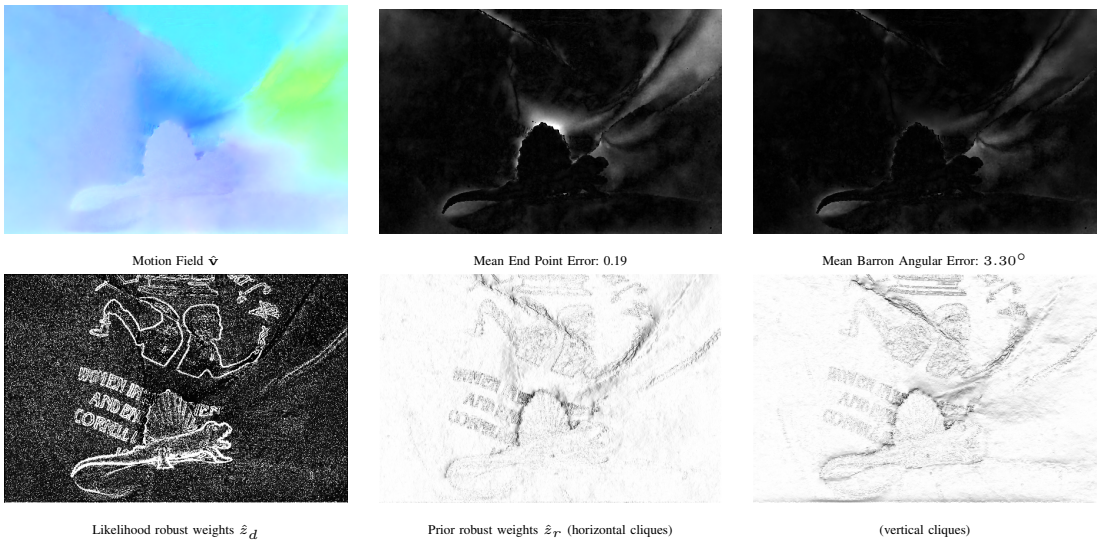


Fig. 14. **Evaluation of most likely motion $\hat{\mathbf{v}}$ estimate.** Color visualization, end point and angular error maps of the motion estimate obtained with the selected *likelihood* model ($\mathcal{D}$ = OFC model with $\mathcal{L}^1$ norm) and *prior* model ($\hat{\mathcal{R}}$ = first order regularization with $\mathcal{L}^1$ norm) and with the most likely hyper-parameters $(\hat{\gamma}, \hat{\tau}_d, \hat{\tau}_r) = (1.69, 20.0, 10.0)$. The bottom row displays spatial maps of robust weights $\hat{z}_d$ and $\hat{z}_r$ corresponding to the Leclerc's norm parameters $\hat{\tau}_d$ and $\hat{\tau}_r$