International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 22, No. 5 (2008) 1015–1029

World Scientific
www.worldscientific.com

WHAT THE DRAUGHTSMAN'S HAND TELLS THE DRAUGHTSMAN'S EYE: A SENSORIMOTOR ACCOUNT OF DRAWING

RUBEN COEN CAGLI* and PAOLO CORAGGIO†

Department of Physics
Università di Napoli "Federico II"
via Cinthia, Napoli 80100, Italy
**coen@na.infn.it*
†*pcoraggio@na.infn.it*

PAOLO NAPOLETANO‡ and GIUSEPPE BOCCIGNONE§

Natural Computation Lab
Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica
Universitá di Salerno, via Ponte Melillo 1
84084 Fisciano (SA), Italy
‡*boccig@unisa.it*
§*pnapoletano@unisa.it*

In this paper we address the challenging problem of sensorimotor integration, with reference to eye-hand coordination of an artificial agent engaged in a natural drawing task. Under the assumption that eye-hand coupling influences observed movements, a *motor continuity* hypothesis is exploited to account for how gaze shifts are constrained by hand movements. A Bayesian model of such coupling is presented in the form of a novel Dynamic Bayesian Network, namely an Input–Output Coupled Hidden Markov Model. Simulation results are compared to those obtained by eye-tracked human subjects involved in drawing experiments.

Keywords: Dynamic Bayesian networks; sensorimotor integration; active vision; biologically-inspired robots.

1. Introduction

The problem of eye-hand coordination in performing a given task, is considered[2] a paradigmatic one with respect to the more general question of sensorimotor integration. This, in turn, is reputed to be a crucial issue either for designing situated artificial agents and for the investigation of the underlying cognitive mechanisms in biological agents. Recent approaches to sensorimotor coordination in primates claim that motor preparation has a direct influence on subsequent eye movements,[19] sometimes turning coordination into competition. Complementary, eye movements

come into play in generating motor plans, as suggested by the existence of *look ahead* fixations in many natural tasks.[14]

Differently from the problem of modeling eye movements in purely visual tasks, contending with visuomotor tasks requires a shift of perspective. The main difference in this case is that eye movements should not be treated as entirely independent from movements of other parts of the body. In fact, it is the basic tenet of Active Vision[10] that eye movements depend on the task at hand, and if the task is a sensorimotor one, it is reasonable to expect a dependence on body movements as well.

Our main motivation is to develop a model of the coupling between the processes that give rise to eye and hand movements in a visuomotor task; yet, the model can provide the bare bones of a general framework for the integration of Active Vision and Motor Control.

In Ref. 5, we chose the task of realistic drawing, namely the activity of representing an original scene by means of visible traces on a canvas, trying to render the contours defining objects within the observed scene as faithfully as possible. Since copying an original image on a white canvas requires a quite regular alternation of eye and hand movements,[8,20] this task provides a good example of the "looped" influence between active vision and motor planning/control. A functional model of the sensorimotor processing involved in the drawing behavior was developed on the basis of eye-hand tracking experiments. Eventually, with the aim of providing in a principled way a computational theory (in the sense of Marr[15]) of the underlying processes, we conjectured that such model could be formalized in terms of a novel type of Dynamic Bayesian Network[16] (DBN), which we have denoted the *Input–Output Coupled Hidden Markov Model* (IOCHMM).

In this paper, building on such previous work, we have provided a detailed account of the IOCHMM for modelling eye-hand coordination along drawing, and compare simulation results with eye-hand tracking experiments. Before moving to the following sections, it is worth remarking on two points.

First, the choice of probabilistic graphical models is primarily motivated by the well-known fact that motor and perceptual neural signals are inherently noisy,[12] and that there is a long tradition of statistical modeling of eye movements. Early and seminal attempts were provided by Ellis and Stark, who described the sequence of gaze points in terms of Markov chains,[6,9] and by Rimey who adopted Hidden Markov Models[18] (HMM). Recent models of eye movements in reading[7] have adopted the Input–Output HMM (IOHMM[3]) to account for the fact that variability in gaze sequences reflects not only random fluctuations in the system but also factors such as moment-to-moment changes in the visual input, cognitive influences, and the state of the oculomotor system. The IOCHMM we describe in Sec. 2 treats both eye and hand movements as driven by IOHMMs, but the main point here is that the two are not independent, but rather coupled; the structure of the network reflects our assumption, namely that both eye and hand movements at any given time depend on both eye and hand movements at the previous step.

Second, most computational models of motor control cast the issue of movement planning and execution as an optimization problem,[21] where optimality means minimization or maximization of a scalar function (e.g. jerk, energy, variance) that depends on control signals as well as on the current state of the musculo–skeletal system and environment. Recently, the problems of motor control and optimization have been considered from a stochastic, Bayesian standpoint.[12] Although the question of Bayesian integration of sensorimotor capabilities has been addressed with particular reference to learning,[13] yet, we lack a well-defined framework for integrating an active approach to vision with motor control strategies.

In the present paper we take a step further, and consider the problem of how motor optimization can influence the visual system. To this aim, in Sec. 3, we assume that maximizing the continuity of hand movements represents a constraint for eye movements as well. We test this hypothesis — and its consequences on the observable behavior — by recording human eye-hand movements in a drawing task. Then, in Sec. 4, we detail the implementation of our model; we show that after a learning phase performed on a suitable training set, the system is able to generate both continuous hand strokes and eye movements that are fairly consistent with experimental recordings from human subjects. These results, together with the comparison against models of eye movements that do not consider motor issues, indicate that the proposed model can suitably account for motor constraints and their effects on the visual system.

