International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 20, No. 6 (2006) 835–847

World Scientific
www.worldscientific.com

BAYESIAN PROPAGATION FOR PERCEIVING MOVING OBJECTS

GIUSEPPE BOCCIGNONE*, ANGELO MARCELLI† and PAOLO NAPOLETANO‡
Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica
Universitá di Salerno, via Ponte Melillo 1, 84084 Fisciano (SA), Italy
**boccig@unisa.it*
†*amarcelli@unisa.it*
‡*pnapoletano@unisa.it*

VITTORIO CAGGIANO
Dipartimento di Informatica e Sistemistica
Universitá di Napoli Federeico II via Claudio 21, 80125 Napoli, Italy
vcaggian@unina.it

GIANLUCA DI FIORE
CoRiTeL via Ponte Melillo 1, 84084 Fisciano (SA), Italy
difiore@coritel.it

In this paper we address the issue of how form and motion can be integrated in order to provide suitable information to attentively track multiple moving objects. Such integration is designed in a Bayesian framework, and a Belief Propagation technique is exploited to perform coherent form/motion labeling of regions of the observed scene. Experiments on both synthetic and real data are presented and discussed.

Keywords: Bayesian belief propagation; motion estimation and segmentation; visual attention; tracking.

1. Introduction

Visual attention not only restricts various types of visual processing to certain spatial areas of the visual field,[8] but also accounts for object-based information, so that attentional limitations are characterized in terms of the number of discrete objects which can be simultaneously processed.[15] Several theories have been concerned with how object-based individuation, tracking and access are realized and, in particular, Pylyshyn's FINST (FINgers of INSTantiation) proposal has complemented such theories.[15] The model is based on a finite number, say $k \simeq 4, 5$, of visual indexes (fingers, inner pointers) that can be assigned to various items and serve as means of access to such items for higher level processes that allocate focal attention. The visual indexes bestow a processing priority, insofar as they allow focal

attention to be shifted to indexed items, possibly moving, either under volitional control or due to habituation factors, without first searching for them by spatial scanning.

In Ref. 2, a general model was discussed that, grounding in the functional architecture of biological vision, provides a computational account of FINST theory within a Bayesian approach. The Bayesian perspective has been gaining some currency in vision science since Helmholtz conjecture of perception as unconscious inference[9] and is currently the focus of serious investigation (e.g. see Ref. 6, 12 and 13).

In a nutshell, the FINST conjecture may find its Bayesian computational counterpart in the framework of multiple hypotheses tracking coupled with a suitable, top-down modulation of gaze shift.[2] To this end the first issue is the design of a mechanism for instantiating "inner pointers" to each moving object k, in order to keep track of his current state y_t^k (e.g. position and dimension at time t). It is worth remarking that here we use the term "object" in a broad sense, to indicate a coherent region or visual pattern, which is likely to be associated to a physical object in the world (in some way close to the "proto-object" concept in cognitive science[15]). It has been shown[2] that such pointers can be realized as a set of hypotheses that are kept alive in parallel with time. Then, the indexed items can be pursued by Bayesian recursive filtering

$$p(y_t^k|\mathcal{Z}_{t_0:t}^k) \propto p(\mathcal{Z}_t^k|y_t^k) \int p(y_t^k|y_{t-1}^k) p(y_{t-1}^k|\mathcal{Z}_{t_0:t_{n-1}}^k) dy_{t-1}^k, \tag{1}$$

where $p(y_t^k|\mathcal{Z}_{t_0:t}^k)$ is the probability that object k is in state y_t^k at time t, given the sequence of observations $\mathcal{Z}_{t_0:t}^k = \{\mathcal{Z}_t^k, \mathcal{Z}_{t-1}^k, \ldots, \mathcal{Z}_{t_0}^k\}$ and $\mathcal{Z}_t^k$ denotes the set of features observed on the same object. In particular, Eq. (1) can be implemented via the Condensation algorithm.[2,11]

The second issue is the ability to select one object k among other objects $j \neq k$ under volitional control. The winner-take-all strategy has been proposed,[15] which can be implemented[2] via MAP rule on the posterior probabilities $p(y_t^k|\mathcal{Z}_t^k, y_t^j, \mathcal{Z}_t^j)_{j\neq k}$ of gazing, at time t, object k in state y_t^k, given the state and average features of each surrounding object indexed in the scene. The posterior grows as a function of the "feature contrast" of $\mathcal{Z}_t^k$ against $\mathcal{Z}_t^j, j \neq k$ (likelihood) and the commitment of observing object k within a given task or context (prior knowledge). The posterior thus defines a top-down focus of attention (FOA) eventually used to modulate a bottom-up saliency density map, in order to take the final decision (motor command) of setting the gaze at a location (state) y_t^{FOA}.

Clearly, at the heart of this approach [cfr. Eq. (1)], there is the capability of consistently deriving a suitable prediction based on dynamics $p(y_t^k|y_{t-1}^k)$, embodying knowledge about how the object might evolve from time $t-1$ to t, and to perform an update relying upon the likelihood $p(\mathcal{Z}_t^k|y_t^k)$ of the current observation $\mathcal{Z}_t^k$. In this respect, it is worth noting that many approaches use simplified dynamics (e.g. first order models) and observations (e.g. color histograms), while in a complex

vision system the richness of information made available by other visual modules (optical flow, segmentation, etc.) should be exploited.[2]

Here, the very issue we address is that object dynamics, to compute prediction and feature observations to evaluate the likelihood, can be more effectively derived and handled by dynamic integration of form and motion information into consistent percepts of moving forms, which we obtain by resorting to Bayesian propagation machinery.[19,6] Indeed, progress in motion analysis has shown that motion estimation and form segmentation are tightly coupled and that mechanisms of spatial form analysis must be incorporated into the motion estimation procedure. This has led to a new generation of algorithms that iterate between optic flow estimation and segmentation, and namely the Expectation-Maximization (EM) algorithm has been devised as a suitable tool.[18,17]

Here we take one step further by exploiting the Belief Propagation (BP) algorithm to integrate motion and form information. Processing of visual motion in biological systems undergoes two levels of processing, a motion data level and an object-relevant level.[16] The motion data level, primarily involving cortical area V1, uses image filtering mechanisms to extract motion signals, and it has been generally viewed as a purely stimulus-driven filtering process. The object-relevant level is needed to account for motion perception of complex stimuli and is likely to integrate and segment motion information collected from the motion data level into discrete object representations. The dorsal extrastriate cortex, especially the human analogue to monkey MT/MST complex is thought to be a critical cortical site for this type of integrative motion processing. On the other hand, measurements of the color sensitivity in cortical areas linked to the perception of motion, particularly the MT or V5 area, have shown measurable responses to moving isoluminant stimuli containing only chromatic contrast, suggesting that color contributes to moving image segmentation, and that other neurons, perhaps ones with more explicit chromatic signals such as those in V4, are recruited for segmentation purposes.[4]

An emerging consensus is that object-based perceptual and attentional mechanisms may interact with integrative motion processing at this level.[16,4] In the following section we will discuss how Bayesian BP can be suitably adopted to account for such issues and to infer information that eventually could better fit the needs of Bayesian filtering [Eq. (1)].

2. Overview of the Method and Definitions

Assume that K colored objects are observed in a scene, and each object can be described by a vector of parameters $\boldsymbol{\theta}_k$, e.g. the average color $\boldsymbol{\mu}_k$. Such objects undergo different kinds of motion, which can be described by L motion models $\Lambda = \{\boldsymbol{v}_l\}_{l=1}^{L}$; here we denote the motion model $\boldsymbol{v}_l$ as the pair (v_l, ρ_l), speed and direction, respectively, taking values among three possible speeds (slow, average, fast) and eight different directions. In this context, a consistent percept of a moving

form can be defined as a region in which any point of that region is assigned the same label/state s indexing one among $K \times L$ possible motion/form states. Namely, the label represents a "pointer" to access motion and shape features that uniquely defines the object as "that" moving form.

What we propose here is that one such labeling can be formulated as an inference of the "hidden" motion/form state, which relies upon joint observations of motion and shape features.

The input to our system is represented by a pair of subsequent frames $(\mathcal{Z}_{t-1}, \mathcal{Z}_t)$, where each frame is a field $\mathcal{Z}_t = \{\mathbf{z}_{i,t}^{\text{color}}\}_{i=1}^{N}$ of vector-valued random variables $\mathbf{z}_{i,t}^{\text{color}}$ defined in a suitable color space, and index $i \in \Omega$ identifies a site (pixel) in the frame support, the square lattice $\Omega \subseteq \mathbb{Z}^2$.

Let $l \in \mathcal{L} = \{1, 2, \ldots, L\}$ denote motion labels , $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ segmentation labels; labels l, k are used to assign a site i to one of the L motion models and to one of the K objects, respectively. Let $s \in \mathcal{S} = \{1, 2, \ldots, M\}$ denote motion/form labels. $\mathcal{S}$ is named the motion/form state space, defined as the cartesian product $\mathcal{K} \times \mathcal{L}$, of dimension $|\mathcal{S}| = K \times L = M$. In other terms, since l indexes motion models $\{\boldsymbol{v}_l\}_{l=1}^{L}$ and k indexes object parameters $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$, label s is an index for the table $\boldsymbol{m}(s) = [\boldsymbol{v}_l(s), \boldsymbol{\theta}_k(s)]$ representing all combinations of motion models and object parameters describing the observed scene. Let $\mathbf{z}_{i,t}^{OF}$ denote an optical flow vector at a site i. Define motion features as the random variables $z_{i,t}^{\text{motion}}$ that can take values in the motion label set $\mathcal{L}$, and form features as the random variables $z_{i,t}^{\text{form}}$ taking values in the segmentation label set $\mathcal{K}$. Motion and form features can be collected in the random fields $\mathcal{Z}_t^{\text{motion}} = \{z_{i,t}^{\text{motion}}\}_{i=1}^{N}$ and $\mathcal{Z}_t^{\text{form}} = \{z_{i,t}^{\text{form}}\}_{i=1}^{N}$, respectively; a realization of $\mathcal{Z}_t^{\text{motion}}$ is denoted *motion map*, while a *segmentation map* is a realization of $\mathcal{Z}_t^{\text{form}}$.

Motion and form features can be combined into a joint observation $z_{i,t}^{\text{obs}}$, given motion and form observations $z_{i,t}^{\text{motion}} = \hat{l}$, $z_{i,t}^{\text{form}} = \hat{k}$, by assigning $z_{i,t}^{\text{obs}} = s$ so that $\boldsymbol{m}(s) = [\boldsymbol{v}_{l \equiv \hat{l}}(s), \boldsymbol{\theta}_{k \equiv \hat{k}}(s)]$ holds. Such variables can define the random field $\mathcal{Z}_t^{\text{obs}} = \{z_{i,t}^{\text{obs}}\}_{i=1}^{N}$; a realization of the latter will be named *joint observation map.*

Eventually, let $\mathcal{X}_t = \{x_{i,t}\}_{i=1}^{N}$ denote the random field of *hidden* random variables $x_{i,t} \in \mathcal{S}$. Thus, the problem we address here is to infer the most likely motion/form state $\mathcal{X}_t$ on the basis of the joint observation $\mathcal{Z}_t^{\text{obs}}$. The method can be summarized in the following steps.

For each pair of subsequent frames $(\mathcal{Z}_{t-1}, \mathcal{Z}_t)$:

(1) Compute optical flow field $\{\mathbf{z}_{i,t}^{\text{OF}}\}_{i=1}^{N}$. Obtain motion map $\mathcal{Z}_t^{\text{motion}}$ by assigning to each site i the most probable velocity model as $z_{i,t}^{\text{motion}} = \arg\max_l \{p(\mathbf{z}_{i,t}^{\text{OF}} | l, \boldsymbol{v}_l)\}$.
(2) Compute the form map $\mathcal{Z}_t^{\text{form}}$ by assigning to each site i, $z_{i,t}^{\text{form}} = \arg\max_k \{p(\mathbf{z}_{i,t}^{\text{color}} | k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$.
(3) Given $\mathcal{Z}_t^{\text{motion}}$ and $\mathcal{Z}_t^{\text{form}}$, compute the joint map $\mathcal{Z}_t^{\text{obs}}$ by assigning to each site i the state $z_{i,t}^{\text{obs}} = s$ consistent with motion and form observations at that site.

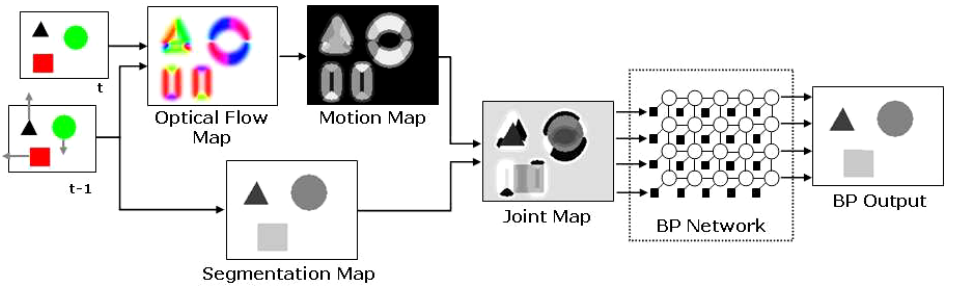


Fig. 1. Functional outline of the method and intermediate results rendered as gray level maps. From left to right: the input pair of frames with superimposed velocity vectors; optical flow map coded in HSV color space ($V = \text{const}$) to represent speeds (S) and directions (H) followed by its coarse motion coding by using three speeds and eight directions. The latter and the segmentation result are combined into the joint observation map $\mathcal{Z}_t^{\text{obs}}$. The bipartite graph represents the set of hidden variables $\mathcal{X}_t$ (circle nodes), coupled with the joint observation map.

(4) Use a loopy Belief Propagation algorithm to infer the most likely "hidden" map $\mathcal{X}_t$, through the joint density $p(\mathcal{X}_t, \mathcal{Z}_t^{\text{obs}})$ represented via a graphical model with a pairwise Markov network topology.

Note that step (1) results in a quantization of the motion field, while step (2) performs a segmentation of the observed scene. Eventually, the BP step integrates such information by taking into account spatial constraints and thus inferring a coherent moving form. Intermediate results of the different processing steps are illustrated in Fig. 1 by using a simple example of synthetic moving objects: namely, a black triangle, a green disk and a red square that are moving in different directions and with different speeds. The same example will be exploited throughout this section to detail the proposed approach. It is easy to note that, even in this "toy" example, features derived from motion analysis, although quantized, and segmentation are *per se* unreliable for characterizing a moving form, and the joint map itself could not be straightforwardly used for such purpose. This remark motivates the introduction of an inference step performed by resorting to Belief Propagation.

