Lattice Models for Context-driven Regularization in Motion Perception

Silvio P. Sabatini, Fabio Solari, and Giacomo M. Bisio

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

Abstract. Real-world motion field patterns contain intrinsic statistic 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 can make their interpretation more confident. We propose a context-sensitive recurrent filter capable of evidencing motion Gestalts corresponding to 1st-order elementary flow components (EFCs). 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. By proper specification of the interconnection scheme, the approach can be straightforwardly extended to model any type of multimodal spatio-temporal relationships (i.e., multimodal spatiotemporal context).

1 Introduction

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). The distinction between proximal stimulus and distal stimulus touches on something fundamental to sensory processes and perception. The proximal stimulus, not the distal stimulus, actually sets the receptors' responses in motion. Considering the ill posedness of such inverse problem, one should include *a priori* constraints to reduce the dimension of the allowable solutions, or, in other terms, to reduce the uncertainty on visual measures. These considerations apply both if one tackles the problem of perceptual interpretation as a whole, and if one considers the confidence on single feature measurements. Each measure of an observable property of the stimulus is indeed affected by an uncertainty that can be removed, or, better, reduced by making use of context information. *Early cognitive vision* can be related to that segment of perceptual vision that takes care of reducing the uncertainty on visual measures by capturing coherent properties (Gestalts) over large, overlapping, retinal locations, a step that precedes the true understanding of the scene.

In this perspective, we formulate a probabilistic, model-based approach to image motion analysis, which capture, in each local neighborhood, coherent motion properties to obtain context-based regularized patch motion estimation. Specifically, given motion information represented by an optic flow field, we want to recognize if a group of velocity vectors belongs to a specific pattern, on the basis of their relationships in a spatial neighborhood. Casting the problem as a generalized Kalman filter (KF)[1], 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 KF. Due to its recurrent formulation, KF appears particularly promising to design *context-sensitive filters* (CSFs) that mimic recurrent cortical-like interconnection architectures.

2 Kalman-based perceptual inference

In general, 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, and (3) linearity and Gaussian normal densities. The first condition can be motivated by the fact that the a priori contraints necessary to regularize the solution can be described in probabilistic terms. Bayes' rule allows the computation of the *a posteriory* probability as $p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})/p(\boldsymbol{y})$, where $p(\boldsymbol{x})$ is the *a priori* probability densities for the distal stimulus and represents *a priori* knowledge about the visual scene; $p(\boldsymbol{y}|\boldsymbol{x})$ is the likelihood function for $\boldsymbol{x}$. This function represents the transformation from the distal to proximal stimulus and includes information about noise in the proximal stimulus. Finally, $p(\boldsymbol{y})$ is the probability of obtaining the proximal stimulus. The inverse problem of determining the distal stimulus can be solved by finding $\hat{\boldsymbol{x}}$ that maximizes the *a posteriori* probability, $p(\boldsymbol{x}|\boldsymbol{y})$. Such $\hat{\boldsymbol{x}}$ is called a maximum a posteriori (MAP) estimator. Although the Bayesian framework is more general than the standard regularization, there exist a relationship between the deterministic and stochastic methods of solving inverse problems. Under the assumption of normal probability densities, maximizing the *a posteriory* probability $p(\boldsymbol{x}|\boldsymbol{y})$ is, indeed, equivalent to minimizing the Tikhonov functional. The second concept, the Markovianity, captures the step-by-step local nature of the interactions in a cooperative system, and makes possible Kalman recursion, by allowing to express *global* properties of the state in terms of its *local* properties. Under these hypotheses the conditional probability that the system is in a particular state at any time is determined by the distribution of states at its immediately preceding time. That is, the conditional distribution of that states of a system given the present and past distributions depends only upon the present. Specifically, considering the visual signal as a random field, the Markovianity hypothesis implies that the joint probability distribution of that random field has associated positive-definite, translational invariant conditional probabilities that are spatially Markovian (cf. Markov Random Fields). The third assumption represents the necessary conditions to achieve the exact, analytical solution of the KF.

3 Local 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* (linear deformations). Velocity gradients provide important cues about the 3-D layout of the visual scene. On a local scale, velocity gradients caused by the motion of objects provide perception of their 3-D structure (structure from motion and motion segmentation), whereas, on a global scale, they specify the observer's position in the world, and his/her heading.

