Vision Research 49 (2009) 810–818

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Visuomotor characterization of eye movements in a drawing task

Ruben Coen-Cagli [a,*], Paolo Coraggio [e], Paolo Napoletano [b], Odelia Schwartz [a], Mario Ferraro [c], Giuseppe Boccignone [d]

[a] Department of Neuroscience, Albert Einstein College of Medicine of Yeshiva University, 1410 Pelham Pkwy S., Rm 921, Bronx, New York 10461, USA
[b] Natural Computation Lab, Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica, Università di Salerno, via Melillo 1, 84084 Fisciano, SA, Italy
[c] Dipartimento di Fisica Sperimentale, Università di Torino, via Giuria 1, 10125 Torino, Italy
[d] Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, via Comelico 39/41, 20135 Milano, Italy
[e] DISI, Università di Genova, via Dodecaneso 35, 16146 Genova, Italy

ARTICLE INFO

Article history:
Received 11 October 2008
Received in revised form 19 February 2009

Keywords:
Visuomotor
Drawing
Bayesian

ABSTRACT

Understanding visuomotor coordination requires the study of tasks that engage mechanisms for the integration of visual and motor information; in this paper we choose a paradigmatic yet little studied example of such a task, namely realistic drawing. On the one hand, our data indicate that the motor task has little influence on which regions of the image are overall most likely to be fixated: salient features are fixated most often. Viceversa, the effect of motor constraints is revealed in the temporal aspect of the scanpaths: (1) subjects direct their gaze to an object mostly when they are acting upon (drawing) it; and (2) in support of graphically continuous hand movements, scanpaths resemble edge-following patterns along image contours. For a better understanding of such properties, a computational model is proposed in the form of a novel kind of Dynamic Bayesian Network, and simulation results are compared with human eye–hand data.

1. Introduction

In this paper we address the challenging problem of characterizing the visuomotor behavior of an agent engaged in a natural drawing task. Making a realistic portrait of a visual scene requires accurate attentional control of fixations, and imposes rigid constraints on eye–hand coordination, and as such is a paradigmatic example of a visuomotor task.

The issue of eye–hand coordination in drawing has been addressed by a number of authors (Cohen, 2005; Gowen & Miall, 2006; Land, 2006; Tchalenko, 2007; Tchalenko & Miall, 2008; Viviani & Flash, 1995). On a global behavioral level, a consistent feature of human drawing strategies is the following execution cycle: *fixation on the original image – saccade – fixation(s) on the canvas – saccade – fixation on the original image* (Tchalenko, 2007). The specific kind of visual processing that takes place when fixating on the original is still unclear in general, but two main positions have been outlined: (i) fixations on the original serve to encode image features to visual working memory, and such mental image is later recalled and converted to a motor plan (Tchalenko, Dempere-Marco, Hu, & Yang, 2003); (ii) the visuomotor mapping from image features to hand motor activity takes place during fixations on the original image, without the need to invoke working memory (Coen-Cagli, Coraggio, Boccignone, & Napoletano, 2007; Tchalenko & Miall, 2008).

This last view is consistent with results from the eye tracking experiments presented in this paper, which explore how the scanpaths observed in human subjects involved in drawing, differ from those obtained in free viewing control experiments. The results discussed in Section 2.2 can be summarized by the observation that not only are eye movements in drawing strongly biased by the task, but a precise dependency can be established between the peculiar motor constraints and the recorded scanpaths (this has been reported for other motor tasks with low memory load and reduced stimulus complexity; see e.g. Aivar, Hayhoe, Chizk, & Mruczek (2005), Brouwer & Knill (2007), & Stritzke & Trommershäuser (2007)). In particular, we show that the observed eye movements represent a precise strategy to help satisfy the hand motor constraint of graphical continuity.

In order to make this notion more precise, we also address the issue of outlining a computational model of eye movements based on dynamical eye–hand coupling (currently, to the best of our knowledge, there exist no theoretical model of the processes underlying drawing). A model at the computational level (Marr, 1982) should account for what is the goal of the computation, and what is the logic of the strategy by which it can be carried out, thus abstracting from algorithmic and physical realization details. To this end, the Bayesian approach is exploited as a sound framework (see Carter, Tenenbaum, & Yuille (2006) for an in-depth

* Corresponding author. Fax: +1 718 430 8821.
E-mail address: rcagli@aecom.yu.edu (R. Coen-Cagli).

0042-6989/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2009.02.016

discussion and review). Whether the model proposed here is amenable to neural realization is outside the scope of this paper; however, for a discussion of plausibility of Bayesian computations one can refer to Carter et al. (2006) and Lee and Mumford (2003). The model is confronted with experimental results obtained by human observers in Section 2.3, and an overall discussion is provided in Section 3.

2. Eye movements in drawing and free viewing subjects

2.1. Experimental methods

2.1.1. Participants

Two experimental sessions were realized, during which eye movements were recorded and hand movements monitored. All subjects had normal or corrected to normal vision; none of them had specific previous training in drawing or painting. Subjects consisted of undergraduates, graduate students and research fellows from the University of Salerno, from a range of academic disciplines. The experiments were undertaken with the understanding and written consent of each subject.

For the first session 29 human subjects, five of which were left-handed, participated in the drawing task. The subjects were asked to perform an accurate drawing of an original image; these instructions did not pose any constraints on the execution time. The second session involved six subjects, who were asked to watch the same images, without a specific task (free viewing).

2.1.2. Displays and protocol

The experimental setup for the drawing task is shown in Fig. 1a. Subjects were presented with a rectangular, vertical tablet 30 cm $\times$ 40 cm, viewed binocularly from a distance ranging from 35 cm to 45 cm depending on the subject's arm length. In the left half of the tablet, the original images were displayed, while the right half was initially covered by a white sheet. The original images (Fig. 1 shows the three discussed here), represent simple contours drawn by hand with a black pencil on white paper, that occupy an area of approximately 15 cm $\times$ 15 cm, subtending a visual field of 10–12 deg in both the horizontal and vertical direction. 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 pose any constraints on the execution time. Each subject carried out one trial per image, always in the same order. On completion of a trial, the original and copied images were manually removed by the experimenters, and replaced respectively by a new original and white paper. This allowed for a pause of about 5 s between trials. The execution time varied across individuals and across trials; on average, the time to complete a single trial was 17 ± 9 s.

For the free viewing experiment, original images were digitized with a scanner, and displayed on a 19-inch computer screen for 10 s each, interleaved with a 5 s blank screen. Screen resolution and viewing distance were chosen in such a way that the images subtended a similar visual angle as in the drawing trials.

2.1.3. Eye data acquisition

The subject's left eye movements were recorded with a remote eye tracker (ASL 5000 series) with the aid of a magnetic head tracker (Ascension *Flock of Birds*), 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 deg in optimal light conditions. Fixations were detected from raw data with the standard *dispersion* algorithm, with threshold set to 2.0 deg of visual angle and minimum fixation duration of 100 ms.

2.2. Data analysis

At present, only very few eye tracking studies on drawing humans have been conducted (Coen-Cagli et al., 2007; Gowen & Miall, 2006; Tchalenko, 2007; Tchalenko & Miall, 2008), and no standard measures have been defined for this task. The analyzes presented here are aimed at highlighting the regularities in the observed eye movement during the drawing task, as compared to free viewing. We focus mainly on eye movements related to the segmentation of the image in separate objects, and to the visuomotor mapping from visual features to hand movements, because only fixations on the original image are relevant to these sub-tasks, which allows for a direct comparison with purely visual tasks. Therefore in the drawing task we analyzed only fixations on the left hemifield (*i.e.* the original image).

2.2.1. Object commitment

First we consider the image displayed in Fig. 1b, which is composed by two closed contours that are *spatially separated*. We find that a peculiar feature of the drawing behavior is that the gaze does not move back and forth among different objects, but proceeds sequentially, and most fixations on an object are executed within a time interval in which no fixations occur on other objects. This finding illustrates how the motor task influences object-based visual attention, showing that gaze is directed to an object only when it becomes relevant to the task, namely during the time that it is being copied.

To quantitate this effect, we proceeded as follows. From qualitative analysis of the data collected in drawing, it was clear that all subjects started drawing the second object only after completion of the first one, irrespective of which of the two objects was chosen as the first. Therefore, we were able to define, for each subject, two time intervals, τ_1 and τ_2, corresponding to the two drawing phases; these were found by inspection of the video data. Then we defined two rectangular Regions Of Interest (ROI), R_1, R_2, each one containing one of the two objects; Fig. 2a shows the fixations executed by one subject in each of the two time intervals. Fixations were then classified in each time interval, as falling in R_1, R_2 or outside (*OFF*).

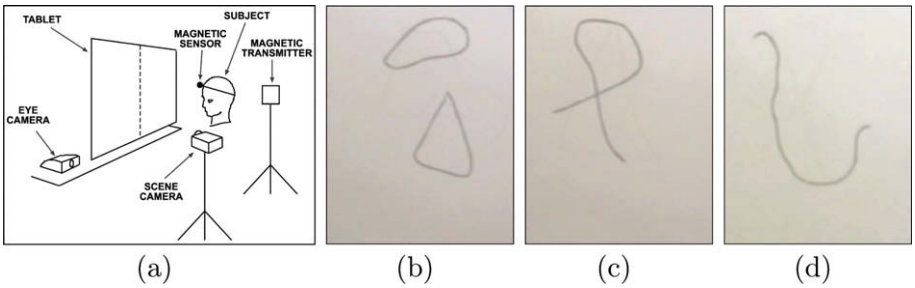

Fig. 1. (a) Layout of the experimental setup. (b–d) The original images adopted in the experiments.

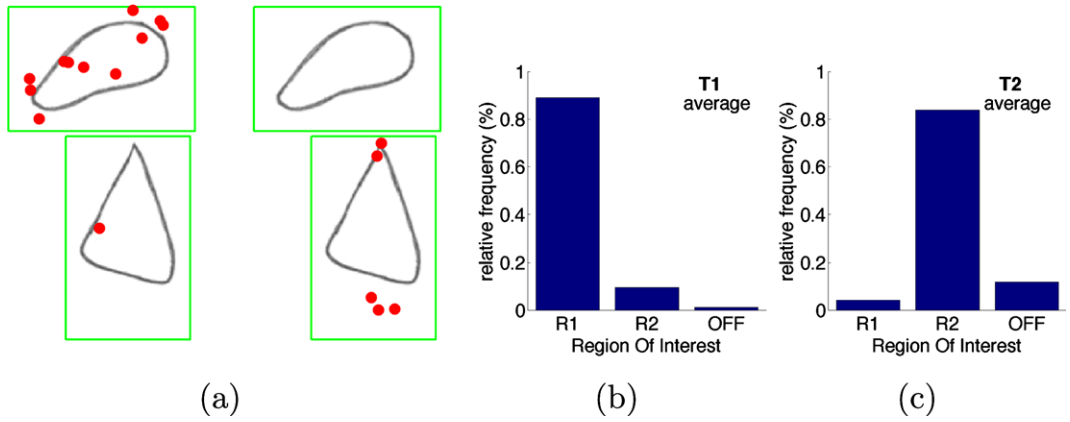

Fig. 2. (a) The cumulative fixations executed by one subject in the drawing task, trial 1, while drawing the first object (left) and the second object (right). Regions of Interest are highlighted by the rectangles. (b and c) The distribution of fixations over the ROI's (R_1, R_2, OFF), averaged across nine subjects, during the first and the second time intervals, respectively.

Fig. 2b and c shows the histogram of the fixations, averaged across nine subjects, over the three ROI's in each of the two time intervals. We found that (a) the maximum of the distribution is always in the ROI corresponding to the time interval considered; and, most notably, (b) the percentage of fixations in the 'wrong' ROI is always below 27% for each subject, and below 10% on average. Notice also that the percentage of fixations in *OFF* increases when moving from τ_1 to τ_2; after one object **O** has been completed, fixations located between **O** and the next object can be used to evaluate information, such as the distance and relative size, that are relevant for an accurate drawing.

In the free viewing task it is not possible to define two time intervals such as τ_1, τ_2 above. Nevertheless, we can define *inter-object* saccades as the saccades between two different objects; as an illustration, the scanpaths displayed in Fig. 3 show that, for a single subject, there are clearly more inter-object saccades (pink dashed lines) in the free viewing task (a) than in the drawing task (b). The mean ratio of inter-object saccades to total saccades confirms this effect, the values being $.36 \pm .19$ for free viewing and $.08 \pm .03$ for drawing (mean and standard deviation, averaged across all subjects; Wilcoxon rank-sum test, $p < .05$).

2.2.2. Saliency drive

Next we try to assess the degree to which image saliency affects fixations. In the following we define the saliency of a point as proportional to the local intensity and orientation contrast, and define the *saliency map* as the collection of saliency values at each location in the image (see Itti & Koch (2001) for implementation details).

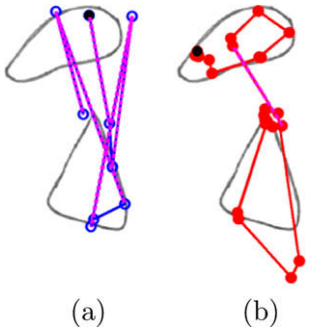

Fig. 3. Scanpaths recorded in trial 2 from a single subject in free viewing (blue) and drawing (red), respectively; dashed pink lines denote inter-object saccades, the black dot represent the first fixation. (For interpretation of color mentioned in this figure legend the reader is referred to the web version of the article.)

Although it is well known that in purely visual tasks, saliency models are able to capture the spatial distribution of fixations (see e.g. Itti & Koch (2001)), it was not obvious *a priori* that this should hold also in motor tasks. Here we show that indeed a similar proportion of fixations in free viewing and in the drawing task are found nearby high saliency regions of the image. This suggests that saliency evaluation is taken into account even in the context of a visuomotor task.

We apply a 'reverse engineering' procedure, which allows us to recover the underlying saliency map (namely the *fixation map*, Wooding (2002)) from the cumulative fixations of all the subjects in each task, and compare it to the saliency map of the image. First we divide the image in a fixed number of cells (12×16), count the total number of fixations per cell by all subjects, assign a 2D Gaussian centered on each cell multiplied by the corresponding value (the covariance matrix is set to .8 times the identity matrix, corresponding to horizontal and vertical standard deviations of ca. 0.9 deg). Then we normalize the resulting matrix under the sum-to-one constraint. The map obtained this way gives an estimate of the probability that a fixation is directed to each region, which in turn is directly related to the saliency of that region. We repeat this procedure for each of the three trials separately; Fig. 4 shows the fixation and saliency maps on trial 2 (Fig. 1c) as an example.

To compare the maps in a given trial, let p^{sal} be the saliency map, and p^{free} and p^{draw} the fixation maps for free viewing and drawing respectively, considered as distributions; we measure how narrow a distribution p is by its entropy $H(p)$, and how much it differs from distribution q by the Kullback–Leibler divergence $KL(p\|q)$ (see e.g. Cover & Thomas (1991) for the mathematical definitions). We compute KL with respect to p^{sal}, and find similar

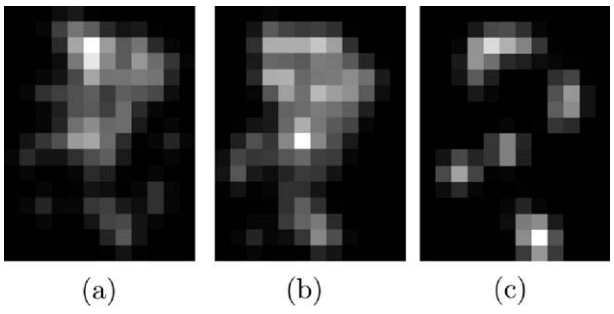

Fig. 4. The fixation maps in free viewing (a) and drawing (b), in the trial corresponding to Fig. 1c. The saliency map is shown in (c).

values for free viewing and drawing ($KL(p^{sal}\|p^{free}) = .78 \pm .08$ and $KL(p^{sal}\|p^{draw}) = .78 \pm .09$, respectively, mean and SEM across the three trials), and significantly smaller than those obtained comparing p^{sal} with random maps ($KL(p^{sal}\|p^{rand}) = 1.71 \pm .11$, averaged over 10,000 random distributions). This small *KL* values show that fixations are distributed preferentially near high saliency regions. Notice also that the *KL* value is not closer to zero because the entropy of the empirical distributions of fixations is slightly larger than the entropy of the saliency map (the ratios being $H(p^{free})/H(p^{sal}) = 1.20 \pm .03$ and $H(p^{draw})/H(p^{sal}) = 1.19 \pm .03$), since the actual fixations are scattered away from the image contours.

2.2.3. Motor continuity

In the drawing experiments, analysis of video recordings indicated that all of our subjects used graphically continuous hand strokes (note that this was not required by experimental instructions); this is, we hypothesize, a specific motor constraint that subjects had to contend with by means of some eye–hand coordination strategy. We explored the specific strategy adopted by subjects, and found that it is remarkably similar across subjects: first of all, there is a clear effect of the drawing task on the length of saccades as well as on the distance between fixations separated by more than one saccade; in addition, most drawing scanpaths resemble an approximated edge-following of the image contours, which parallels the sequence of hand strokes.

To assess the effect on the distance between pairs of fixations we show in Fig. 5 that the mean distance (shown along with 95% confidence intervals, averaged across all subjects and trials) between fixation points is significantly smaller in the drawing task when the number of saccades that separates the two fixations is smaller than 5 (Wilcoxon rank-sum test, $p < .0005$); in particular, saccade length (corresponding to the first point on the horizontal axis) is almost halved in drawing. While this fact could be thought of as a strategy to obtain a higher-resolution sampling of the image, which may be needed to accurately reproduce it, we argue that instead this effect is, at least in part, a consequence of the constraint posed by motor continuity.

Fig. 6 depicts the cumulative plot of fixations, and the corresponding hand position of one subject, at four subsequent stages, during trial 3 (Fig. 1d). The snapshots correspond to the following observed sequence: *hand stops – fixation(s) on the left – saccade – fixation(s) on the right – hand moves*. We interpret the points where the hand stops as key points, at which the hand's action needs to be reprogrammed and thus fixations on the original image become necessary. Qualitative inspection of Fig. 6 shows a general tendency of the gaze to move orderly along the image contour. Such trend is confirmed by the scanpaths of different subjects in trials 2 and 3, plotted in Fig. 7 (with the noticeable exception of the last subject in both trials, where the edge-following is often interrupted by fixations offset from the line towards the curvature center). Quantification of this effect is postponed to the end of Section 2.3. This observations suggests that this peculiar form of the scanpaths is a precise eye–hand coordination strategy in support of graphical continuity of drawing gestures; the effect on saccade length is implied by this kind of scanpaths.

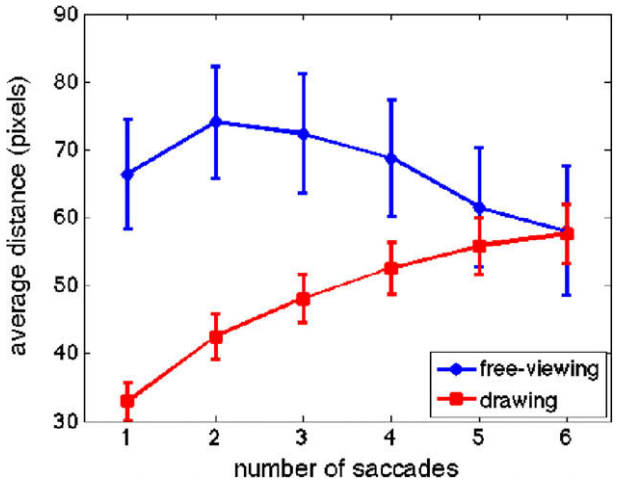

Fig. 5. Average distance between pairs of fixation points as a function of the number of saccades, in the two conditions, across all subjects and trials. Distance is expressed in pixels, with 1 pixel corresponding to ca. 0.05 deg. Error bars denote 95% confidence interval.

2.3. A computational model for sensorimotor coupling

The scanpaths discussed in the previous section are the result, we argue, of a dynamical coupling between eye and hand movements. Here we introduce a computational model of such coupling, and show that under minimal assumptions it produces edge-following scanpaths; subsequently, we give a quantitative comparison of such scanpaths with human data, to assess how well the latter are described as edge-followers.

Probabilistic approaches have a long history in models of eye movements (early and seminal attempts were provided by Ellis & Smith (1985), Hacisalihzade, Stark, & Allen (1992), & Rimey & Brown (1991) who described the sequence of gaze-points in terms of Markov chains and Hidden Markov Models), primarily motivated by the fact that motor and perceptual neural signals are inherently noisy (Kording & Wolpert, 2006). We adopt a Bayesian framework to model the dependencies between eye and hand directions of movement on the basis of sensory inputs, and to derive the optimal movements; these are then combined with saliency information, to simulate the actual scanpaths.

We represent sensory inputs (visual and proprioceptive) and eye–hand motor outputs as two pairs of random variables, (u^e, u^h) and (y^e, y^h), respectively; in addition, the underlying dynamics of the process will be described in terms of the "hidden" random variables (x^e, x^h). Notice that the introduction of hidden variables is a standard approach in generative models of processes whose detailed mechanisms are not observable (Bishop, 2007). In the simulations presented here, proprioceptive input is provided in the form of an estimate of the hand direction of movement (radians), $u^h \in \{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\}$; visual input is coded as a discrete angular value corresponding to the orientation of the image contour in the currently fixated region, computed as the weighted average of the orientation histogram, and discretized to the values (radians), $u^e \in \{0, \frac{\pi}{8}, \ldots, \frac{7\pi}{8}\}$. State variables code for the direction of movement (either of the eye or of the hand) relative to the current position, and take values in the same discrete set as u^h. The outputs of the model, y^e, y^h, which are based on estimates of the hidden states perturbed with Gaussian noise, code for the continuous direction of eye and hand movements.

The dynamics on state variables is introduced by suitable dependencies in time, represented by the Dynamic Bayesian Network (DBN, Murphy (2002)) shown in Fig. 8. Note that such a DBN represents the evolution in time of states in terms of discrete-time slices, two of which at times $t, t+1$ are shown.

The dynamics is composed by the following dependencies: (1) the temporal dependency of the current eye state variable on the previous one $(x^e_t \to x^e_{t+1})$; (2) the analogous dependency for hand state variables $(x^h_t \to x^h_{t+1})$; (3) the dependency of current eye state on the previous hand state $(x^h_t \to x^e_{t+1})$; and (4) the coupling of current hand and eye states $(x^e_{t+1} \to x^h_{t+1})$. Notice in (4) the causal relation between eye and hand: under normal conditions, eye movements typically precede hand movements (e.g. Ballard, Hayhoe, Li, & Whitehead (1992) & Neggers & Bekkering (2000)).

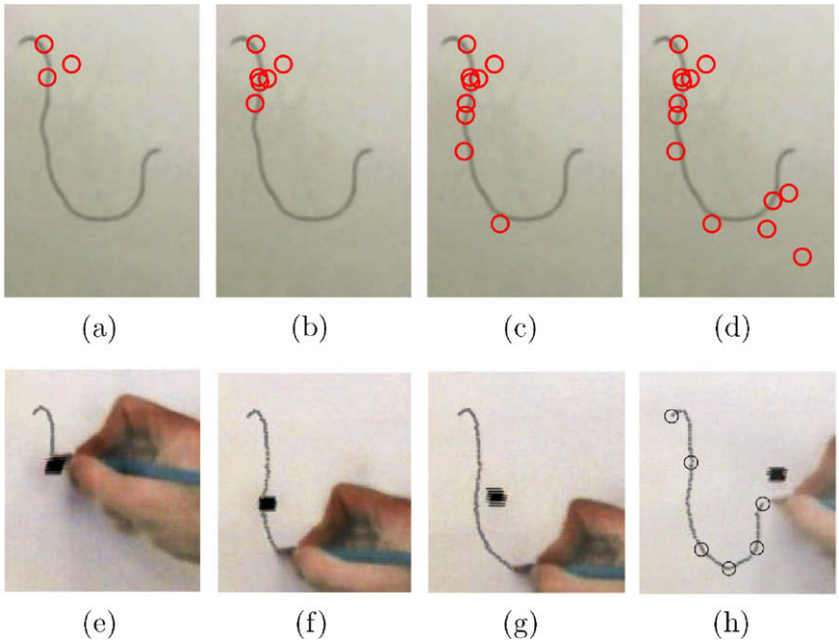

Fig. 6. The sequence of eye and hand movements by one subject in the drawing task, trial 1. In the upper row, cumulative fixations on the original image are represented by red circles. In the lower row the solid black square denotes the gaze point. In (h) the circles denote the endpoints of each trajectory segment.

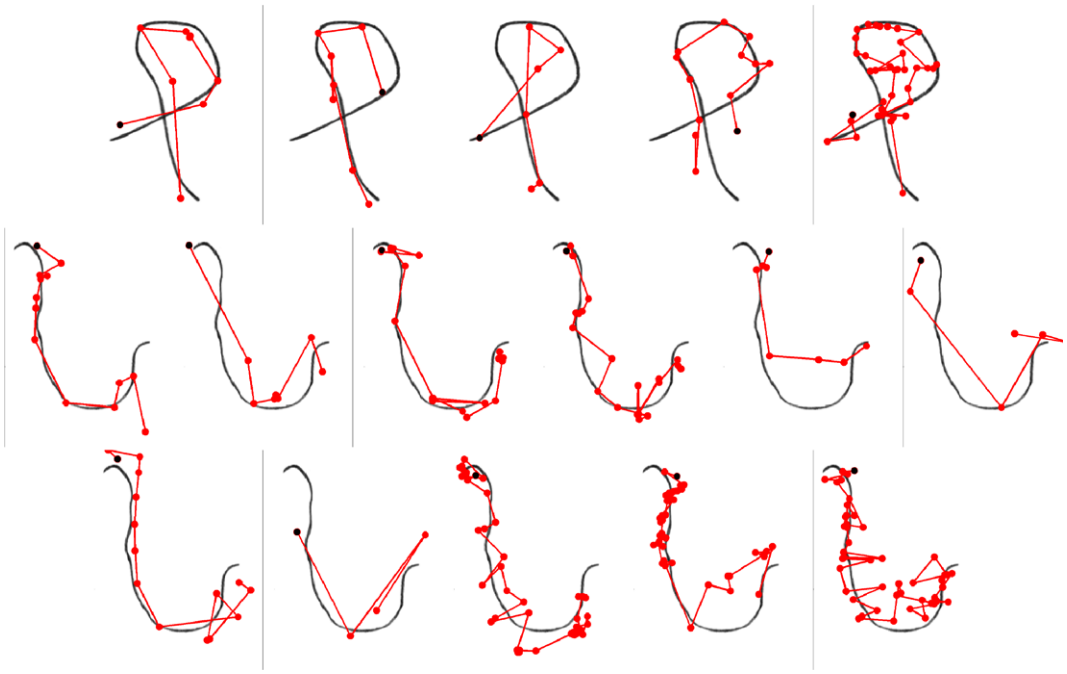

Fig. 7. The scanpath executed during the drawing task by five subjects in trial 2 (first row), and 11 subjects in trial 3 (last two rows). The black circle represents the first fixation point.

At a given time instant t, to compute the actual outputs of the model, we first infer the optimal eye–hand state pair $\left(x_t^{e\star}, x_t^{h,\star}\right)$, and then sample the outputs from their conditional distribution $p\left(y_t^e, y_t^h | x_t^{e\star}, x_t^{h\star}\right)$. Appendix A discusses how to learn the distribution parameters and perform inference.

Eventually, we combine the outputs of the DBN with saliency information. We multiply the saliency map by an oriented 2D Gaussian whose orientation is provided by the inferred eye movement direction $y_{t+1}^{e\star}$, and centered at distance d from the previous fixation point along the chosen direction; d is set to the experimental average saccade length ($d = 33$ pixels, see Fig. 5), the variance along the main axis is set to d and the variance along the orthogonal axis to $d/2$ (see Torralba (2003) for a probabilistic interpretation of a similar interplay between bottom–up and top–down information). The actual fixation point is then chosen as the location of the maximum of the resulting map.

2.3.1. Results

Comparison between scanpaths produced by distinct human behaviors, a saliency-based algorithm and the proposed method. We

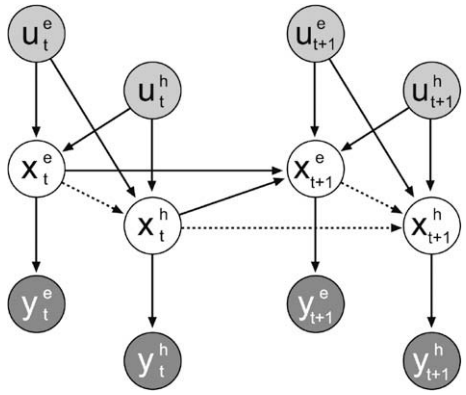

Fig. 8. The DBN representing two time slices (t, left, and $t+1$, right) of the evolution in time of eye and hand direction of movement. Nodes stand for random variables, edges denote conditional dependencies between variables. See text for the meaning of the variables. Dotted connections in the hidden layer highlight the dependence of the hand on the eye, while continuous connections denote the reverse dependence.

have run the DBN on the original image shown in Fig. 1d. The resulting time sequences of eye and hand states are shown in the two top rows of Fig. 9, and the final scanpath is depicted in Fig. 10d. The latter can be compared with a pure bottom–up scanpath, obtained by feeding the saliency map to a *winner-takes-all* network endowed with *Inhibition Of Return* (IOR; Itti & Koch, 2001). In Fig. 10 the saliency-based (a) and the DBN (d) scanpaths are shown along with data from a human subject in the free viewing (b) and drawing (c) conditions. Notice that the IOR mechanism is not explicitly assumed in our model, but gained via motor conditioning on eye movements: the joint distribution of eye–hand states is learned from a data set that promotes graphical continuity of hand movements, and penalizes eye movements that are inconsistent with the hand state. The resulting scanpaths follow the contours of the image without turning back, therefore implementing IOR.

Levenstein measure of similarity between scanpaths: the proposed model approximately fits the human behavior in the drawing task. To assess the similarity between human and simulated scanpaths, we start by partitioning the image in a regular grid composed by $N \times M$ rectangular cells. For all K fixations, each fixation occurring at spatial coordinates $\vec{r}_k = (x,y), k = 1 \ldots K$, is assigned to the corresponding cell. We can now define a discrete version of the scanpath in a given trial, as the one dimensional time-ordered sequence of fixated cells $s(t) = (i_t - 1)M + j_t$ where t is a discrete-time parameter taking values in $1,2,\ldots,K$, and (i_t, j_t) are the matrix coordinates of the cell to which the tth fixation belongs (i and j range from 1 to N and M, respectively). Then a temporal sequence of fixations is grouped into a single *event* if they all fall in the same cell. This procedure replaces the scanpath with a sequence of events, each one belonging to a single cell of the grid. Then each cell is labeled with a symbol (an ASCII character in the interval 'A' to 'e'), and each sequence of events is converted to a string; this enables a straightforward comparison between different scanpaths, by means of a string matching algorithm (Privitera & Stark, 2000). This allows us to evaluate the *string similarity* index as the Levenshtein distance (Levenshtein (1965), but see Privitera & Stark (2000) for implementation details). The final result is then normalized by the string length, to obtain a value in the [0,1] interval.

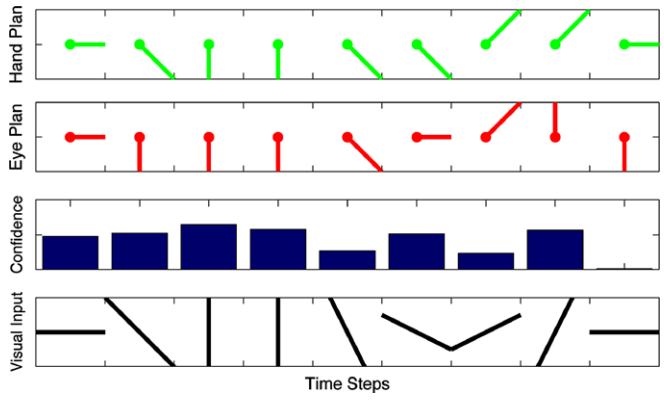

Fig. 9. The simulated discrete-time evolution. From bottom to top: the bottom row represents the sequence of visual inputs, namely the orientation of the foveated image region (graphically coded as an oriented bar); the second row shows the joint probability values associated to the Maximum A Posteriori eye–hand pair states, namely $p(x_t^e, x_t^h | u_t^e, u_t^h)$ (see Appendix A); the third and fourth rows show the DBN outputs, namely the eye and hand movement directions, respectively (depicted as oriented bars departing from the central circular spot).

Fig. 11 summarizes scanpath similarity values obtained by comparing either 11 scanpaths during the drawing task (Fig. 11a) or six free viewing scanpaths (Fig. 11b), against: (1) ([1]red bars) the sequence of gaze-points generated by our model, which implements the edge-following (details are given in Section 2.3); (2) (blue bars) scanpaths generated by the saliency-based algorithm (described in Section 2.2.1 and in Itti & Koch (2001)); and (3) (green bars) random scanpaths (averaged over 10,000 cases). Notice that our model performs significantly better than chance, as well as better than a purely bottom–up saliency-based model, in the drawing task (Wilkoxon rank-sum test, $p < .0001$); vice versa, the control experiment shows that both our and the saliency models are, on average, as poor as chance in modeling free viewing scanpaths ($p > .1$).

3. Discussion

In this paper we have analyzed the visuomotor behavior of subjects involved in a drawing task, which, we argued, can be considered as a paradigmatic one as regards the problem of eye–hand coordination. We recorded the eye movements of human subjects confronted with the task of copying simple shapes. The results obtained, when confronted with data from free viewing control experiments, confirm that there is a strong influence of hand motor constraints on saccade direction (see Aivar et al. (2005), Brouwer & Knill (2007), & Stritzke & Trommershäuser (2007) for other reports of similar effects). To provide a formal account for this phenomenon, we have introduced a computational model of eye–hand coupling, capable of learning the appropriate sensorimotor mapping for the drawing task, and of generating synthetic scanpaths that can be directly compared to human data. A summary of most relevant results of this paper follows.

3.1. The relevance of objects

When confronted with multiple shapes, subjects fixate on each object mostly in the period during which they are drawing that specific object (Figs. 2 and 3). The fact that objects play an important role for gaze shift is well known, (e.g. Desimone & Duncan (1995); see Scholl (2001) for a review); in the context of a motor task, however, we find an enriched notion of object: not only what subjects visually segment as an object, but the units of visuomotor manipulation, namely what they can draw (manipulate) in the course of time.

[1] For interpretation of color in Figs. 1–3, 5–7, 9, and 10, the reader is referred to the web version of this article.

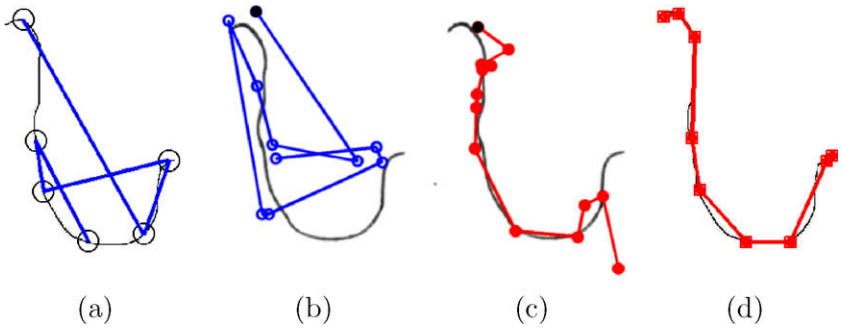

Fig. 10. The scanpaths produced by: (a) a saliency-based algorithm; (b) a free viewing human subject; (c) a human subject in the drawing task; (d) our computational model. The first fixation for the models is on the top left; for the human scanpaths, it is denoted by the black circle.

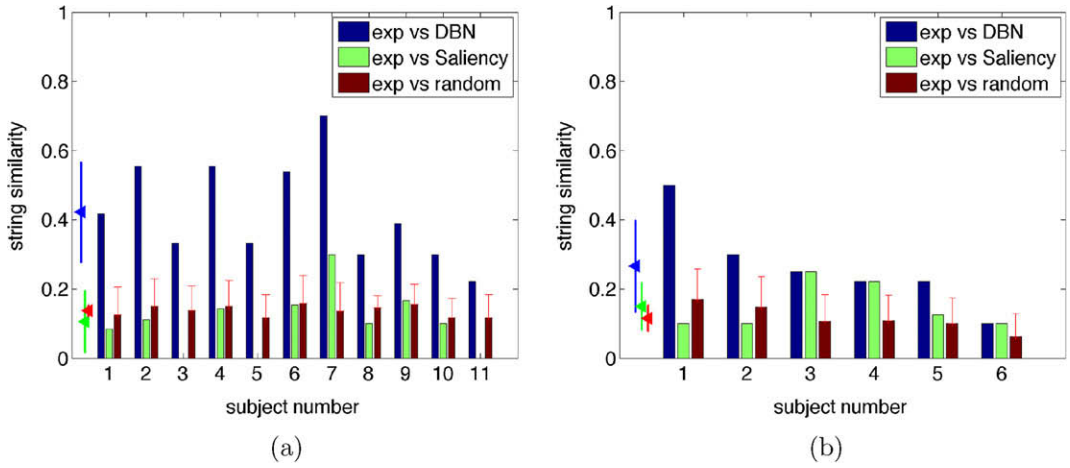

Fig. 11. (a) The similarity, as measured by the Levenstein distance, between experimental scanpaths in the drawing task and those simulated by our model (blue), by a saliency-based algorithm (green), and randomly generated (red); bars denote the values for each subject, while triangles denote mean value and standard deviation across subjects. (b) Same as above, but with human data obtained in the free viewing condition. (For interpretation of color mentioned in this figure legend the reader is referred to the web version of the article.)

3.2. The role of saliency

The interplay between bottom–up and top–down mechanisms in determining attentional selection under natural viewing conditions has been for a long time under debate (see Einhauser, Rutishauser, & Koch, 2008; Findlay & Walker, 1999; Parkhurst, Law, & Niebur, 2002). Bottom–up computational models have been successful in reproducing the spatial distribution of human fixations in static and dynamic scenes (Carmi & Itti, 2006; Itti, 2006; Itti & Koch, 2001). On the other hand, several measures of natural image statistics at fixation locations reveal a more complex scenario, where subsequent fixations are chosen so as to reduce uncertainty about the stimulus (Najemnik & Geisler, 2005, 2008; Nelson & Cottrell, 2007; Raj, Geisler, Frazor, & Bovik, 2005; Renninger, Verghese, & Coughlan, 2007), which in turn depends on subjects knowledge of where relevant information is likely to be located (Chen & Zelinsky, 2006; Droll, Gigone, & Hayhoe, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006).

The degree to which bottom–up mechanisms still contribute to eye movements in highly constrained motor tasks, is still an open issue. We have explored this issue using information-theoretic measures to compare the spatial distribution of fixations in a drawing task and in free viewing, with the saliency map of the images. What can be summarized from our results is that the fixation map, when used as a "reverse engineered" saliency map, exhibits little more information than that related to conspicuity of regions such as the crossing, end points and varying curvature portions of the drawing. This finding suggests that the motor task we studied has little effect on the spatial locations in the image that are more likely to be fixated: overall, subjects look at salient regions most of the time. Vice versa, and most important, the effect of motor constraints is clearly revealed in the temporal sequences of fixations, the scanpaths (see Foulsham & Underwood (2008) & Privitera & Stark (2000) for a discussion of spatial *vs.* sequential aspects of eye movements and idiosyncrasies of scanpaths in free viewing).

3.3. The signature of motor constraints

The drawing task reveals a peculiar oculomotor strategy, that is quite regular across different subjects, and that had not been reported before, to the best of our knowledge. Most of the observed scanpaths are well described as *edge-following* patterns (Fig. 7). We argued that this is the result of the tight coordination of eye movements and drawing gestures: in fact, the interplay of task-related limb movements and oculomotor behavior has been documented in several cases (Ballard et al., 1992; Johansson, Westling, Backstrom, & Randall Flanagan, 2001; Land, 1992; Pelz, Hayhoe, & Loeber, 2001; Rothkopf, Ballard, & Hayhoe, 2007 e.g.), and there is evidence from recordings in the pre-motor areas of the primate brain, that hands and eye compete for motor resources (Rizzolatti, Riggio, Dascola, & Umiltà, 1987; Rizzolatti, Riggio, & Sheliga, 1994; Sheliga, Craighero, Riggio, & Rizzolatti, 1997).

On the other hand, different computational models have been conceived to account for saccadic behavior in a probabilistic setting

(see, for instance, Boccignone & Ferraro (2004), Feng (2006), Hacisalihzade et al. (1992), Itti et al. (2006), Rimey & Brown (1991), Torralba et al. (2006), & Torralba (2003)). Differently from those works, we have provided a principled way to chain eye movements with hand movements (but see Hayhoe & Ballard (2005) for a model of visuomotor coordination in the algorithmic framework of reinforcement learning). The model (Fig. 8) formalizes the dynamic eye–hand coupling by suitable cross-connections between eye and hand-related random variables. Due to the observers' behavior discussed above, the problem here was to contend with motor mechanisms peculiar of the drawing task. As shown in Fig. 9, simulated eye and hand movements are qualitatively similar to the human behavior; furthermore quantitative comparison of the scanpaths, indicates that our model performs significantly better than chance, as well as better than a purely bottom–up saliency-based model, in the drawing task (Fig. 11a; as a control, Fig. 11b shows that both our and the saliency model are, on average, as poor as chance in modeling free viewing scanpaths).

Acknowledgments

The authors are grateful to the Referees and the Associate Editor, for their enlightening and valuable comments that have greatly improved the quality and clarity of an earlier version of this paper.

Appendix A. DBN

To model eye–hand movements in a motor task, 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, and with $\bar{y} = (y^e, y^h)$ the pair of variables accounting for eye and hand output signals; $\bar{x} = (x^e, x^h)$ denotes the corresponding pair of eye and hand (hidden) state variables, that we will use to model the temporal dynamics. The simplest dynamics on the state variables is introduced by a dependency between the eye states at two subsequent time steps $(x^e_t \to x^e_{t+1})$, and similarly for the hand states; this corresponds to two Input–Output Hidden Markov Models (IOHMM, Bengio & Frasconi, 1996; Feng, 2006). However, the most important point here is that the two processes are not independent but rather modeled as coupled chains: at a given time step the hand state depends on the eye state $(x^e_{t+1} \to x^h_{t+1})$, which in turn depends on the previous hand state $(x^h_t \to x^e_{t+1})$; indeed, these are the conditional dependencies that model the very visuomotor nature of eye–hand coordination. The resulting graphical model (Fig. 8) unifies the IOHMM and another kind of model known in the literature as the Coupled HMM (Murphy, 2002). We call the resulting network an *Input–Output Coupled Hidden Markov Model*.

By generalizing the two time slices snapshot of Fig. 8 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:

$$\begin{aligned} p(\bar{x}_{1:T}, \bar{y}_{1:T} | \bar{u}_{1:T}) = {} & p(x^e_1 | u^e_1, u^h_1) p(y^e_1 | x^e_1) p(x^h_1 | u^e_1, u^h_1, x^e_1) p(y^h_1 | x^h_1) \\ & \cdot \prod_{t=1}^{T-1} \{ p(x^e_{t+1} | u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t) p(y^e_{t+1} | x^e_{t+1}) \\ & \cdot p(x^h_{t+1} | u^e_{t+1}, u^h_{t+1}, x^e_{t+1}, x^h_t) p(y^h_{t+1} | x^h_{t+1}) \} \end{aligned} \tag{A.1}$$

With the specification of variables introduced in Section 2.3, parameter learning amounts to estimating the discrete conditional distribution over hidden states, and the parameters of the output conditional distributions, by adapting the Baum–Welch variant of the EM algorithm (Rabiner, 1989) to our network (details provided in Coen-Cagli, Coraggio, Napoletano, & Boccignone, 2008). To reduce the number of parameters, we assumed the output distributions to be Gaussian with mean equal to the corresponding hidden value and $\sigma = 10$ deg; in addition, we assumed $u^h_{t+1} = x^h_t$ and mirror symmetry with respect to u^e, which leaves us with a total of 8000 parameters for the distributions over hidden states. The training set was composed by 100,000 synthetic two-step sequences, created by imposing graphical continuity and edge-following on randomly generated short line contours.

To generate the outputs, given the inputs at time step $t+1$, we first evaluate the distribution of hidden states $p(\bar{x}_{t+1} | \bar{u}_{1:t+1})$, then take the Maximum A Posteriori estimate $(x^{e\star}_{t+1}, x^{h\star}_{t+1}) = \arg\max \{ p(x^e_{t+1}, x^h_{t+1} | \bar{u}_{1:t+1}) \}$, and eventually we generate the outputs by sampling $p(y^e_{t+1}, y^h_{t+1} | x^{e\star}_{t+1}, x^{h\star}_{t+1})$. Notice that, according to the network structure, the hidden state at time $t+1$ depends only on the input subsequence $\bar{u}_{1:t+1}$; thus, making use of Eq. (A.1) we find:

$$\begin{aligned} p(\bar{x}_{t+1} | \bar{u}_{1:t+1}) = {} & \sum_{\bar{x}_{1:t}} p(x^e_{t+1} | u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t) p(x^h_{t+1} | u^e_{t+1}, u^h_{t+1}, x^e_{t+1}, x^h_t) \\ & \cdot p(\bar{x}_{1:t} | \bar{u}_{1:t}) \end{aligned} \tag{A.2}$$

where the first two terms on the r.h.s. have been estimated in the learning stage. Eq. (A.2) is a particular case of recursive Bayesian filtering (Bishop, 2007).

References

Aivar, M., Hayhoe, M., Chizk, C., & Mruczek, R. (2005). Spatial memory and saccadic targeting in a natural task. *Journal of Vision, 5*(3), 177–193.

Ballard, D., Hayhoe, M., Li, F., & Whitehead, S. (1992). Hand–eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 337*, 331–338.

Bengio, Y., & Frasconi, P. (1996). Input–output hmm's for sequence processing. *IEEE Transactions on Neural Networks, 7*, 1231–1249.

Bishop, C. (2007). *Pattern recognition and machine learning*. Berlin: Springer.

Boccignone, G., & Ferraro, M. (2004). Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications, 331*(1–2), 207–218.

Brouwer, A., & Knill, D. (2007). The role of memory in visually guided reaching. *Journal of Vision, 7*(5), 6.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*(9), 898–914.

Carter, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Where next? *Trends in Cognitive Sciences, 10*(7), 292–293.

Chen, X., & Zelinsky, G. (2006). Real-world visual search is dominated by top–down guidance. *Vision Research, 46*, 4118–4133.

Coen-Cagli, R., Coraggio, P., Napoletano, P., & Boccignone, G. (2008). What the draughtsmans hand tells the draughtsmans eye: A sensorimotor account of drawing. *International Journal of Pattern Recognition and Artificial Intelligence, 5*, 1015–1029.

Coen-Cagli, R., Coraggio, P., Boccignone, G., & Napoletano, P. (2007). The bayesian draughtsman: A model for visuomotor coordination in drawing. In *Advances in brain vision and artificial intelligence*, LNCS (Vol. 4729, pp. 161–170). Berlin: Springer-Verlag.

Cohen, D. J. (2005). Look little, look often: The influence of gaze frequency on drawing accuracy. *Perception Psychophysics, 67*, 997–1007.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York, NY: Wiley and Sons.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Droll, J., Gigone, K., & Hayhoe, M. (2007). Learning where to direct gaze during change detection. *Journal of Vision, 7*(14), 1–12.

Einhauser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2), 1–19.

Ellis, S. R., & Smith, J. D. (1985). Patterns of statistical dependency in visual scanning. In R. Groner, G. W. McConkie, & C. Menz (Eds.), *Eye movements and human information processing* (pp. 221–238). Amsterdam: Elsevier.

Feng, G. (2006). Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research, 7*, 70–95.

Findlay, J., & Walker, R. (1999). A model of saccadic eye movement generation based on parallel processing and competitive inhibition. *Behavioral and Brain Science, 22*, 661–674.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2), 1–17.

Gowen, E., & Miall, R. (2006). Eye–hand interactions in tracing and drawing tasks. *Human Movement Science, 25*, 568–585.

Hacisalihzade, S. S., Stark, L. W., & Allen, J. S. (1992). Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems, Man and Cybernetics, 22*, 474–481.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4).

Itti, L. (2006). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition, 14*, 959–984.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems* (Vol. 18, pp. 1–8). MIT Press.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews – Neuroscience, 2*, 1–11.

Johansson, R., Westling, G., Backstrom, A., & Randall Flanagan, J. (2001). Eye–hand coordination in object manipulation. *Journal of Neuroscience, 21*, 6917–6932.

Kording, K., & Wolpert, D. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10*(7), 319–326.

Land, M. (1992). Predictable eye–head coordination during driving. *Nature, 359*, 318–320.

Land, M. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research, 25*, 296–324.

Lee, T., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434–1448.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR, 163*, 845–848.

Marr, D. (1982). Vision. Francisco, CA: Freeman, S.

Murphy, K. (2002). Dynamic Bayesian Networks: Representation, inference and learning. PhD dissertation, Berkeley, University of California, Computer Science Division.

Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387–391.

Najemnik, J., & Geisler, W. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision, 8*(3), 1–14.

Neggers, S., & Bekkering, H. (2000). Ocular Gaze is anchored to the target of an ongoing pointing movement. *Journal of Neurophysiology, 83*(2), 639–651.

Nelson, J., & Cottrell, G. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing, 70*, 2256–2272.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107–123.

Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research, 139*, 266–277.

Privitera, C., & Stark, L. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(9), 970–982.

Rabiner, L. (1989). A tutorial on hmm and selected applications in speech recognition. In *Proceedings of IEEE* (pp. 257–286).

Raj, R., Geisler, W., Frazor, R., & Bovik, A. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 22*, 2039–2049.

Renninger, L., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision, 7*(3), 1–17.

Rimey, R. D., & Brown, C. M. (1991). Controlling eye movements with hidden markov models. *International Journal of Computer Vision, 7*(1), 47–65.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltà, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25*, 31–40.

Rizzolatti, G., Riggio, L., & Sheliga, B. (1994). Space and selective attention. In C. Umiltà & M. Moscovitch (Eds.). *Attention and performance* (Vol. XV, pp. 231–265). Cambridge, MA: MIT Press.

Rothkopf, C., Ballard, D., & Hayhoe, M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14), 1–20.

Scholl, B. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46.

Sheliga, B., Craighero, L., Riggio, L., & Rizzolatti, G. (1997). Effects of spatial attention on directional manual and ocular responses. *Experimental Brain Research, 114*, 339–351.

Stritzke, M., & Trommershäuser, J. (2007). Eye movements during movement under risk. *Vision Research, 47*, 2000–2009.

Tchalenko, J. (2007). