An Early Cognitive Approach to Visual Motion Analysis

Silvio P. Sabatini and Fabio Solari

Department of Biophysical and Electronic Engineering
University of Genova - Via Opera Pia 11/a
16145 Genova - ITALY
{silvio, fabio}@dibe.unige.it
http://www.pspc.dibe.unige.it

Abstract. Early cognitive vision can be related to the segment of perceptual vision that takes care of reducing the uncertainty on visual measures through a visual context analysis, by capturing regularities over large, overlapping retinal locations, a step that precedes the true understanding of the scene. In this perspective, we defined a general framework to specify context sensitive motion filters based on elementary descriptive components of optic flow fields. The resulting regularized patch-based motion estimation obtained in real-world sequences validated the approach.

1 Introduction

Computer vision procedes through several stages, ranging from low-level (early vision) processes, mainly devoted to feature extraction, to high-level (visual cognitive) processes, dedicated to recognition and dynamic 3D shape inference, up to the extraction of spatio-temporal relationships between the perceptual agent and the scene's objects. In general, there is a gap between early and cognitive vision paradigms. This gap is not only due to their different position in the hierarchical bottom-up scheme of visual processing, but also relates to the different computational paradigms they adopt. Early vision processes are usually based on distributed computation (cf. parallel distributed processing), that can be directly associated to neuronal mechanisms (cf. neuromorphic approach). On the other hand, cognitive processes are traditionally associated to the AI approach, based on symbolic processing and logic, operating in terms of symbols and propositions, and aimed to the understanding of the scene. This leads to systems in which visual feature (like edges, depth, motion, etc.) are computed from early-vision algorithms and those features are then subjected to a relational analysis. In this way, there is a risk of "jumping to conclusions", leaving a distributed representation of visual features too fast, for an hazardous integrated description of cognitive entities. Considering that visual features computed from early vision algorithms are usually error-ridden, it is rather complicated to subject them directly to a relational analysis. Each measure of an observable property of the visual stimulus is, indeed, affected by an uncertainty (not only due to

the additive noise, but also to the fact that the visual properties are themselves random processes) that can be removed, or, better, reduced by making use of additional information (context information, a priori knowledge, etc.). *Early cognitive vision* can be related to the segment of perceptual vision that takes care of reducing the uncertainty on visual measures through a visual context analysis, that is by capturing coherent properties (regularities) over large, overlapping retinal locations (Gestalts [1]), a step that precedes the true understanding of the scene. Following a conventional AI approach, the use of contextual information occurs through the application of specific knowledge-based rules to establish consistency relationships among the extracted visual features. By contrast, in visual cortex, contextual modulation of the sensorial input occurs through dense intra- and inter-area feedback interconnections that integrate context information by modulating cells responses, adapting their tuning and refining their selectivity. We challenged the goal of mimicking cortical computational paradigms to develop parallel distributed processing systems to implement adaptive visual filters, which are fully data-driven and avoid explicit use of AI rules. This would allow to define *context-sensitive filters* (CSFs) based on structural computation rather than on mere calculus. It is important to stress the necessity of re-thinking about cognitive aspects in *structural* terms, by evidencing novel strategies to allow a more direct (i.e., structural) interaction between early vision and cognitive processes, that can be employed by new artificial vision systems. In this perspective, we defined a general framework to specify context sensitive motion filters based on deterministic (i.e., geometric) spatial motion Gestalts. In particular, the geometric properties of the optic flow field have been described through a specific set of elementary gradient-type patterns, as cardinal components of a linear deformation space. By checking the presence of such Gestalts in optic flow fields, we make the interpretation of visual motion more confident. Given motion information represented by an optic flow field, we recognize if a group of velocity vectors belong to a specific pattern, on the basis of their relationships in a spatial neighborhood. Casting the problem as a Kalman filer (KF), the detection occurs through a spatial recurrent filter that checks the consistency between the spatial structural properties of the input flow field pattern and a structural rule expressed by the process equation of the Kalman filter.

2 A Kalman filtering approach to early-cognitive vision

Basic concepts Perception can be viewed as an inference process to gather properties of real-world, or *distal*, stimuli (e.g., an object in space) given the observations of *proximal* stimuli (e.g., the object's retinal image). In this perspective, early cognitive vision can be cast as an *adaptive filter* in which some kind of early-cognitive algorithm plays the role of the *adaptive process*. A general adaptive filtering system is shown in Fig. 1, where $\boldsymbol{x}^*[k]$ is the *unknown*

[1] In this paper, Gestalts are defined as pixel groups with a shared and persistent properties in space and/or time. This concept goes beyond that of a visual "feature", because Gestalts capture the relationships existing among features.