Formally, first-order deformations can be described by a 2×2 velocity gradient tensor

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \partial v_x/\partial x & \partial v_x/\partial y \\ \partial v_y/\partial x & \partial v_y/\partial y \end{bmatrix} \; . \tag{1}$$

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:

$$\boldsymbol{v} = \bar{\boldsymbol{v}} + \bar{\mathbf{T}}\boldsymbol{x} \tag{2}$$

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 *cardinal* EFCs:

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \boldsymbol{d}^x_x \left.\frac{\partial v_x}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}^x_y \left.\frac{\partial v_x}{\partial y}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}^y_x \left.\frac{\partial v_y}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}^y_y \left.\frac{\partial v_y}{\partial y}\right|_{\boldsymbol{x}_0} \tag{3}$$

where $\boldsymbol{\alpha}^x : (x, y) \mapsto (1, 0)$, $\boldsymbol{\alpha}^y : (x, y) \mapsto (0, 1)$ are pure translations and $\boldsymbol{d}^x_x : (x, y) \mapsto (x, 0)$, $\boldsymbol{d}^x_y : (x, y) \mapsto (y, 0)$, $\boldsymbol{d}^y_x : (x, y) \mapsto (0, x)$, $\boldsymbol{d}^y_y : (x, y) \mapsto (0, y)$ represent cardinal deformations, basis of the linear deformation 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. 1. More complex local flow descriptors such as the divergence, the curl and the two components of shear, can be straightforwardly obtained by linear combination of such basic templates.

4 The context sensitive filter

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

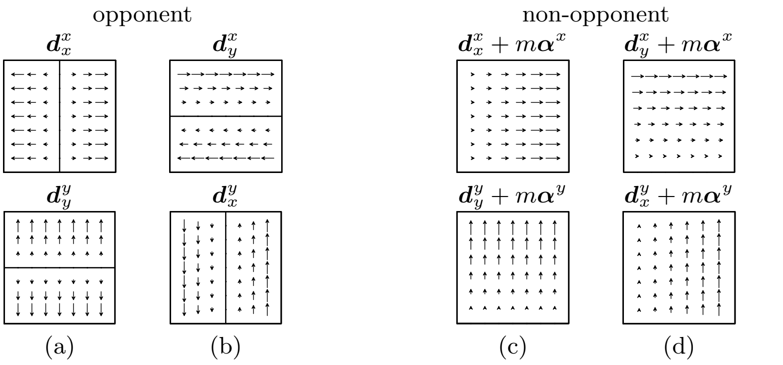
opponent
non-opponent
$\boldsymbol{d}_x^x$
$\boldsymbol{d}_y^x$
$\boldsymbol{d}_x^x + m\boldsymbol{\alpha}^x$
$\boldsymbol{d}_y^x + m\boldsymbol{\alpha}^x$
$\boldsymbol{d}_y^y$
$\boldsymbol{d}_x^y$
$\boldsymbol{d}_y^y + m\boldsymbol{\alpha}^y$
$\boldsymbol{d}_x^y + m\boldsymbol{\alpha}^y$
(a)
(b)
(c)
(d)

Fig. 1. 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 tranlations and cardinal deformations: $\boldsymbol{d}_j^i + m\boldsymbol{\alpha}^i$, where m is a proper positive scalar constant.

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 $\boldsymbol{\varepsilon}(i, j)[k]$:

$$\tilde{\boldsymbol{v}}[k] = \boldsymbol{v}[k] + \boldsymbol{\varepsilon}[k] \ . \tag{4}$$

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*:

$$\mathbf{v}[k] = \tilde{\boldsymbol{v}}[k] + \boldsymbol{n}_1[k] = \boldsymbol{v}[k] + \boldsymbol{\varepsilon}[k] + \boldsymbol{n}_1[k] \tag{5}$$

where $\boldsymbol{n}_1$ represents the uncertainty associated with a neuron's response. The Gestalt is formalized through a *process equation*:

$$\boldsymbol{v}[k] = \boldsymbol{\Phi}[k, k-1]\boldsymbol{v}[k-1] + \boldsymbol{n}_2[k-1] + \boldsymbol{s} \ . \tag{6}$$