3. Computation of Motion Features

Results presented in Fig. 1 (cfr. the optical flow map) give evidence of the general problem that optical flow fields derived from multiple motions usually display discontinuities (motion edges) and sparseness. This poses a severe issue on direct exploitation of the flow map to characterize motion at the object level.[18,17]

To overcome such drawback, we assume that the input to the network should capture tuning properties of MT neurons in terms of their velocity selectivity.[20,7] Rather than model all of the details in the neural circuits that might be responsible to achieve such tuned responses,[7] we instead use a simpler system (similarly to Ref. 20) to compute a quantized velocity encoding (Fig. 2). To this end, the initial

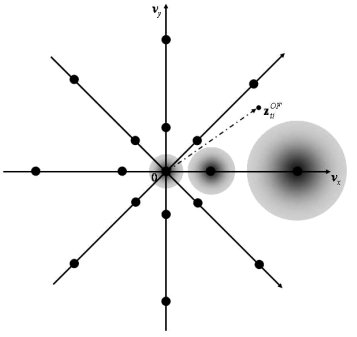


Fig. 2. Motion models located as points within the velocity frame, denoting a discrete set of velocities $\Lambda = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L\}$.

velocity flow field $\{\mathbf{z}_{t,i}^{\mathrm{OF}}\}_{i=1}^{N}$ is obtained by using Horn–Shunk algorithm;[10] an example is provided in Fig. 1. Then, we assume that a number L of possible velocities (motion models) exists, each characterized by different speed and direction. The latter are represented by a finite set of locations $\Lambda = \{\boldsymbol{v}_l\}_{l=1}^{L}$ in a velocity reference frame, where index l labels a motion model (location) and axes represent components v_{lx} and v_{ly} as in Fig. 2; in other terms, each location is tuned to a different velocity. Three speeds (slow, average, fast) and eight different directions are used as illustrated in Fig. 2; speed quantization is adaptively determined on the basis of $\mathbf{z}_{i,t}^{\mathrm{OF}}$ distribution (histogram). The actual velocity vector $\mathbf{z}_{i,t}^{\mathrm{OF}} = [zx_{i,t}^{\mathrm{OF}}, zy_{i,t}^{\mathrm{OF}}]^T$ at an image point i, as obtained by optical flow, is encoded by a finite mixture of velocity receptor units (2D Gaussian functions) centered on frame points $\boldsymbol{v}_l \in \Lambda$:

$$p(z_{i,t}^{\mathrm{motion}}|\Lambda) = \sum_{l=1}^{L} P(l)p(\mathbf{z}_{i,t}^{\mathrm{OF}}|l, \boldsymbol{v}_l), \tag{2}$$

with

$$p(\mathbf{z}_{i,t}^{\mathrm{OF}}|l, \boldsymbol{v}_l) = \frac{1}{(2\pi)^{(D/2)}\boldsymbol{\sigma}_l^{1/2}} \exp\left(-\frac{(zx_{i,t}^{\mathrm{OF}} - v_{lx})^2 + (zy_{i,t}^{\mathrm{OF}} - v_{ly})^2}{2\boldsymbol{\sigma}_l^2} \right), \tag{3}$$

where $D = 2$ and $P(l)$ represents the prior probability of observing a kind of motion. In the absence of context (e.g. a cognitive bias), $P(l)$ can be retained as uniform. Each point in the velocity space thus encodes the degree to which the local velocity matches its preferred velocity. Note that parameter σ_l, which is responsible for the corresponding velocity "receptive field" width, increases with speed in order to provide a uniform covering of nonuniform sampling space (Fig. 2).

In order to associate a model l to each pixel, we have to find the $\max_l\{p(\mathbf{z}_{i,t}^{\mathrm{OF}}|l, \boldsymbol{v}_l)\}$ probability. Eventually, we obtain the motion map $\mathcal{Z}_t^{\mathrm{motion}} = \{z_{i,t}^{\mathrm{motion}}\}_{i=1}^{N}$at time t, by setting at each site i

$$z_{i,t}^{\mathrm{motion}} = \arg\max_l \{p(\mathbf{z}_{i,t}^{\mathrm{OF}}|l, \boldsymbol{v}_l)\}. \tag{4}$$

An example rendered as a gray level map is provided in Fig. 1.

4. Computation of Form Features

Initial form features are derived through segmentation, that is by assigning a label k to each site i, given the observed data $\mathbf{z}_{i,t}^{\text{color}} = [Y_{i,t}, Cb_{i,t}, Cr_{i,t}]^T$ in the $YCbCr$ color space. Segmentation is accomplished via Diffused Expectation Maximization (DEM),[3] a variant of the expectation maximization (EM) algorithm. The method models an image/frame as a finite mixture, where each mixture component corresponds to a region class and uses a maximum likelihood approach to estimate the parameters of each class, via the EM algorithm, coupled with anisotropic diffusion on classes, in order to account for the spatial dependencies among pixels.

To this end, the probabilistic model is assumed to be the mixture

$$p(\mathbf{z}_{i,t}^{\text{color}}|\Theta) = \sum_{k=1}^{K} P(k)p(\mathbf{z}_{i,t}^{\text{color}}|k, \boldsymbol{\theta}_k), \tag{5}$$

where $\Theta = \{\boldsymbol{\theta}_k, k\}_{k=1}^K$ and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the vector of the parameters (mean vectors and covariance matrices) associated to label $k \in \mathcal{K}$. Each label k defines a particular region/form, and $p(\mathbf{z}_{i,t}^{\text{color}}|k, \boldsymbol{\theta}_k)$ are multivariate gaussians

$$p(\mathbf{z}_{i,t}^{\text{color}}|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\exp(-\frac{1}{2}(\mathbf{z}_{i,t}^{\text{color}} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{z}_{i,t}^{\text{color}} - \boldsymbol{\mu}_k)}{(2\pi)^{(D/2)}|\boldsymbol{\Sigma}_k|^{1/2}}, \tag{6}$$

weighted by mixing proportions $P(k)$. Note that, we can consider the covariance matrices being diagonal because of the choice of the $YCrCb$ color space, and, furthermore we assume K fixed, in that we are not concerned here with the problem of model selection. Parameters of each object are estimated via DEM.[3] After the parameter estimation stage has been completed, segmentation is achieved by assigning to each site i, the label k for which $\max_k\{p(\mathbf{z}_{i,t}^{\text{color}}|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$ holds:

$$z_{i,t}^{\text{form}} = \arg\max_k\{p(\mathbf{z}_{i,t}^{\text{color}}|k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}. \tag{7}$$

The assignment produces the segmentation map at time t, $\mathcal{Z}_t^{\text{form}}$ (see Fig. 1).