stimulus (the *state*) at time step k, $\boldsymbol{y}[k]$ is the observation of the stimulus (the *measure*), $\hat{\boldsymbol{x}}[k]$ is the estimated stimulus, and $\boldsymbol{x}[k]$ is the reference signal (i.e., what we know about $\boldsymbol{x}^*[k]$). The purpose of a general adaptive system is to filter the input signal $\boldsymbol{y}[k]$ (measure) to invert (in some sense) the measure operator and gain an estimation of $\boldsymbol{x}^*[k]$ by making use of the knowledge $\boldsymbol{x}[k]$. Such a

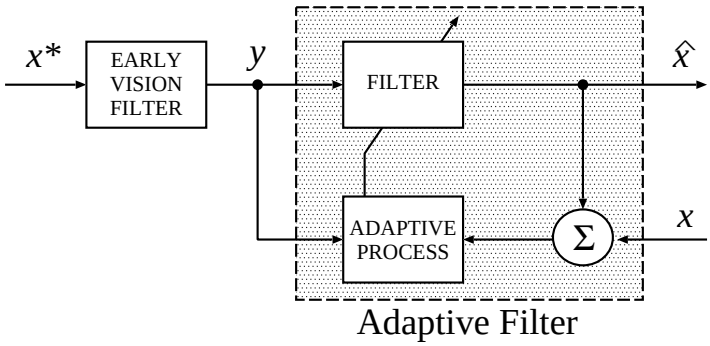

Fig. 1. Schematic representation of an adaptive early vision filter.

knowledge can be provided by:

1. the *visual context*
   - the relationships among the feature values of a single modality in a spatial neighborhood (e.g., responses outside the early vision filter): spatial context;
   - the relationships among the feature values of a single modality in a (spatio-)temporal neighborhood (e.g., the constraints posed by rigid body motion): (spatio-)temporal context;
   - the interdependences of punctual/local feature values from different modalities: multimodal context;
2. the *state of the perceptual agent* (e.g., alert state, task dependency, expectation, etc.)
3. *a priori* information (e.g., cognitive models such as shading, familiarity, perspective, etc.)

In this paper we focus only on data-driven (exogenous) information provided by the visual context, disregarding the other two model-driven (endogenous) components.

Kalman estimator The Kalman Filter is an optimal recursive linear estimator [1], in the sense that it can iteratively process new measures as they arrive, on the basis of the knowledge about the system accrued by previous measurements. Accordingly, a recursive process equation is required to describe the reference signal

(the model). Due to its recurrent formalization it appears particularly promising to design *context-sensitive filters* based on recurrent cortical-like interconnection architectures. Formally, the two inputs to the filter are:

the *process equation*

$$\boldsymbol{x}[k] = \boldsymbol{\Phi}[k, k-1]\ \boldsymbol{x}[k-1] + \boldsymbol{S}[k-1]\ \boldsymbol{s}[k-1] + \boldsymbol{n}_1[k-1] \tag{1}$$

and the *measurement equation*

$$\boldsymbol{y}[k] = \boldsymbol{C}[k]\ \boldsymbol{x}[k] + \boldsymbol{n}_2[k] \tag{2}$$

The matrix $\boldsymbol{\Phi}[k, k-1]$ is a known state transition matrix that relates the state at the previous time step $k-1$ to the state at the current step k. The matrix $\boldsymbol{S}[k]$ takes into the account an optional control input to the state. The matrix $\boldsymbol{C}[k]$ is a known measurement matrix. The process and measurement uncertainty are represented by $\boldsymbol{n}_1[k] = N(0, \boldsymbol{\Lambda}_1[k])$ and $\boldsymbol{n}_2[k] = N(0, \boldsymbol{\Lambda}_2[k])$. The space spanned by the observations $\boldsymbol{y}[1], \boldsymbol{y}[2], \cdots, \boldsymbol{y}[k-1]$ is denoted by $\boldsymbol{\mathcal{Y}}_{k-1}$.

Casting Let us interpret the meaning of the input/output signals of the KF in relation with our perceptual problem.

Measurement equation - The linear operator $\boldsymbol{C}$ represents a general "early-vision filter" providing a noisy measure of an observable property of the visual stimulus.

Process equation - Assuming $\boldsymbol{x}$ a vector containing the values of a bunch of visual features over a fixed spatial region, Eq. 1 models the temporal evolution of the relationships among such features, according to specific rules embedded in the transition matrix $\boldsymbol{\Phi}$. By example, if we consider just one feature (e.g., motion velocity), $\boldsymbol{x}[k]$ will represent the "model" optic flow values at time step k, for all the (discrete) locations of the considered spatial regions (the velocity state). If $\boldsymbol{\Phi}$ has a diagonal structure, the process equation will describe the "model" temporal evolution of punctual velocities, independently of the spatial neighborhood values (temporal context). On the other hand, if $\boldsymbol{\Phi}$ shows a non-diagonal structure, the process equation models a "model" temporal evolution of the state that takes into account *also* spatial relationships (spatio-temporal context). More generally, if we build a state vector that collects more multiple features (e.g., motion, stereo, etc.), by proper specification of the transition matrix $\boldsymbol{\Phi}$, the process equation can potentially model any type of *multimodal* spatio-temporal relationships (multimodal context).