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

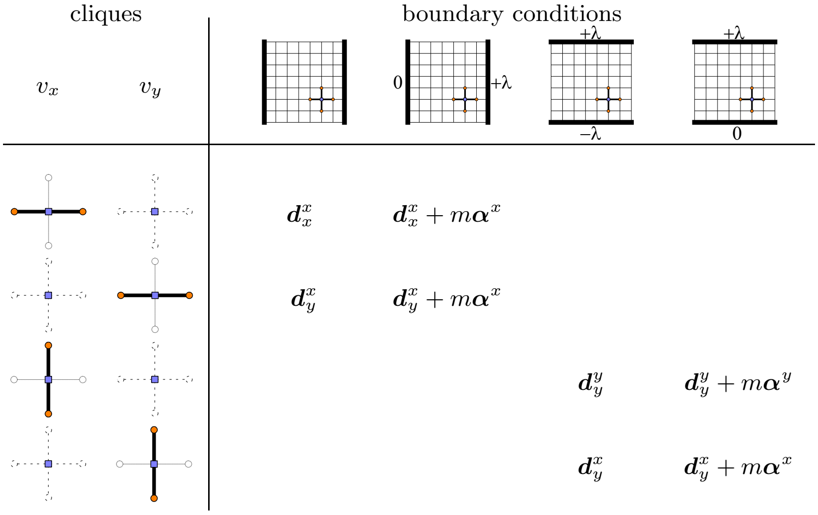
cliques
boundary conditions
v_x
v_y
0
+λ
+λ
−λ
+λ
0
$\boldsymbol{d}_x^x$
$\boldsymbol{d}_x^x + m\boldsymbol{\alpha}^x$
$\boldsymbol{d}_y^x$
$\boldsymbol{d}_y^x + m\boldsymbol{\alpha}^x$
$\boldsymbol{d}_y^y$
$\boldsymbol{d}_y^y + m\boldsymbol{\alpha}^y$
$\boldsymbol{d}_y^x$
$\boldsymbol{d}_y^x + m\boldsymbol{\alpha}^x$

Fig. 2. Basic lattice interconnection schemes for the linear deformation templates considered. The boundary value λ controls the gradient slope.

the velocity vectors (v_x, v_y), through anisotropic interconnection schemes:

$$\begin{aligned} v_{x/y}(i,j)[k] = {} & w_N^{x/y} v_{x/y}(i,j-1)[k-1] + w_S^{x/y} v_{x/y}(i,j+1)[k-1] + \\ & w_W^{x/y} v_{x/y}(i-1,j)[k-1] + w_E^{x/y} v_{x/y}(i+1,j)[k-1] + \\ & w_T^{x/y} v_{x/y}(i,j)[k-1] + n_2^{x/y}(i,j)[k-1] + s_{x/y}(i,j) \end{aligned} \tag{7}$$

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 (see Fig. 2). Given Eqs. (5) and (6), 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:

$$\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k] = \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}] + \boldsymbol{G}[k] \ (\mathbf{v}[k] - \hat{\mathbf{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]) \tag{8}$$

The difference term in Eq. (8) is the *innovation* $\boldsymbol{\alpha}[k]$ that takes into account the discrepancy between the current measurement $\mathbf{v}[k]$ and the predicted measurement $\hat{\mathbf{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]$. The matrix $\boldsymbol{G}[k]$ is the Kalman gain that minimizes the *a posteriori* error covariance:

$$\boldsymbol{K}[k] = E\left\{(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k])(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k])^T\right\} \ . \tag{9}$$

Eqs. 8 and 9 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):

$$NIS_k = \boldsymbol{\alpha}^T[k] \ \boldsymbol{\Sigma}^{-1}[k] \ \boldsymbol{\alpha}[k] \tag{10}$$

where $\boldsymbol{\Sigma}$ is the covariance of the innovation. It is possible to exploit Eq. (10) to detect if the current observations are an instance of the model embedded in the KF [3].

To assess the performances of the CSFs, we applied them to real world optic flows. A “classical” algorithm [4] 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 twenty-four 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 Figs. 3 and 4).