With respect to previous work in the literature, the IOCHMM proposed here provides a general high level mechanism for the dynamic integration of eye and hand motor plans, and enables the use of information coming from multiple sensory modalities. It also accounts for the task-dependence of eye and hand plans, by learning a sensorimotor mapping that is suitable for the drawing task. To the best of our knowledge the IOCHMM architecture represents a novelty with respect to computational models of drawing, and more generally for sensorimotor coordination.

2. DBN for Eye-Hand Coupling

In a previous work[5] we introduced a functional model for an artificial drawing agent. We argued that the core of the model could be implemented as a DBN, whose inputs are collected from external sensory modules, that feeds premotor information to the subsequent modules responsible for the control of detailed eye and hand motor signals. In the following we develop further and more formally such proposal. In our "minimal" model we introduce two variables that account for sensory inputs, two state variables and two outputs. Specifically, we denote with $\bar{u} = (u^e, u^h)$ the pair of variables representing the visual and hand proprioceptive inputs, respectively, while $\bar{x} = (x^e, x^h)$ denotes the pair of eye and hand (hidden) state variables; eventually, $\bar{y} = (y^e, y^h)$ is the pair of variables accounting for eye and hand output signals (see Sec. 4.1 for details on the state spaces). Further, since sensorimotor coupling evolves in time, say from $t = 1$ to T, we will consider the

discrete time indexed pair sequences $\bar{u}_{1:T}$, $\bar{x}_{1:T}$ and $\bar{y}_{1:T}$. In order to provide a Bayesian generative model of eye-hand coordination, we need to specify the joint pdf, $p(\bar{x}_{1:T}, \bar{y}_{1:T}|\bar{u}_{1:T})$.

To this end, the dynamics of the system presented in Ref. 5 can be summarized as follows: at time t, when visual and hand proprioceptive inputs are fed into the network, the hand state is influenced by the eye state, and motor outputs are generated accordingly; successively, at time $t+1$, the new eye state will be influenced by previous states of both hand and eye, while the hand state depends on its previous state and on the eye current state; thus, on the basis of current visual and hand inputs, new motor outputs are generated. Such behavior can be formalized in the two temporal slices of the DBN shown in Fig. 1.

Note that ideally, the process corresponding to the temporal evolution of the eye plan alone could be considered as an IOHMM; the same holds for the hand plan. However, the most important point here is that the two processes are not independent but rather modeled as coupled chains: in these terms the resulting graphical model unifies the IOHMM DBN and another kind of DBN known in the literature as the Coupled HMM.[16] We call the resulting DBN an *Input–Output Coupled Hidden Markov Model* (IOCHMM, Fig. 1).

By generalizing the time slice snapshot of Fig. 1 to the time interval $[1, T]$ the time dependent joint distribution of state and output variables, conditioned on the input variables can be written as:

$$\begin{aligned} p(\bar{x}_{1:T}, \bar{y}_{1:T}|\, \bar{u}_{1:T}) = {} & p(x_1^e|\, u_1^e, u_1^h) p(y_1^e|\, x_1^e) p(x_1^h|\, u_1^e, u_1^h, x_1^e) p(y_1^h|\, x_1^h) \\ & \cdot \prod_{t=1}^{T-1} [p(x_{t+1}^e|\, u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h) p(y_{t+1}^e|\, x_{t+1}^e) \\ & \cdot p(x_{t+1}^h|\, u_{t+1}^e, u_{t+1}^h, x_{t+1}^e, x_t^h) p(y_{t+1}^h|\, x_{t+1}^h)]. \end{aligned} \tag{1}$$

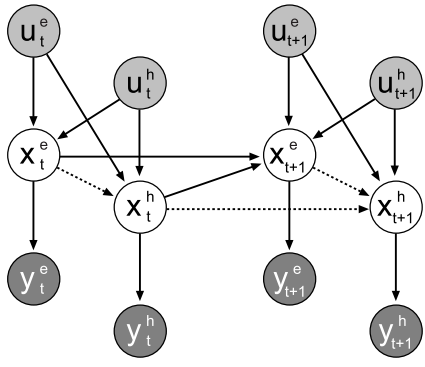


Fig. 1. IOCHMM's for combined eye and hand movements. Dotted connections in the hidden layer highlight the dependence of the hand on the eye, while continuous connections denote the reverse dependence.

The hidden variables x^e, x^h take values in $\{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\}$, and represent the planned eye and hand movement directions, with respect to the current position. The visual input u^e is chosen as the orientation of the attended region, taking values in $\{0, \frac{\pi}{8}, \ldots, \frac{7\pi}{8}\}$. The proprioceptive information u^h, which concerns the direction of the previous hand movement, is encoded using the same values as x^h.

To use the IOCHMM as a control device for an artificial draughtsman, we must contend with three problems: (1) learning the parameters of the model; (2) using the model for inference (i.e. to compute the expected hidden states for each time slice); (3) exploiting inferences to make decisions. For what concerns the inference process, the joint pdf of Eq. (1) can be rewritten as:

$$p(\bar{x}_{1:T}, \bar{y}_{1:T} | \, \bar{u}_{1:T}) = p(\bar{y}_{1:T} | \bar{x}_{1:T}, \bar{u}_{1:T}) p(\bar{x}_{1:T} | \bar{u}_{1:T}) \, . \tag{2}$$

Since output $\bar{y}_{1:T}$ is conditionally independent from input $\bar{u}_{1:T}$ (see Fig. 1), then $p(\bar{y}_{1:T} | \bar{x}_{1:T}, \bar{u}_{1:T}) = p(\bar{y}_{1:T} | \bar{x}_{1:T})$. The latter term describes the mechanism for the generation of eye and hand movements through appropriate pre-motor information, that would be eventually processed by the oculomotor and hand actuator controllers.