5. Inference of Moving Forms via Belief Propagation

At this stage, local observations of both motion and form features are available at each point of the observed scene, and collected into the motion and segmentation maps, $\mathcal{Z}_t^{\text{motion}}$ and $\mathcal{Z}_t^{\text{form}}$. Then, the integration of such features into consistent percepts can be formulated in terms of the inference, for each point i, of the most likely joint motion/form state $x_{i,t}$, given $\mathcal{Z}_t^{\text{motion}}$ and $\mathcal{Z}_t^{\text{form}}$.

Here we show how such inference can be accomplished through Belief Propagation.[19,14] BP algorithms can best be understood by imagining that each node in a Markov net, which is responsible for a local observation, communicates by "messages" with other connected nodes about what their beliefs should be. The messages converge after a finite number of steps, when each node has correctly computed its own belief $b(x_{i,t})$ (posterior distribution).

Formally, we want to estimate the joint probability $p(\mathcal{X}_t, \mathcal{Z}_t^{\text{form}}, \mathcal{Z}_t^{\text{motion}})$, where $\mathcal{X}_t = \{x_{i,t}\}_{i=1}^N$ is the field of *hidden* random variables $x_{i,t}$ taking values in $\mathcal{S}$.

To this end, we use the *joint observation* $\mathcal{Z}_t^{\text{obs}} = \{z_{i,t}^{\text{obs}}\}_{i=1}^N$ derived from the pair $(\mathcal{Z}_t^{\text{color}}, \mathcal{Z}_t^{\text{motion}})$ as described in Sec. 2, by assigning $z_{i,t}^{\text{obs}} = s$ so that $\boldsymbol{m}(s) = [\boldsymbol{v}_{l\equiv\hat{l}}(s), \boldsymbol{\theta}_{k\equiv\hat{k}}(s)]$ holds, and where $\hat{l}$ and $\hat{k}$ are consistent with motion and form observation $z_{i,t}^{\text{motion}} = \hat{l}$, $z_{i,t}^{\text{form}} = \hat{k}$ at the same site i. For what concerns object parameters, we only retain the vector mean $\boldsymbol{\mu}_k$ and omit covariance $\boldsymbol{\Sigma}_k$. Also, each motion model $\boldsymbol{v}_l$ is represented in terms of speed and direction (v_l, ρ_l). Thus label s provides access to features $(v_l(s), \rho_l(s), \boldsymbol{\mu}_k(s))$ in the look-up table $\boldsymbol{m}(s)$, namely the quantized speed and direction of motion, and average color of the kth region. Also it is worth remarking that at this stage, the state space $\mathcal{S}$ is dynamically reduced to those state/models that have been actually employed; in other terms the cardinality of the space is $|\mathcal{S}| = \widehat{K} \times \widehat{L} = \widehat{M}$, where $\widehat{M} \leqslant M$.

The random field $\mathcal{Z}_t^{\text{obs}}$ represents the set of *observed* variables to estimate the density $p(\mathcal{X}_t, \mathcal{Z}_t^{\text{form}}, \mathcal{Z}_t^{\text{motion}})$ via $p(\mathcal{X}_t, \mathcal{Z}_t^{\text{obs}})$, where $\mathcal{Z}_t^{\text{motion}}, \mathcal{Z}_t^{\text{form}}, \mathcal{Z}_t^{\text{obs}}, \mathcal{X}_t$ share the same support (topology), the connected grid Ω. Then, coupling between motion and form modules can be represented via a graphical model, with a pairwise Markov network topology as illustrated in Fig. 1. Define E the corresponding set of edge indexes of the set $\mathcal{X}_t$; two nodes, say $i, j \in \Omega$ are correlated if and only if the index associated to the edge, in this case (i, j), exists in the set E. The overall or "joint" probability that defines a generative model on this graph is

$$p(\mathcal{X}_t, \mathcal{Z}_t^{\text{obs}}) = \frac{1}{Z_Q} \prod_{(i,j)\in E} \psi_{i,j}(x_{i,t}, x_{j,t}) \prod_{i=1}^{N} \phi_i(x_{i,t}, z_{i,t}^{\text{obs}}), \tag{8}$$

where $\phi_i(x_{i,t}, z_{i,t}^{\text{obs}})$ represents the compatibility function between $x_{i,t}$ and $z_{i,t}^{\text{obs}}$, also called the evidence for $x_{i,t}$, and $\psi_{i,j}(x_{i,t}, x_{j,t})$ represents the compatibility function between $x_{i,t}$ and $x_{j,t}$, also called the interaction between i and j.[19] The main goal is to find the belief $b(x_{i,t}) = p(x_{i,t}, \mathcal{Z}_t^{\text{obs}})$, that is the marginal probability distribution of each node to be in a state $x_{i,t}$.

The belief at each node could be obtained by marginalizing $p(\mathcal{X}_t, \mathcal{Z}_t^{\text{obs}})$; unfortunately, marginalization is not an efficient method due to exponential in the size of the graph. To turn an exponential inference computation into one which is linear, Belief Propagation (BP) algorithms were proposed,[19] that calculate beliefs by local message-passing where each message is defined as:[19]

$$m_{ij}(x_{j,t}) = \beta \sum_{x_{i,t}\in\mathcal{S}} \left[\psi_{j,i}(x_{j,t}, x_{i,t}) \phi(x_{i,t}, z_{i,t}^{\text{obs}}) \times \prod_{s\in\Gamma(i)\setminus j} m_{si}(x_{i,t}) \right], \tag{9}$$

where $\Gamma(i) \triangleq \{j|(i, j) \in E\}$ defines the neighborhood of node i. For graphs which are acyclic the BP algorithm gives the exact marginal probability distribution[14]

$$b(x_{i,t}) = p(x_{i,t}, \mathcal{Z}_t^{\text{obs}}) = \alpha\phi(x_{i,t}, z_{i,t}^{\text{obs}}) \prod_{j\in\Gamma(i)} m_{ji}(x_{i,t}), \tag{10}$$

where α is a normalization constant, and $\sum_{x_{i,t}\in\mathcal{S}} b(x_{i,t}) = 1$. Notwithstanding the grid topology we are exploiting, strong empirical results and recent theoretical work

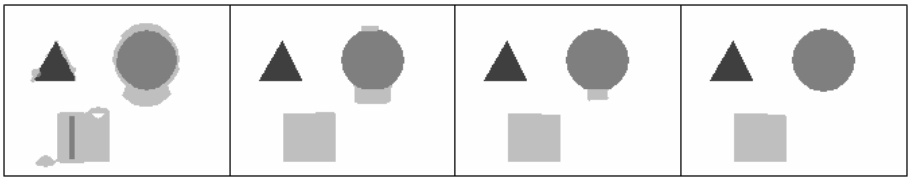


Fig. 3. BP evolution at iterations $1, 15, 35, 60$. The right-most map represents as gray level the final form/motion labeling $\mathcal{X}_t$ achieved.

provide support for a very simple approximation: applying the propagation rules above even in a network with loops.[5] Yet, we have to solve the problem of designing suitable compatibility functions ϕ and ψ.

5.1. *Compatibility functions*

In order to model compatibility functions $\phi(x_{i,t}, z_{i,t}^{\text{obs}})$ and $\psi(x_{i,t}, x_{j,t})$, recall that, according to the discrete formulation of the BP algorithm we have provided, both the observations $z_{i,t}^{\text{obs}}$ and the hidden states $x_{i,t}$ take values within the set $\mathcal{S}$ labeling the $\widehat{M}$ form/motion models. Compatibilities can be determined[5] as $\phi(x_{i,t}, z_{i,t}^{\text{obs}}) \propto p(x_{i,t}, z_{i,t}^{\text{obs}}), \psi(x_{i,t}, x_{j,t}) \propto p(x_{i,t}, x_{j,t})$, that is in both cases, due to our representation, as $p(s, s')$, $s, s' \in \mathcal{S}$ indexing a pair of models. In the vein of Ref. 5, we assume a Gaussian penalty

$$p(s, s') = \prod_{q=1}^{3} \exp\left(-\frac{(m_q(s) - m_q(s'))^2}{2\sigma_q^2} \right), \tag{11}$$

where $m_q(s)$ represents one of three fields of table $\boldsymbol{m}(s) = [v_l(s), \rho_l(s), \boldsymbol{\mu}_k(s)]$ indexed by s and σ_q^2 is a penalty parameter.

By providing initialization and compatibility functions obtained as described above, the BP algorithm iterates message passing among nodes [see Eq. (9)] until convergence to a final state map $\mathcal{X}_t$ (Fig. 1). Convergence condition[6] is obtained as $\frac{1}{N}\sum_{i=1}^{N} |b(x_{i,t}) - b(x_{i,t-1})| < \epsilon$, where ϵ is experimentally determined ($\epsilon = 0.004$). In Fig. 3, an excerpt of intermediate outputs of BP evolution is shown.

6. Experimental Work

Different clips have been produced to simulate different conditions, one synthetically generated and three representing fixed-camera outdoor sequences. Due to limitations of space, we present here results obtained on a single outdoor clip, which is the most critical with respect to motions and lighting conditions, with people walking at different distances from the camera, at different speeds and directions. Figure 4 illustrates results of the proposed method obtained on a pair of frames of the sequence; the top row shows the different maps as described in Fig. 1, while

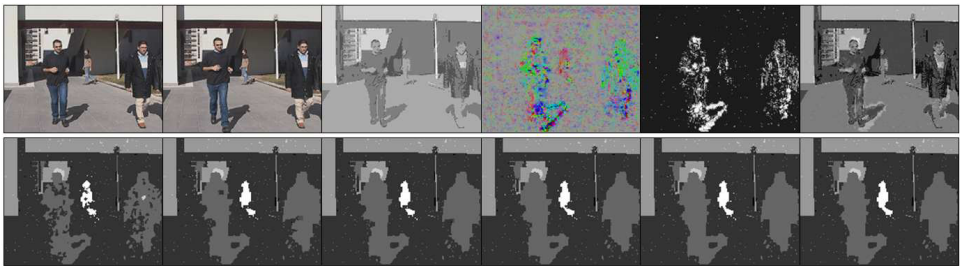


Fig. 4. Example of results on a real sequence. Top row, from left to right: input frames, maps from segmentation, optical flow, motion coding, joint observation. Bottom row: BP evolution at iterations $1, 10, 15, 20, 25, 30$. The right-most map represents the form/motion labeling achieved.

the bottom row of the image shows BP evolution, converging after 30 iterations ($\epsilon = 0.004$). Segmentation was performed by using initially $K = 6$ object classes, while all motion models $L = 17$ were allowed ($M = 104$). After these steps only $\widehat{M} = 6$ models survived. Also, segmentation was obtained after only a single iteration of the DEM algorithm. The motivation for limiting the segmentation process to a broad initialization is grounded in the fact that the "optimal" perception of a moving form can be conceived as the best trade-off achieved by integration of the two processes, motion estimation and segmentation, as provided by the BP algorithm, which needs not be equivalent to either the best segmentation result or motion estimate *per se*.

Note that the two people walking towards the camera and dressing clothes that are similar with respect to the average color are equally labeled, while a different label is assigned to the one walking leftwards behind. Other parts of the scene (ground and building) having null velocity are nevertheless distinguished due to difference in color. It is worth remarking that occlusions are implicitly handled, provided that occluded objects are characterized by different color and/or motion models; clearly, a moving object partially occluded by another object of similar color and motion, will be merged with the latter. On the other hand, the occlusion issue should be more appropriately addressed at the tracking stage.

The next example (Fig. 5) summarizes at a glance results obtained on the whole video by integrating form/motion estimation within the attentive tracking system presented in Ref. 2. In particular, the row in the middle collects results obtained by the Condensation tracking[2]; this relies upon the form/motion estimation and cooperates with the face detection module; the bottom row shows how attention is deployed in terms of FOA setting. Also, experiments have been performed with human observers to compare model-generated gaze-shifts. The subjects involved were 39 students (19 to 26 years old), with normal or corrected-to-normal vision, and naive with respect to the purpose of the experiment. Each subject was sitting in front of the display of the eye-tracking system (ASL 5000) at a distance of 60 cm. Results eye-tracked from five subjects have been preliminary used to

Fig. 5. Top, from left to right: an excerpt of the input sequence. Center: corresponding person and face tracking. Bottom: produced fixation points (FOAs).

train the model, and derive prior probabilities (estimated as fixation frequencies of specific objects, e.g. faces, moving persons, etc.); the other 34 subjects were eye-tracked to compute a "reference" scanpath to include fixations common to many observers (average observer). Results, in terms of overlap between observed and model-generated FOA areas, achieve on the average 54% of successful hits (more than 80% overlap), in the absence of a given task, reaching 90% when a task (e.g. observe people) was given.

7. Final Remarks

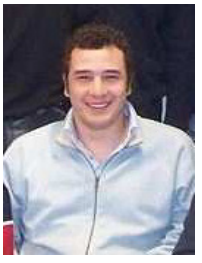
The method proposed relies on Belief Propagation to integrate form and motion information into coherent percepts of moving objects, thus providing a suitable basis for tracking within an attentive system.[2] When compared to the motion segmentation step adopted in Ref. 2, the proposed method not only achieves better results in terms of effectiveness, but also exhibits higher independence from optical flow and segmentation input. This allows to avoid the use of more sophisticated algorithms[2] for correcting optical flow drawbacks and to reduce the number of iterations performed along the DEM segmentation. Further, the discrete label-based representation exploited by BP, makes joint estimation of motion and shape more efficient than the method adopted in Ref. 2. One limitation of the work presented here is the fixed camera setting, and current efforts are spent to adapt the model in order to deal with camera motion, by taking into account feedback as provided by active camera control (e.g. pan, tilt commands). Also, the sequential nature of video analysis is not taken into account here, while it could be embedded within the method in order to exploit at frame $\mathcal{Z}_{t+1}$, estimates of parameters computed on $\mathcal{Z}_t$.[6] On-going research is investigating a possible generalization via nonparametric BP techniques.[13]

References

1. S. Amari, Information geometry of the em and em algorithms for neural networks, *Neur. Networks* **8** (1995) 1379–1408.
2. G. Boccignone, V. Caggiano, G. Di Fiore, A. Marcelli and P. Napoletano, A Bayesian approach to situated vision, *Brain, Vision and Artificial Intelligence 2005*, eds. M de Gregorio, V. Di Maio, M. Frucci, C. Musio, Lecture Notes in Computer Science, Vol. 3704 (2005), pp. 367–376.
3. G. Boccignone, M. Ferraro and P. Napoletano, Diffused expectation maximisation for image segmentation, *Electron. Lett.* **40** (2004) 1107–1108.
4. K. H. Britten, Motion perception: how are moving images segmented? *Current Biol.* **9** (1999) 728–730.
5. W. T. Freeman, E. C. Pasztor and O. T. Carmichael, Learning low-level vision, *Int. J. Comput. Vis.* **40** (2000) 25–47.
6. B. J. Frey and N. Jojic, A comparison of algorithms for inference and learning in probabilistic graphical models, *IEEE Trans. PAMI* **27** (2005) 1392–1416.
7. S. Grossberg, E. Mingolla and C. Pack, A neural model of motion processing and visual navigation by cortical area MST, *Cerebral Cortex* **9** (1999) 878–895.
8. M. M. Hayhoe, D. H. Ballard and D. Bensinger, Task constraints in visual working memory, *Vis. Res.* **38** (1998) 125–137.
9. H. Helmholtz, *Physiological Optics, The Perception of Vision*, Vol. III (Optical Society of America, Rochester, NY, 1925).
10. B. K. P. Horn, *Robot Vision* (MIT Press, Cambridge, MA, 1986).
11. M. Isard and A. Blake, Condensation-conditional density propagation for visual tracking, *Int. J. Comput. Vis.* **29** (1998) 5–28.
12. D. C. Knill, D. Kersten and A. Yuille, A Bayesian formulation of visual perception, in *Perception as Bayesian Inference*, eds. D. C. Knill and W. Richards (Cambridge University Press, 1996).
13. T. S. Lee and D. Mumford, Hierarchical Bayesian inference in the visual cortex, *J. Opt. Soc. Am. A* **20** (2003) 1434–1448.
14. J. Pearl, *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference* (Morgan Kaufmann, 1988).
15. Z. Pylyshyn, Situating vision in the world, *Trends Cogn. Sci.* **4** (2000) 197–207.
16. J. E. Raymond, Attentional modulation of visual motion perception, *Trends Cogn. Sci.* **4** (2000) 42–50.
17. N. Vasconcelos and A. Lippman, Empirical Bayesian motion segmentation, *IEEE Trans. PAMI* **23** (2001) 217–220.
18. Y. Weiss and E. Adelson, A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models, *Proc. IEEE Conf. Comput. Vision Patt. Recognition* (IEEE Computer Society Press, 1996), pp. 321–326.
19. J. S. Yedidia, W. T. Freeman and Y. Weiss, Understanding belief propagation and its generalizations, *Exploring Artificial Intelligence in the New Millennium* (Morgan Kaufmann, San Francisco, CA, 2003), pp. 239–269.
20. R. S. Zemel and T. J. Sejnowski, A model for encoding multiple object motions and self-motion in area MST of primate visual cortex, *J. Neurosci.* **18** (1998) 531–547.

Giuseppe Boccignone received the Laurea degree in theoretical physics from the University of Torino, Italy, in 1985. He worked at Olivetti Corporate Research, Ivrea, as a chief researcher of the Computer Vision and Artificial Intelligence Lab at CRIAI Naples, as a Research Consultant at Research Labs of Bull HN, Milan, Italy. In 1994, he joined as Assistant Professor in the Department of Electrical and Information Engineering, University of Salerno, Italy, where he currently is an Associate Professor of Computer Science. He is a member of the IEEE, IEEE Computer Society and IAPR.

His research interests lie in active vision and theoretical models for computational vision.

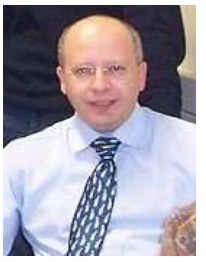


Paolo Napoletano received the Laurea degree in telecommunication engineering from the University of Naples Federico II, Italy, in 2003. He currently is a Ph.D. student in information engineering at the University of Salerno, Italy. He is a student member of the IEEE, IEEE Computer Society.

His research interests lie in active vision, theoretical models for computational vision, medical imaging and image processing.

Gianluca Di Fiore received the Laurea degree in computer engineering from the University of Naples Federico II, Naples, Italy, in 2003. Currently, he is a Research Consultant at CoRiTel Labs, Salerno, Italy.

His research interests lie in video analysis and compression, and software engineering.



Angelo Marcelli received the M.Sc. degree in electronic engineering (cum laude) and the Ph.D. in electronic and computer engineering both from the University of Napoli "Federico II", Italy, in 1983 and 1987, respectively. From 1987 to 1989, he was chief researcher of the Computer Vision and Artificial Intelligence Lab at CRIAI, Napoli, Italy, where he also founded and directed the Italy–Russian Laboratory for Image Analysis and Processing. From 1989 to 1992, he has held a Researcher position at the Department of Computer and System Engineering, School of Engineering, University of Napoli "Federico II". Since 1998, he has been with the Department of Electrical and Information Engineering of the University of Salerno, where he is currently Associate Professor. Dr. Marcelli serves as Area Editor for the *Int. J. Document Analysis and Recognition*. He is a member of the IEEE, IEEE Computer Society, IEEE Systems, Man and Cybernetics Society, IEEE Education Society, IAPR. He is the President-elect of International Graphonomics Society.

His current research interests include handwriting recognition, theory and application of evolutionary algorithms, active vision model and natural computation.

Vittorio Caggiano received the Laurea degree in electronic engineering from the University of Salerno, Italy, in 2004. He currently is a Ph.D. student in computer engineering at the University of Naples "Federico II", Italy.

His research interests lie in active vision, biological vision, medical imaging, image and video databases.