Acknowledgments We wish to thank Martin Mülenberg of Hella for having provided the driving sequences. This work was partially supported by EU Project IST-2001-32114 *ECOVISION “Artificial vision systems based on early cognitive cortical processing”*.

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. Y. Bar-Shalom and X.R. Li. *Estimation and Tracking, Principles, Techniques, and Software.* Artech House, 1993.
4. 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.

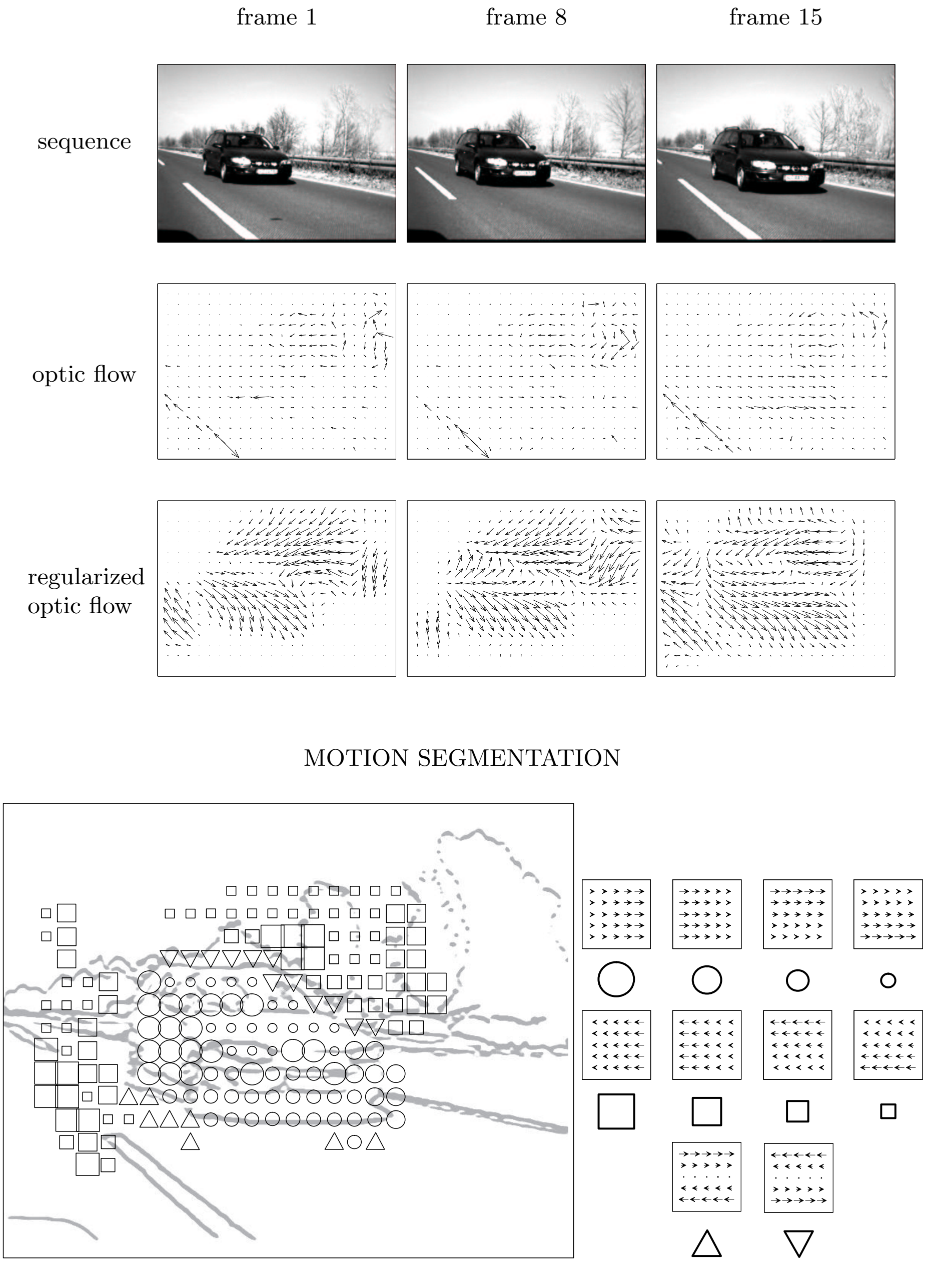

Fig. 3. Results on a driving sequence showing a road scene taken by a rear-view mirror of a moving car under an overtaking situations: Gestalt detection in noisy flows and the resulting motion segmentation (context information reduces the uncertainty on the measured velocities). Each symbol indicates 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.