Such mechanism, which is actually plagued with noise,[12] can be simplified for the strict purposes of this paper as an ideal, non-noisy mapping, $p(\bar{y}_{1:T} | \bar{x}_{1:T}) = \delta_{\bar{y},\bar{x}}$. Under such assumption, the inference process reduces to the computation of $p(\bar{x}_{1:T} | \bar{u}_{1:T})$. Thus, the expected internal states for each time slice, can be computed as

$$p(\bar{x}_{t+1} | \bar{u}_{1:T}) = \sum_{x^e_{1:T} - x^e_{t+1}} \; \sum_{x^h_{1:T} - x^h_{t+1}} p(\bar{x}_{1:T} | \bar{u}_{1:T}). \tag{3}$$

Note that, according to the network structure, the expected state at time $t+1$ depends only on the input subsequence $\bar{u}_{1:t+1}$; thus, making use of Eq. (1) together with the simplifying assumption discussed above, we can rewrite Eq. (3) as follows:

$$\begin{aligned} &p(\bar{x}_{t+1} | \bar{u}_{1:t+1}) \\ &\quad = \sum_{\bar{x}_{1:t}} p(x^e_{t+1} | u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t) p(x^h_{t+1} | u^e_{t+1}, u^h_{t+1}, x^e_{t+1}, x^h_t) p(\bar{x}_{1:t} | \bar{u}_{1:t}) \end{aligned} \tag{4}$$

which represents a particular case of recursive Bayesian filtering.[4]

The explicit computation of Eq. (4) requires knowledge of the network's dynamics, namely the state transition probability distributions, which can be gained through the learning stage. Following a classical approach, this consists in evaluating the parameters by maximizing the log-likelihood $\log p(\bar{x}_{1:T} \, | \, \bar{u}_{1:T})$.

Recalling that $p(\bar{y}_t \, | \, \bar{x}_t) = \delta_{\bar{y},\bar{x}}$, when we provide the DBN with an appropriate data set, i.e. a set of input–output sequences $\{\bar{u}_{1:T}, \bar{y}_{1:T}\}$, we can set $\bar{x}_{1:T} = \bar{y}_{1:T}$, and by considering Eq. (1), we can write the likelihood function in matrix form (see Appendix A for details):

$$\begin{aligned} L_c = {} & x_1^{e\perp} \log(\Phi^e) u_1^e u_1^h + x_1^{h\perp} \log(\Phi^h) u_1^e u_h^h x_1^e \\ & + x_{t+1}^{e\perp} \log(\Gamma^e) u_{t+1}^e u_{t+1}^h x_t^e x_t^h + x_{t+1}^{h\perp} \log(\Gamma^h) u_{t+1}^e u_{t+1}^h x_{t+1}^e x_t^h \end{aligned} \tag{5}$$

where $\perp$ denotes the transpose, Φ, Γ denote respectively the *input state* and *transition* probability distributions.

In this work, we make no assumption on the parametric functional form of such pdf's, but rather consider them as Conditional Probability Tables (CPT), i.e. matrices whose entries are the parameters that should be learned. This is done by adapting the Baum–Welch[4] algorithm to our specific DBN.

Eventually, to use the DBN as a control system, we apply a decision rule to inference and learning results. According to Bayesian Decision theory, different choices can be made for the decision rule; in the simulations presented in this work we used the Maximum a Posteriori (MAP) criterion, which consists in selecting the pair (x^e_{t+1}, x^h_{t+1}), such that

$$(x^{e\star}_{t+1}, x^{h\star}_{t+1}) = \arg\max[p(x^e_{t+1}, x^h_{t+1}|\bar{u}_{1:t+1})]. \tag{6}$$

3. Gaze Analysis in a Drawing Task

Our experiments have addressed a drawing task, where the subjects were asked to draw a copy of an original image. Previous behavioral analysis of draughtsmen at work[20] have revealed the existence of a regular execution cycle, where two main phases can be distinguished. During one phase, which corresponds to either the selection of what to draw next or the evaluation of the emerging result, the hand is not drawing, and globally distributed eye movements can be observed; the other phase is the one during which drawing hand strokes are observed, and the gaze is moved orderly and locally on the original image.

Elsewhere we have considered the overall role of the two phases[5]; here, we are concerned with characterizing fixations on the original image during the drawing phase, and understanding how eye and hand movements are related along this phase.

3.1. *Experimental setup, subjects and instructions*

Eye scan records were obtained from 25 subjects, aged between 18 and 33, without previous specific experience in drawing. Subjects were presented with a rectangular, vertical tablet 40 cm $\times$ 30 cm. As shown in Fig. 2, original images were displayed in the left half of the tablet, while the right half was covered by a white sheet. The original images represented simple contours drawn by hand with a black pencil on white paper with an area of approximately 15 cm $\times$ 15 cm.

One image per trial was shown, and the subjects were instructed to copy its contours as faithfully as possible, drawing on the right-hand sheet. These instructions did not give constraints on the execution time. Each subject carried out six trials, one per image.

The subject's left eye movements were recorded with a remote eye tracker (ASL Model 504) with the aid of a magnetic head tracker, with the eye position sampled at the rate of 60 Hz. The instrument can integrate eye and head data in real time and can deliver a record with an accuracy of less than 1°. Here we present the

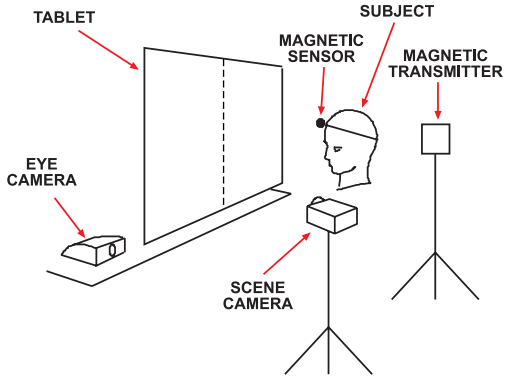


Fig. 2. Experimental setup for eye tracking recordings during the drawing task. The subject sits in front of a vertical Tablet. In the left half of the Tablet hand-drawn images are displayed, and the subject is instructed to copy the images on the right half. The eye tracker integrates data from the Eye Camera and the Magnetic Sensor and Transmitter; eye position is then superimposed on the Scene Camera video stream, which takes the approximate subjective point of view.

analysis of data corresponding to the left hemifield (the original image). In the following we refer to the scanpath as the sequence of saccades and fixations on the scene, minus saccades and fixations on the right hemifield: thus a sequence *fixation on the left–saccade–fixation(s) on the right–saccade–fixation on the left* becomes *fixation on the left–saccade–fixation on the left.*

The analysis of the recorded eye data is performed under the following hypothesis:

Motor Continuity. *The sequence of fixations on the original scene is constrained to maximize graphical continuity of tracing hand movements.*

In order to explore the correctness and the implications of this assumption, we analyze the scanpaths recorded in a trial where the original image is a single line shape. Fixations are found by means of the standard *dispersion* algorithm, with thresholds set to 2° and 100 msec. Figure 3 depicts the cumulative plot of fixations, and the corresponding hand position, at four subsequent stages. The times of the snapshots correspond to the moments during which the following sequence is observed: *hand stops–fixation(s) on the left–saccade–fixation(s) on the right–hand–moves.* We interpret the points where the hand stops as keypoints, at which the hand's action needs to be reprogrammed and thus fixations on the original image become necessary.

A qualitative inspection of Fig. 3 shows a general tendency of the gaze to move orderly along the image contour, as confirmed by the scanpaths of four different subjects, plotted in Fig. 4; furthermore, all of our subjects used graphically continuous hand strokes. This evidence suggests that the strategy that humans adopt in the drawing task, to facilitate graphical continuity of hand movements, is to move the gaze according to a *coarse grained edge-following* along the contours of the original image.

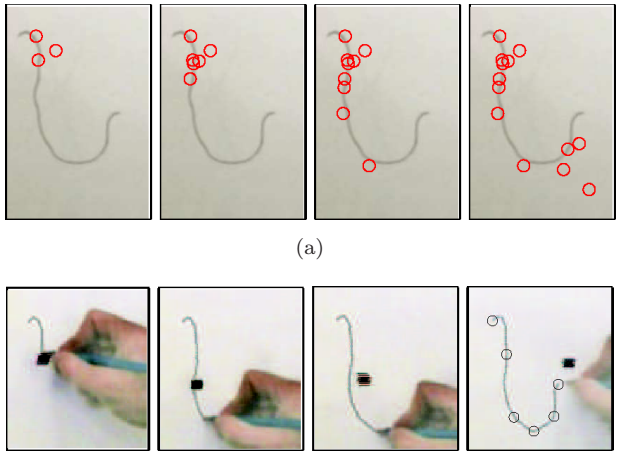


Fig. 3. The performance of subject **AP** in the drawing task. (a) Cumulative fixations on the original image, represented by circles. (b) Manual execution. The solid square denotes the gazepoint, while circles denote the endpoints of each trajectory segment.

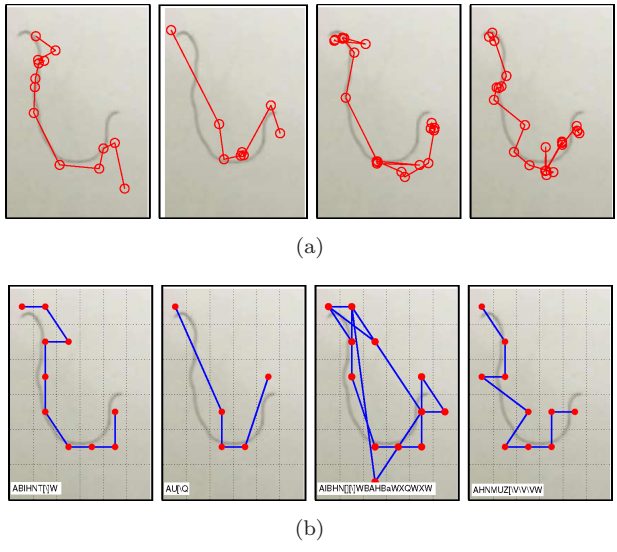


Fig. 4. (From left to right) The top row shows the scanpaths recorded from subjects **AP**, **AS**, **AC**, **MJG**; the bottom row, their clustered version.

Thus, we define a procedure[17] to evaluate in a quantitative manner the similarity of the recorded scanpaths to the *coarse grained edge-following*; the same procedure can be used then to make a comparison with the scanpaths generated by our DBN as well as other computational models.

As a first step we superimpose an ordered grid on the original image, and then we cluster together all subsequent fixations that fall within a single cell as one

single event. At the end of this procedure, instead of the scanpath we have an ordered sequence of events, each one belonging to a single cell of the grid, as shown in Fig. 4(b). Then, each cell is labeled with a symbol (an ASCII character in the interval "A" to "e"), so that each sequence of events is coded as a string; this enables to compare through a unique string similarity algorithm either strings produced by two algorithms, or two human subjects or an algorithm and a subject. The final similarity value can be normalized on the basis of the string length.

The string similarity index can be defined through an optimization algorithm, with a cost unit based on three different operations: deleting, inserting and substituting. By sequentially processing the first string to obtain the second string, we get the similarity index as the minimum total cost (known as Levensthein distance). The numerical results are plotted in Fig. 7, and discussed in Sec. 4.2.

4. Simulation Results and Comparison with Experimental Data

4.1. *Implementation details and simulation results in the drawing task*

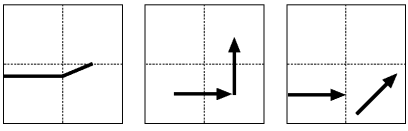
For the simulations presented here, discrete state spaces were chosen for all the variables. The visual input represents the dominant orientation of the fixated image patch, i.e. $u^e \in \{0, \frac{\pi}{8}, \ldots, \frac{7\pi}{8}\}$. The proprioceptive input provides an estimate of the previous hand movement direction $u^h \in \{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\}$; hidden and output variables take values in the same set as u^h, and are interpreted respectively as the proposed direction of the next saccade (x^e, y^e) and of the next hand movement (x^h, y^h).

The training examples we use are sequences that reflect the experimental observations on eye-tracked human subjects: hand movements are graphically continuous and correspondingly the scanpath is a coarse-grained edge-following along the contours of the original image. An example from the training set, whose values are reported in Table 4, is illustrated in Fig. 5(a). As a result of the learning stage followed by the decision step, we obtain a sensory motor map that encodes the eye and hand directions x_t^e and x_t^h for each given input pair. In Fig. 5(b), we show an instance of this map in the case of $x_{t-1}^e = 0$.

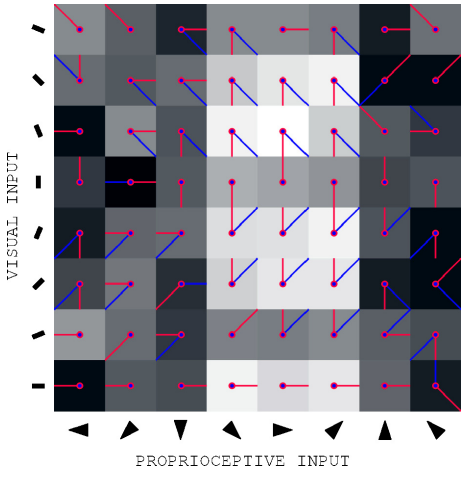
After training the DBN as described above, we have run it on a binarized version of the original image shown to the subjects [Fig. 3(a)]. The resulting time sequences of eye and hand plans $\bar{y}^e, \bar{y}^h$ are provided in the two top rows of Fig. 6(a). The corresponding scanpath is given in Fig. 6(b), and it can be directly compared to the human eye movement recordings shown in Fig. 4. Figure 6(c) shows the trajectories planned according to the DBN outputs, with the endpoints evidenced by blue circles; these trajectories are computed as splines passing through the points corresponding to the position of each eye fixation, with a slope defined by the associated hand plans.

It is worth noting that a pure bottom-up, uncoupled scanpath generation would provide a very different result. This can be easily shown, for instance, by feeding the salient points to a winner-takes-all network combined with the inhibition of

	$t=1$	$t=2$
u_t^e	0	$\frac{\pi}{8}$
u_t^h	π	π
y_t^e	0	$\frac{\pi}{2}$
y_t^h	0	$\frac{\pi}{4}$

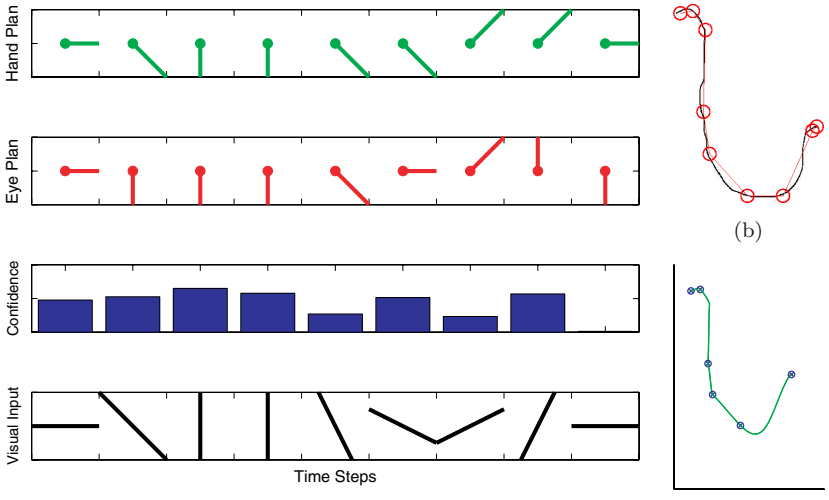


(a)



(b)

Fig. 5. (a) An input/output example: the bottom row depicts the visual input (left) and the eye (middle) and hand (right) outputs corresponding to the sequence given in the table above. (b) A graphical representation of the eye-hand policy obtained by applying Bayesian Decision Theory to the trained DBN, in the specific case that $x_{t-1}^e = 0$: light and dark arrows denote the direction of the eye and hand plan respectively, for each input pair. The level of confidence has been coded as a gray-level (white = 100%, black = 0%).



(a)

(b)

(c)

Fig. 6. On the left, Fig. 6(a), the simulated discrete-time evolution. From bottom to top: the bottom row represents the sequence of visual inputs, namely the orientation of the foveated image region; the second row shows the confidence level assigned to the chosen eye–hand plan; the third and fourth rows show the DBN outputs, namely the eye and hand movement plans, respectively. On the right, Fig. 6(b) shows the generated scanpath, while Fig. 6(c), the planned hand trajectory where circles denote the starting and ending points of each trajectory portion. Both eye and hand movements start from the upper left corner.

return[11] in order to obtain the bottom-up fixation sequence; an evaluation of how a bottom-up scanpath differs from scanpaths either generated by our approach or recorded via eye-tracking, is presented in Sec. 4.2.

4.2. *Comparison with experimental data*

The comparison between each recorded scanpath (precisely 11 subjects) and four different simulated scanpaths (Random, Saliency, Edge Following, DBN) is reported in Fig. 7. Such comparison is obtained by measuring the similarity between simulated and experimental scanpath by performing the well-known Levensthein string similarity algorithm.[17] The simulated scanpaths are obtained as follows:

(1) *Random*: 10,000 random strings are generated and compared with each experimental scanpath. Each random string is formed considering only the cells containing the pattern, and their adjiacent cells.
(2) *Saliency*: Fixations are generated by using a bottom-up, saliency-based algorithm.[11]
(3) *Edge Following*: Obtained through a perfect edge following of the pattern.
(4) *DBN*: Fixations are generated by the proposed DBN.

Note that with respect to the *Random* case, we considered, for each subject, the mean of the resulting 10,000 string similarity measures.

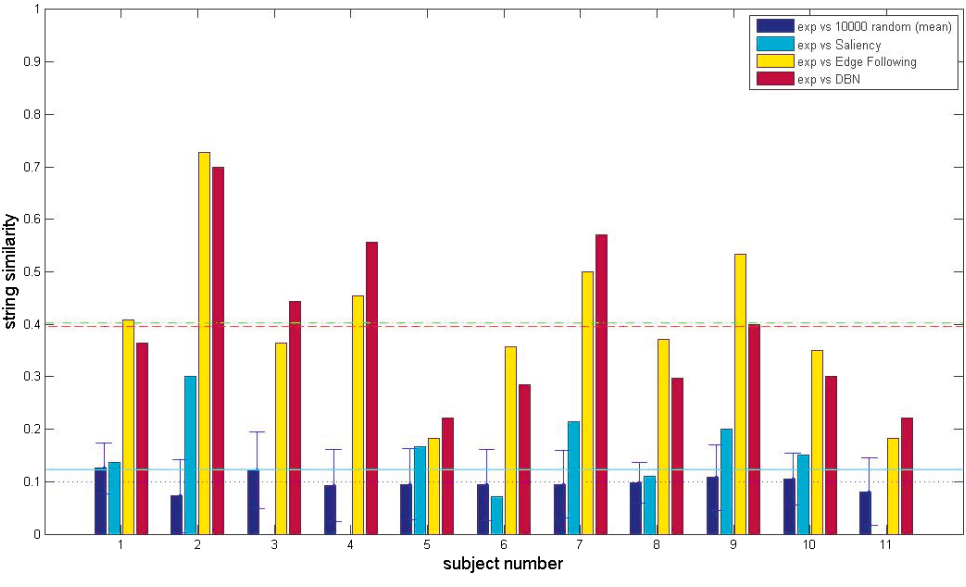


Fig. 7. For each subject, the mean similarity of the observed scanpath to 10,000 random scanpaths (dark gray with error bar); to a preattentive scanpath à la Itti (intermediate gray); to a perfect coarse-grained edge-following (light gray); and to the scanpath simulated by the DBN (dark gray without error bars).

Table 1. Comparison with experimental data: mean and standard deviation.

Exp vs Random	Exp vs Saliency	Exp vs Edge Following	Exp vs DBN
0.098 ± 0.015	0.1227 ± 0.0097	0.40 ± 0.15	0.39 ± 0.16

The mean and standard deviations of the similarity measures between the simulated and the experimentally recorded scanpaths are reported in Table 1.

The results show that the random scanpaths are responsible for the lowest string similarity value; a higher similarity is demonstrated by the perfect edge following and the scanpaths generated by the DBN. It is worth noting that the Saliency performance (bottom-up fixations) is quite similar to the Random one.

5. Final Remarks

In this paper we have presented a computational model of realistic drawing in order to investigate the issue of visuomotor coordination. This issue indeed poses a challenging question at the leading edge of current research in neuroscience, Active Vision, Artificial Intelligence and Robotics: what strategies are to be adopted by any agent situated in the world to coordinate vision and action in order to succeed in a task of interest?

The strategies adopted to coordinate sensorimotor processes of eye and hand movement generation, during the drawing task, are inferred by a Dynamic Bayesian Network, namely an Input-Output Coupled Hidden Markov Model (IOCHMM). To the best of our knowledge such model has never been discussed before in the sensorimotor coordination literature.

Simulations of the IOCHMM behavior have been compared to those obtained by eye-tracked human subjects involved in drawing experiments. Experiments showed that both the simulated trajectory and the gazing points have patterns quite similar to those obtained by human draughtsmen.

As future work we prefigure to remove the assumption of dealing with an ideal motor output so as to extend the simulation to a realistic setting by using a 7-DOF anthropomorphic manipulator together with an active Pan/Tilt/Zoom camera for perfoming actual drawing.

Appendix A. Learning in the Discrete State Space

The likelihood function $L_c \doteq \log p(\bar{y}_{1:T}, \bar{x}_{1:T} | \bar{u}_{1:T})$ can be derived from Eq. (1) while assuming ideal motor output condition $p(\bar{y}_{1:T} | \bar{x}_{1:T}) = \delta_{\bar{y},\bar{x}}$:

$$L_c = \log p(x_1^e | u_1^e, u_1^h) + \log p(x_1^h | u_1^e, u_1^h, x_1^e) + \sum_{t=1}^{T-1} \log p(x_{t+1}^e | u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h)$$
$$+ \sum_{t=1}^{T-1} \log p(x_{t+1}^h | u_{t+1}^e, u_{t+1}^h, x_{t+1}^e, x_t^h). \tag{A.1}$$

Define with M, N, L, K the dimensionality of the hand and eye movement hidden and input space respectively. We encode discrete variables in the canonical basis,[4] e.g. if $x^e \in \{x^{e,1} \dots x^{e,M}\}$, then we have $x^{e,1} = (1, 0 \dots 0)$ and so on. With this choice, the eye-related pdf's in the log-likelihood become:

$$p(x_1^e | u_1^e, u_1^h) = \prod_{i=1}^{M} \prod_{j=1}^{L} \prod_{p=1}^{K} (\Phi_{ijp}^e)^{x_{1,i}^e u_{1,j}^e u_{1,p}^h}$$

$$p(x_{t+1}^e | u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h) = \prod_{i=1}^{M} \prod_{j=1}^{L} \prod_{p=1}^{K} \prod_{r=1}^{M} \prod_{s=1}^{N} (\Gamma_{ijprs}^e)^{x_{t+1,i}^e u_{t+1,j}^e u_{t+1,p}^h x_{t,r}^e x_{t,s}^h}$$

where Φ, Γ denote the *input state* and *transition* probability distribution, respectively. Similar equations hold for $p(x_1^h | u_1^e, u_1^h, x_1^e)$ and $p(x_{t+1}^h | u_{t+1}^e, u_{t+1}^h, x_{t+1}^e, x_t^h)$, and the log-likelihood can be recast in matrix form as:

$$\begin{aligned} L_c = {} & x_1^{e\perp} \log(\Phi^e) u_1^e u_1^h + x_1^{h\perp} \log(\Phi^h) u_1^e u_h^h x_1^e \\ & + x_{t+1}^{e\perp} \log(\Gamma^e) u_{t+1}^e u_{t+1}^h x_t^e x_t^h + x_{t+1}^{h\perp} \log(\Gamma^h) u_{t+1}^e u_{t+1}^h x_{t+1}^e x_t^h \end{aligned} \tag{A.2}$$

where $\perp$ denotes the transpose. The maximization step of the Baum–Welch algorithm is done by taking the derivatives of Eq. (A.2) with respect to the parameters, set to zero and solve under the sum-to-one constraint. The solutions give us the parameters in terms of the expected sufficient statistic:

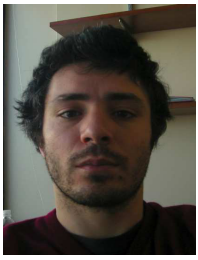
$$\begin{cases} \gamma_{t,i}^e \doteq \langle X_{t,i}^e \rangle \\ \gamma_{t,i}^h \doteq \langle X_{t,i}^h \rangle \\ \gamma_{t,i}^{eh} \doteq \langle X_{t,i}^e, X_{t,i}^h \rangle \\ \xi_{t,ij}^{e,h} \doteq \langle X_{t,i}^e, X_{t-1,j}^h \rangle \\ \xi_{t,ij}^{e,eh} \doteq \langle X_{t,i}^e, X_{t-1,j}^e, X_{t-1,j}^h \rangle \\ \xi_{t,ij}^{eh,h} \doteq \langle X_{t,i}^e, X_{t,i}^h, X_{t-1,j}^h \rangle \end{cases} \Longrightarrow \begin{cases} \Phi_{ijk}^e = \gamma_{1,i}^e u_{1,j}^e u_{1,k}^h \\ \Phi_{ijkl}^h = \gamma_{1,il}^{eh} u_{1,j}^e u_{1,k}^h \\ T_{ijklm}^e = \frac{\sum_{t=2}^T \xi_{t,ilm}^{e,eh} u_{t,j}^e u_{t,k}^h}{\sum_{t=2}^T \gamma_{t,lm}^{eh} u_{t,j}^e u_{t,k}^h}. \\ T_{ijklm}^h = \frac{\sum_{t=2}^T \xi_{t,ilm}^{eh,h} u_{t,j}^e u_{t,k}^h}{\sum_{t=2}^T \xi_{t,lm}^{e,h} u_{t,j}^e u_{t,k}^h} \end{cases} \tag{A.3}$$

Eventually, the γ and ξ terms are found in the E-step via the forward–backward inference algorithm.[4]

References

1. H. Attias, Planning by probabilistic inference, *Proc. 9th Int. Conf. Artificial Intelligence and Statistics* (2003).
2. D. H. Ballard, M. M. Hayhoe, F. Li and S.D. Whitehead, Hand-eye coordination during sequential tasks, *Phil. Trans. R. Soc. Lond. B* **337** (1992) 331–339.
3. Y. Bengio and P. Frasconi, Input–output HMM's for sequence processing, *IEEE Trans. Neural Networks* **7** (1995) 1231–1249.
4. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, Berlin, 2007).
5. R. Coen Cagli, P. Coraggio, G. Boccignone and P. Napoletano, The Bayesian draughtsman: a model for visuomotor coordination in drawing, *Advances in Brain Vision and Artificial Intelligence*, Lecture Notes in Computer Science, Vol. 4729 (Springer, 2007), pp. 161–170.

6. S. R. Ellis and J. D. Smith, Patterns of statistical dependency in visual scanning, *Eye Movements and Human Information Processing*, eds. R. Groner, G. W. McConkie and C. Menz (Elsevier, Amsterdam, 1985) pp. 221–238.
7. G. Feng, Eye movements as time-series random variables: a stochastic model of eye movement control in reading, *Cogn. Syst. Res.* **7** (2006) 7095.
8. E. Gowen and R. C. Miall, Eye-hand interactions in tracing and drawing tasks, *Hum. Mov. Sci.* **25** (2006) 568–585.
9. S. S. Hacisalihzade, L. W. Stark and J. S. Allen, Visual perception and sequences of eye movement fixations: a stochastic modeling approach, *IEEE Trans. Syst. Man Cyber.* **22** (1992) 474–481.
10. M. M. Hayhoe and D. H. Ballard, Eye movements in natural behavior, *Trends Cogn. Sci.* **9** (2000) 188.
11. L. Itti and C. Koch, Computational modelling of visual attention, *Nat. Rev. Neurosci.* **2** (2001) 194–203.
12. K. P. Kording and D. M. Wolpert, Bayesian decision theory in sensorimotor control, *Trends Cogn. Sci.* **10** (2006).
13. K. P. Kording and D. M. Wolpert, Bayesian integration in sensorimotor learning, *Nature* **427** (2004) 244–247.
14. M. Land, N. Mennie and J. Rusted, Eye movements and the roles of vision in activities of daily living: making a cup of tea, *Perception* **28** (1999) 1311–1328.
15. D. Marr, *Vision: A Computational Approach* (Freeman and Co., San Francisco, 1982).
16. K. Murphy, Dynamic Bayesian networks: representation, inference and learning, Ph.D. thesis, Berkeley, University of California (2002).
17. C. M. Privitera and L.W. Stark, Algorithms for defining visual regions-of-interest: comparison with eye fixations, *IEEE Trans. Patt. Anal. Mach. Int.* **22** (2000) 970.
18. R. D. Rimey and C. M. Brown, Controlling eye movements with hidden Markov models, *Int. J. Comput. Vis.* **7** (1991) 47.
19. B. Sheliga, L. Craighero, L. Riggio and G. Rizzolatti, Effects of spatial attention on directional manual and ocular responses, *Exp. Brain. Res.* **114** (1997) 339.
20. J. Tchalenko, R. Dempere-Marco, X. P. Hu and G. Z. Yang, Eye movement and voluntary control in portrait drawing, *The Minds Eye: Cognitive and Applied Aspects of Eye Movement Research* (Elsevier, Amsterdam, 2003), Chap. 33.
21. D. M. Wolpert and Z. Ghahramani, Computational principles of movement neuroscience, *Nat. Neurosci.* **3** (2000) 1212–1217.



Ruben Coen Cagli received the laurea degree cum laude in theoretical physics in 2004 and the Ph.D. degree in 2007 from the University of Napoli (Italy). In January 2008 he joined the Department of Neuroscience at the Albert Einstein College of Medicine of Yeshiva University, New York, where he is currently a Research Associate.

His main research interests are in active vision and visual attention, motor control, image statistics, and the visual arts.

Paolo Coraggio received the laurea degree cum laude in theoretical physics from the University of Naples Federico II (Italy) in 2003, and the Ph. D. degree in computational and information sciences from the University of Naples Federico II in 2007.

He is currently working on robotics, collaborating with the Department of Physical Sciences of the University Federico II, and the design and implementation of algorithms for Gravitational Waves revelation (SCoPE – INFN project).

Paolo Napoletano received the laurea degree in telecommunication engineering from the University of Naples Federico II, Naples, Italy, in 2003, and the Ph.D. degree in information engineering from the University of Salerno, Italy, in 2007. He currently holds a post-doc position at the Natural Computation Lab, Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno.

His current research interests lie in active vision, Bayesian models for computational vision and ontology building. He is Member of the IEEE Computer Society, and GIRPR (the Italian chapter of IAPR).

Giuseppe Boccignone received the laurea degree in theoretical physics from the University of Turin (Italy) in 1985. In 1986, he joined Olivetti Corporate Research, Ivrea, Italy. From 1990 to 1992, he served as a Chief Researcher of the Computer Vision Lab at CRIAI, Naples, Italy. From 1992 to 1994, he held a Research Consultant position at Research Labs of Bull HN, Milan, Italy, leading projects on biomedical imaging. In 1994, he joined the Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, University of Salerno, Salerno, Italy, where he is currently an Associate Professor of Computer Science. He has been active in the field of computer vision, image processing, and pattern recognition.

His current research interests lie in active vision, Bayesian models for computational vision, cognitive science and medical imaging. He is a Member of the IEEE Computer Society, and GIRPR (the Italian chapter of IAPR).