Filter output - Apart from the KF output $\hat{\boldsymbol{x}}$, we could be interested in making the measurements more confident. Accordingly, the output will be $\hat{\boldsymbol{y}}[k|\boldsymbol{\mathcal{Y}}_k]$, to be compared with $\boldsymbol{y}[k]$. The additional (contextual) information will be provided by Kalman innovation. We expect that, if the model is correct, the uncertainty associated to the *a posteriori* estimate of the actual measure $\hat{\boldsymbol{y}}[k|\boldsymbol{\mathcal{Y}}_k]$ is inferior to the uncertainty associated to the actual measure itself $\boldsymbol{y}[k]$.

3 Motion Gestalts

Local spatial features around a given location of a flow field, can be of two types: (1) the average flow velocity at that location, and (2) the structure of the local variation in a the neighborhood of that locality [2]. The former relates to the *smoothness constraint* or *structural uniformity*. The latter relates to *linearity constraint* or *structural gradients*. Velocity gradients provide important cues about the 3-D layout of the visual scene. Formally, they can be described as *linear deformations* by a 2×2 velocity gradient tensor

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \partial v_x/\partial x & \partial v_x/\partial y \\ \partial v_y/\partial x & \partial v_y/\partial y \end{bmatrix} . \tag{3}$$

Hence, if $\boldsymbol{x} = (x, y)$ is a point in a spatial image domain, the linear properties of a motion field $\boldsymbol{v}(x, y) = (v_x, v_y)$ around the point $\boldsymbol{x}_0 = (x_0, y_0)$ can be characterized by a Taylor expansion, truncated at the first order:

$$\boldsymbol{v} = \bar{\boldsymbol{v}} + \bar{\mathbf{T}}\boldsymbol{x} \tag{4}$$

where $\bar{\boldsymbol{v}} = \boldsymbol{v}(x_0, y_0) = (\bar{v}_x, \bar{v}_y)$ and $\bar{\mathbf{T}} = \mathbf{T}|_{\boldsymbol{x}_0}$. By breaking down the tensor in its dyadic components, the motion field can be locally described through 2-D maps representing elementary flow components (EFCs):

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \boldsymbol{d}_x^x \left.\frac{\partial v_x}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}_y^x \left.\frac{\partial v_x}{\partial y}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}_x^y \left.\frac{\partial v_y}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}_y^y \left.\frac{\partial v_y}{\partial y}\right|_{\boldsymbol{x}_0} \tag{5}$$

where $\boldsymbol{\alpha}^x : (x, y) \mapsto (1, 0)$, $\boldsymbol{\alpha}^y : (x, y) \mapsto (0, 1)$ are pure translations and $\boldsymbol{d}_x^x : (x, y) \mapsto (x, 0)$, $\boldsymbol{d}_y^x : (x, y) \mapsto (y, 0)$, $\boldsymbol{d}_x^y : (x, y) \mapsto (0, x)$, $\boldsymbol{d}_y^y : (x, y) \mapsto (0, y)$ represent cardinal deformations, basis of the linear deformation space.

It is worthy to note that the components of pure translations could be incorporated in the corresponding deformation components, thus obtaining generalized deformation components in which motion boundaries are shifted or totally absent. Although this does not affect the significance of the Taylor expansion in Eq. 5, the so-modified elementary components, present very different structural properties. Since a template-based approach cannot be used to extract single components, but only to perform pattern matching operations, the linear decomposition of the motion field has significance only for the definition of a proper representation space. Specific templates would be designed to optimally sample that representation space. In this work, we consider two different classes of deformation templates (opponent and non-opponent), each characterized by two gradient types (stretching and shearing), see Fig. 2. Due to their ability to detect the presence and the orientation of velocity gradients and kinetic boundaries, such cardinal EFCs and proper combinations of them resemble the characteristics of the cell in the Middle Temporal visual area (MT) [3] [4]. It is straightforward to derive that these MT-like components are well suited to provide the building blocks for the more complex receptive field properties encountered in the Medial Superior Temporal visual area (MST) [5] [6]:

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \frac{1}{2}(\boldsymbol{d}_x^x + \boldsymbol{d}_y^y)E + \frac{1}{2}(\boldsymbol{d}_x^x - \boldsymbol{d}_x^y)\omega + \frac{1}{2}(\boldsymbol{d}_x^x - \boldsymbol{d}_y^y)S_1 + \frac{1}{2}(\boldsymbol{d}_y^x + \boldsymbol{d}_x^y)S_2$$

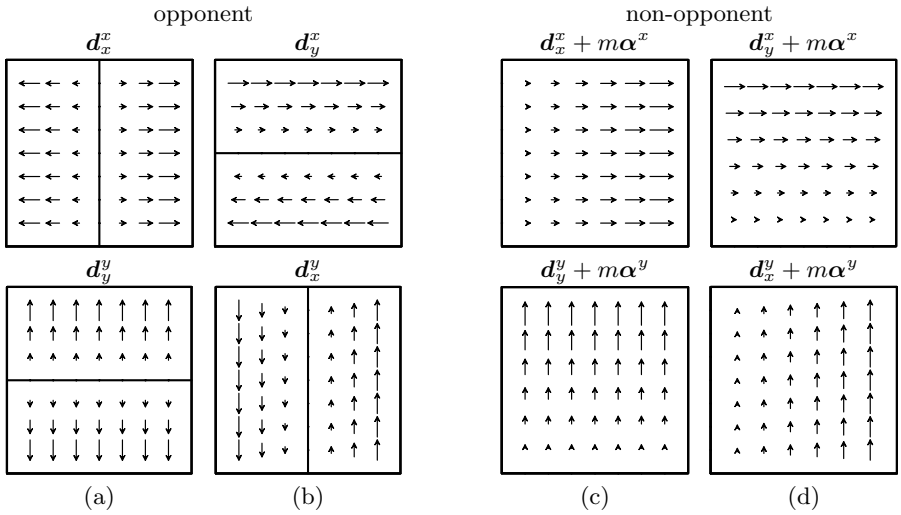

Fig. 2. Basic gradient type Gestalts considered. In stretching-type components (a,c) velocity varies *along* the direction of motion; in shearing-type components (b,d) velocity gradient is oriented *perpendicularly* to the direction of motion. Non-opponent patterns are obtained from the opponent ones by a linear combination of pure translations and cardinal deformations: $\boldsymbol{d}_j^i + m\boldsymbol{\alpha}^i$, where m is a proper positive scalar constant.

where $E = (\bar{T}_{11} + \bar{T}_{22})/2$, $\omega = (\bar{T}_{12} - \bar{T}_{21})/2$, $S_1 = (\bar{T}_{11} - \bar{T}_{22})/2$, $S_2 = (\bar{T}_{12} + \bar{T}_{21})/2$ are the divergence, the curl and the two components of shear deformation, respectively (cf. [2]). These mixed EFCs constitute, together with the pure translations, an equivalent representation basis for the linear properties of the velocity field (see Fig. 3). Yet, they are rather complex since not only the speed, but also the direction of feature motion varies as a function of spatial position. Rigid body motion often generates simpler flow fields characterized by unidirectional patterns, as the cardinal EFCs considered in this study.

4 The context sensitive filter

On the basis of the considerations presented in Section 2, the problem of evidencing the presence of a certain complex feature in the optic flow on the basis of both local and contextual information, is posed as an adaptive filtering problem. Local information act as the input *measurements* and the context acts as the *reference signal*, e.g., representing a specific motion Gestalt.

4.1 Measurement equation

For each spatial position (i, j) and at time step k, let us assume the optic flow $\tilde{\boldsymbol{v}}(i, j)[k]$ as the corrupted measure of the actual velocity field $\boldsymbol{v}(i, j)[k]$. For

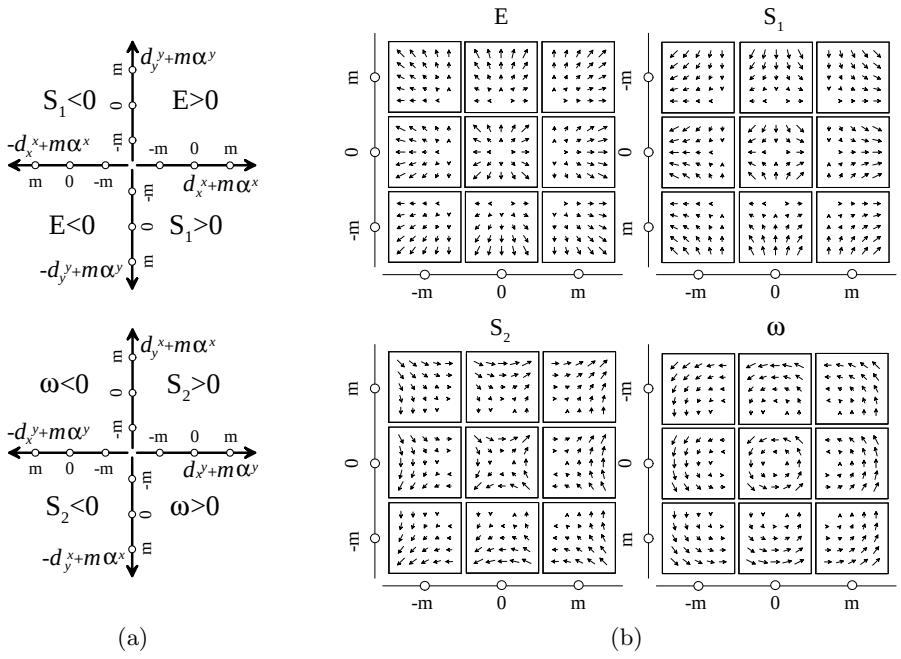

Fig. 3. (a) Two deformation subspaces obtained by the set of cardinal EFCs with different values of the parameter m. The quadrants of each subspace characterize an elementary deformation, as evidenced in (b) for expansion ($E > 0$), horizontal positive shear ($S_1 > 0$), oblique positive shear (S_2), and counterclockwise rotation ($\omega > 0$).

the sake of notation, we drop the spatial indices (i, j) to indicate the vector that represents the whole spatial distribution of a given variable. The difference between these two variables can be represented as a noise term $\boldsymbol{\varepsilon}(i, j)[k]$:

$$\tilde{\boldsymbol{v}}[k] = \boldsymbol{v}[k] + \boldsymbol{\varepsilon}[k] \; . \tag{6}$$

Due to the intrinsic noise of the nervous system, the neural representation of the optic flow $\mathbf{v}[k]$ can be expressed by a *measurement equation*:

$$\mathbf{v}[k] = \tilde{\boldsymbol{v}}[k] + \boldsymbol{n}_1[k] = \boldsymbol{v}[k] + \boldsymbol{\varepsilon}[k] + \boldsymbol{n}_1[k] \tag{7}$$

where $\boldsymbol{n}_1$ represents the uncertainty associated with a neuron's response. In this case the measurement matrix $\boldsymbol{C}$ is the identity operator. The approach can be straightforwardly generalized to consider indirect motion information, e.g., by the gradient equation [7] $-I_t[k] = \boldsymbol{\nabla}^T I[k]\tilde{\boldsymbol{v}}[k] + \boldsymbol{n}_1[k]$ where $\boldsymbol{\nabla}^T I$ and I_t are the spatial image gradient and temporal derivative, respectively, of the image at a given spatial location and time. It is worthy to note that here the linear operator relating the quantity to be estimated to the measurement I_t is *also* a measurement [8].

4.2 Process equation

In the present case, the reference signal should reflect spatio-temporal structural regularities of the input optic flow. These structural regularities can be described statistically and/or geometrically. In any case, they can be defined by a process equation that models spatial relationships by the transition matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{v}[k] = \boldsymbol{\Phi}[k, k-1]\boldsymbol{v}[k-1] + \boldsymbol{n}_2[k-1] + \boldsymbol{s}\ . \tag{8}$$

The state transition matrix $\boldsymbol{\Phi}$ is *de facto* a spatial interconnection matrix that implements a specific Gestalt rule (i.e., a specific EFC); $\boldsymbol{s}$ is a constant driving input; $\boldsymbol{n}_2$ represents the process uncertainty. The space spanned by the observations $\mathbf{v}[1]$, $\mathbf{v}[2]$,..., $\mathbf{v}[k-1]$ is denoted by $\boldsymbol{\mathcal{V}}_{k-1}$ and represents the internal noisy representation of the optic flow. We assume that both $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are independent, zero-mean and normally distributed: $\boldsymbol{n}_1[k] = N(0, \boldsymbol{\Lambda}_1)$ and $\boldsymbol{n}_2[k] = N(0, \boldsymbol{\Lambda}_2)$. More precisely, $\boldsymbol{\Phi}$ models space-invariant nearest-neighbor interactions within a finite region Ω in the (i, j) plane that is bounded by a piece-wise smooth contour. Interactions occur, separately for each component of the velocity vectors (v_x, v_y), through anisotropic interconnection schemes:

$$\begin{aligned} v_{x/y}(i,j)[k] = {} & w_N^{x/y} v_{x/y}(i, j-1)[k-1] + w_S^{x/y} v_{x/y}(i, j+1)[k-1] + \\ & w_W^{x/y} v_{x/y}(i-1, j)[k-1] + w_E^{x/y} v_{x/y}(i+1, j)[k-1] + \\ & w_T^{x/y} v_{x/y}(i,j)[k-1] + n_1^{x/y}(i,j)[k-1] + s_{x/y}(i,j) \end{aligned} \tag{9}$$

where (s_x, s_y) is a steady additional control input, which models the boundary conditions. In this way, the structural constraints necessary to model cardinal deformations are embedded in the lattice interconnection scheme of the process equation. The resulting lattice network has a *structuring effect* constrained by the boundary conditions that yields to structural equilibrium configurations, characterized by specific first-order EFCs. The resulting pattern depends on the anisotropy of the interaction scheme and on the boundary conditions. By example, considering, for the sake of simplicity, a rectangular domain $\Omega = [-L, L] \times [-L, L]$, the cardinal EFC $\boldsymbol{d}_x^x$ can be obtained through:

$$\begin{array}{ll} w_N^x = w_S^x = 0 & w_N^y = w_S^y = 0 \\ w_W^x = w_E^x = 0.5 & w_W^y = w_E^y = 0 \end{array} \quad s_x(i,j) = \begin{cases} -\lambda \text{ if } i = -L \\ \lambda \quad \text{if } i = L \\ 0 \text{ otherwise} \end{cases} \quad s_y(i,j) = 0$$

where the boundary value λ controls the gradient slope. In a similar way we can obtain the other components. Given Eqs. (7) and (8), we may write the optimal filter for optic flow Gestalts. The filter allows to detect, in noisy flows, intrinsic correlations, as those related to EFCs, by checking, through spatial recurrent interactions, that the spatial context of the observed velocities conform to the Gestalt rules, embedded in $\boldsymbol{\Phi}$.

5 Results

To understand how the CSF works, we define the *a priori* state estimate at step k given knowledge of the process at step $k-1$, $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]$, and the *a posteriori* state estimate at step k given the measurement at the step k, $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k}]$. The aim of the CSF is to compute an *a posteriori* estimate by using an *a priori* estimate and a weighted difference between the current and the predicted measurement:

$$\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k}] = \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}] + \boldsymbol{G}[k]\ (\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]) \tag{10}$$

The difference term in Eq. (10) is the *innovation* $\boldsymbol{\alpha}[k]$ that takes into account the discrepancy between the current measurement $\boldsymbol{v}[k]$ and the predicted measurement $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]$. The matrix $\boldsymbol{G}[k]$ is the Kalman gain that minimizes the *a posteriori* error covariance:

$$\boldsymbol{K}[k] = E\left\{(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k}])(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k}])^{T}\right\}\ . \tag{11}$$

Eqs. 10 and 11 represent the mean and covariance expressions of the CSF output.

The covariance matrix $\boldsymbol{K}[k]$ provides us only information about the properties of convergence of the KF and not whether it converges to the correct values. Hence, we have to check the consistency between the innovation and the model (i.e., between observed and predicted values) in statistical terms. A measure of the reliability of the KF output is the Normalized Innovation Squared (NIS):

$$NIS_{k} = \boldsymbol{\alpha}^{T}[k]\ \boldsymbol{\Sigma}^{-1}[k]\ \boldsymbol{\alpha}[k] \tag{12}$$

where $\boldsymbol{\Sigma}$ is the covariance of the innovation. It is possible to exploit Eq. (12) to detect if the current observations are an instance of the model embedded in the KF [9]. Fig. 4 shows the responses of the CSF in the deformation subspaces ($E - S_1$, $\omega - S_2$) for two different input flows. Twentyfour EFC models have been used to span the deformation subspaces shown in Fig. 3a. The grey level in the CSF output maps represents the probability of a given Gestalt according to the NIS criterion: lightest grey indicates the most probable Gestalt. Besides Gestalt detection, context information reduces the uncertainty on the measured velocities, as evidenced, for the circled vectors, by the Gaussian densities, plotted over the space of image velocity.

To assess the performance of the approach to obtain regularized patch-based motion estimation, we applied CSFs to optic flows of real-world driving sequences. Fig. 5 shows a road scene taken by a rear-view mirror of a moving car under an overtaking situations. A "classical" algorithm [10] has been used to extract the optic flow. Regularized motion estimation has been performed on overlapping local regions of the optic flow on the basis of the elementary flow components. In this way, we can compute a dense distribution of the local Gestalt probabilities for the overall optic flow. Thence, we obtain, according to the NIS criterion, the most reliable (i.e. regularized) local velocity patterns, e.g., the patterns of local Gestalts that characterize the sequence (see Fig. 5).

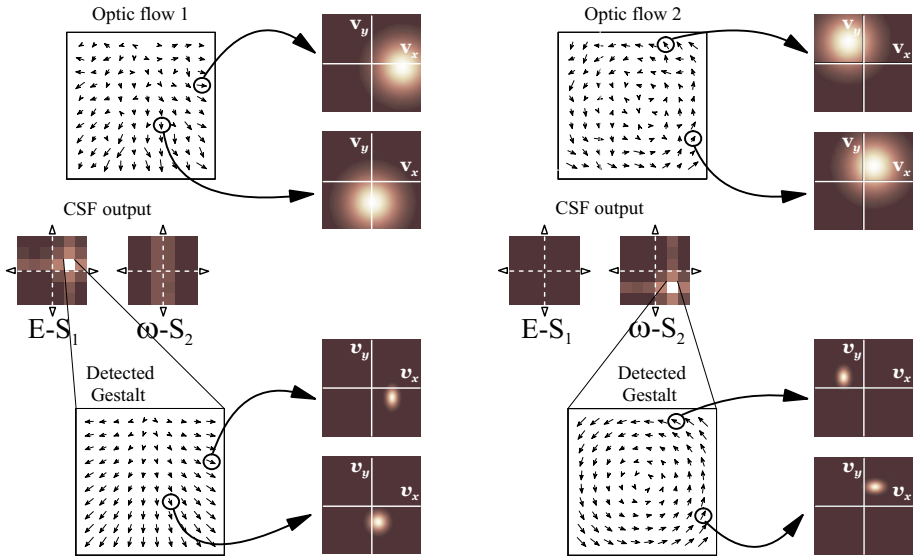

Fig. 4. Example of Gestalt detection in noisy flows.

6 Discussion and conclusions

Measured optic flow fields are always somewhat erroneous and/or ambiguous. First, we cannot compute the actual spatial or temporal derivatives, but only their estimates, which are corrupted by image noise. Second, optic flow is intrinsically an image-based measurement of the relative motion between the observer and the environment, but we are interested in estimating the actual motion field. However, real-world motion field patterns contain intrinsic properties that allow to define Gestalts as groups of pixels sharing the same motion property. By checking the presence of such Gestalts in optic flow fields we obtain context-based regularized patch motion estimation and make the interpretation of the optic flow more confident. We propose an optimal recurrent filter capable of evidencing motion Gestalts corresponding to 1st-order spatial derivatives or elementary flow components. A Gestalt emerges from a noisy flow as a solution of an iterative process of spatially interacting nodes that correlates the properties of the visual context with that of a structural model of the Gestalt. The CSF behaves as a template model. Yet, its specificity lies in the fact that the template character is not built by highly specific feed-forward connections, but emerges by stereotyped recurrent interactions (cf. the process equation). Furthermore, the approach can be straightforwardly extended to consider adaptive cross-modal templates (e.g, motion and stereo). By proper specification of the matrix $\mathbf{\Phi}$, the process equation can, indeed, potentially model any type of multimodal spatio-temporal relationships (i.e., multimodal spatio-temporal context). The presented approach can be compared with Bayesian inference and Markov

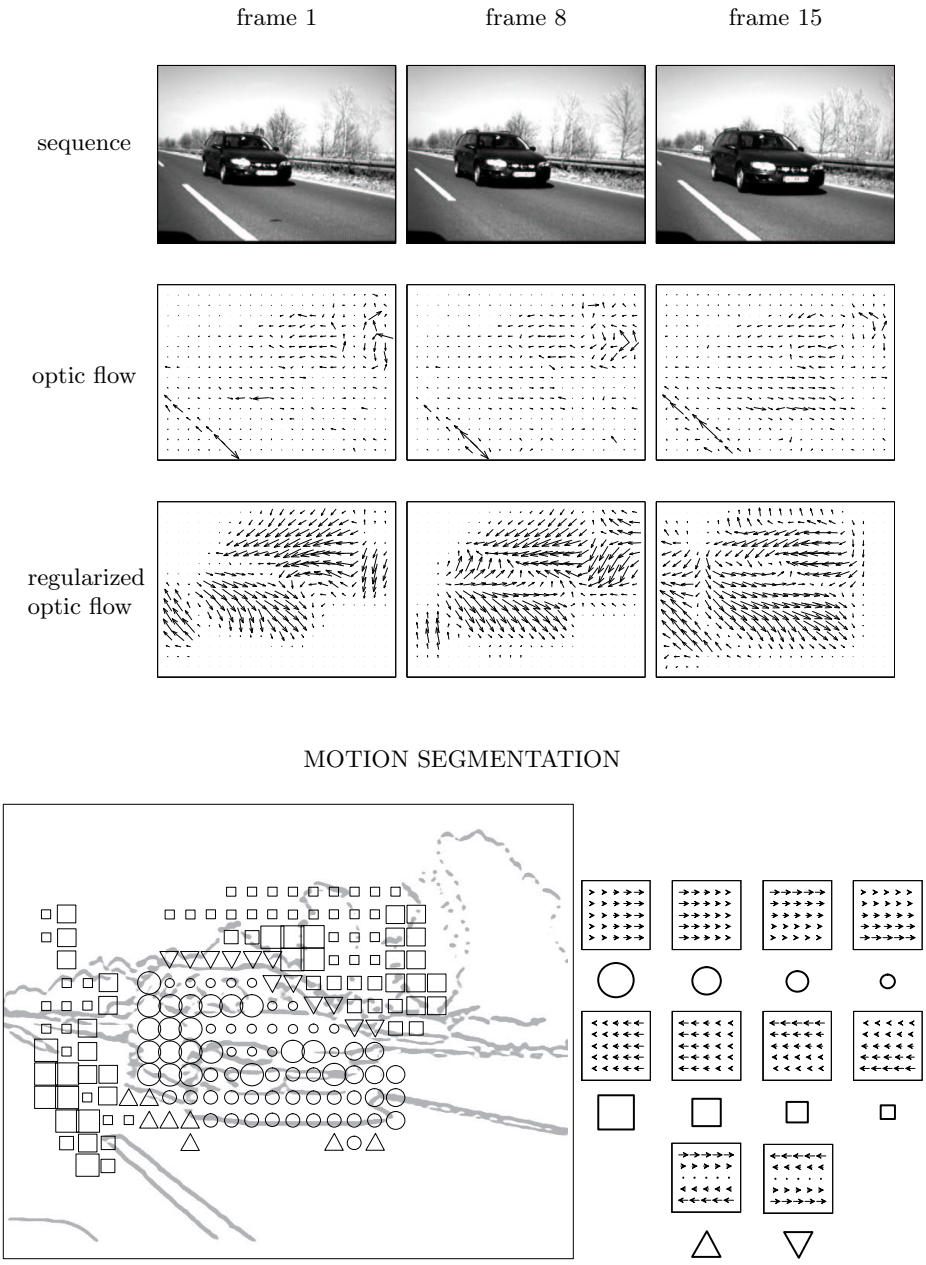

Fig. 5. Results on a driving sequence showing a road scene taken by a rear-view mirror of a moving car under an overtaking situations: Gestalt detection in noisy flows and the resulting motion segmentation (context information reduces the uncertainty on the measured velocities). Each symbol indicates a kind of EFC and its size represents the probability of the given EFC. The absence of symbols indicates that, for the considered region, the reliability of the segmentation is below a given threshold.

Random Fields (MRFs). Concerning Bayesian inference, KF represents a recursive solution to an inverse problem of determing the distal stimulus based on the proximal stimulus, in case we assume: (1) a stochastic version of the regularization theory involving Bayes' rule, (2) Markovianity, (3) linearity and Gaussian normal densities. Concerning MRFs, they are used in visual/image processing to model context dependent entities such as image pixels and correlated features. If one assumes to have not the direct accessibility to the "system", we can refer to dynamic state space models [11] [12] [13] (cf. also Hidden MRF), given by the system's observations and an underlying stochastic process, which is included to describe the distribution of the observation process properly. In this perspective, Kalman's process equation can be related to a MRF. The presence of the measurement equation (observations) makes more evident the distinction between the feed-forward and feed-back components of our CSFs.

Acknowledgments We wish to thank G.M. Bisio and F. Wörgötter for stimulating discussions and M. Müelenberg of Hella for having provided the driving sequences. This work was partially supported by EU Project IST-2001-32114 "ECOVISION".

References

1. S. Haykin. *Adaptive Filter Theory*. Prentice-Hall International Editions, 1991.
2. J.J. Koenderink. Optic flow. *Vision Res.*, 26(1):161–179, 1986.
3. V.L. Marcar, D.K. Xiao, S.E. Raiguel, H. Maes, and G.A. Orban. Processing of kinetically defined boundaries in the cortical motion area MT of the macaque monkey. *J. Neurophysiol.*, 74(3):1258–1270, 1995.
4. S. Treue and R.A. Andersen. Neural responses to velocity gradients in macaque cortical area MT. *Visual Neuroscience*, 13:797–804, 1996.
5. C.J. Duffy and R.H. Wurtz. Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *J. Neuroscience*, 15:5192–5208, 1995.
6. M. Lappe, F. Bremmer, M. Pekel, A. Thiele, and K.P. Hoffmann. Optic flow processing in monkey STS: A theoretical and experimental approach. *J. Neuroscience*, 16:6265–6285, 1996.
7. B.K.P. Horn and B.G. Schunck. Determinig optical flow. *Artificial Intelligence*, 17:185–204, 1981.
8. E.P. Simoncelli. Bayesian multi-scale differential optical flow. In *Handbook of Computer Vision and Applications*, pages 297–322. Academic Press, 1999.
9. Y. Bar-Shalom and X.R. Li. *Estimation and Tracking, Principles, Techniques, and Software*. Artech House, 1993.
10. B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. DARPA Image Understanding Workshop*, pages 121–130, 1981.
11. A.C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge, 1989.
12. M. West and J. Harrison. *Bayesian forecasting and dynamic models*. Springer-Verlag, New York, 1997.
13. H.R. Künsch. State space and hidden Markov models. In *Complex Stochastic Systems, no. 87 in Monographs on Statistics and Applied Probability*, pages 109–173. Chapman and Hall, London, 2001.