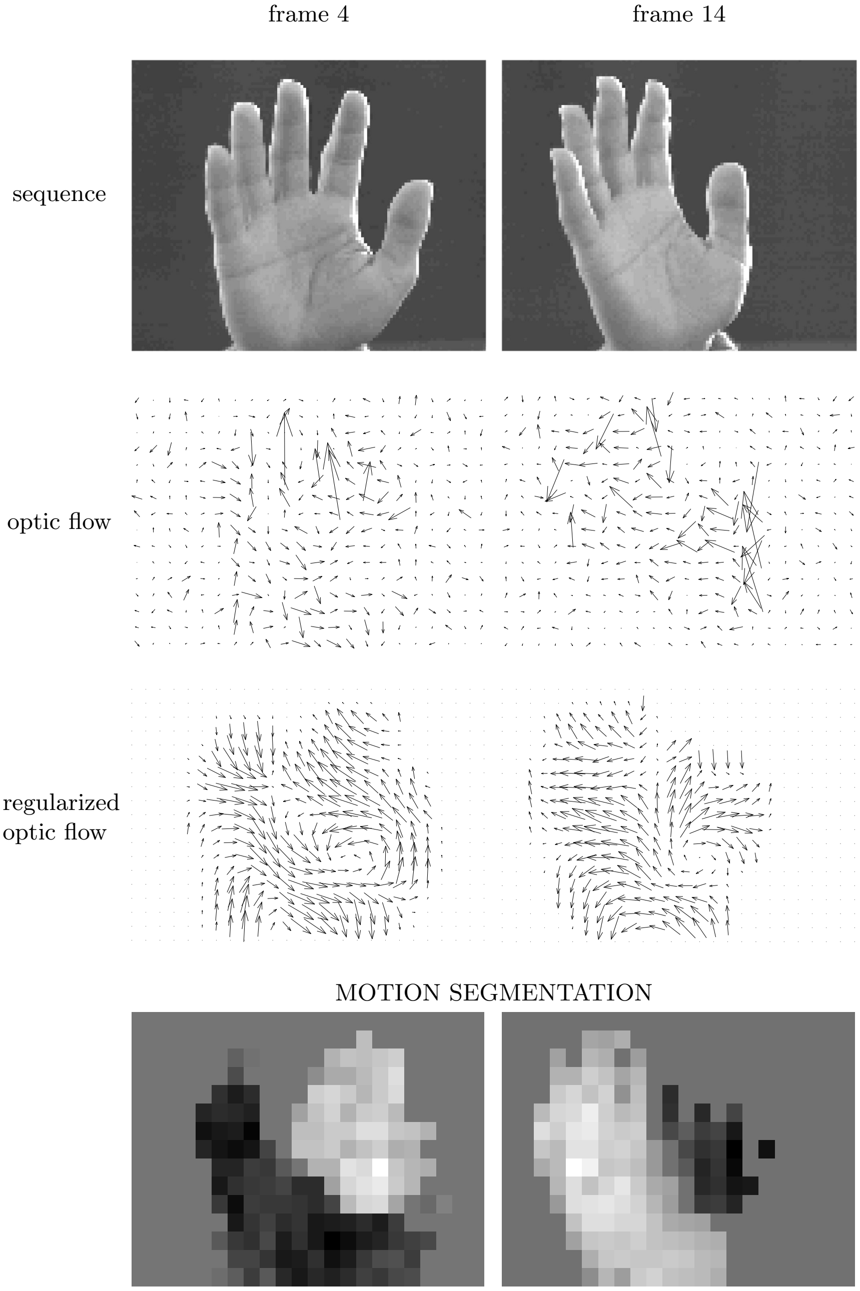

Fig. 4. Context-based patch motion estimation on a sequence showing a hand rotating around its vertical axis. The outputs of the CSFs can be used for motion segmentation evidencing segregation of different motions of each part of the hand: lighter grays indicate leftward motion, whereas darker grays indicate rightward motion.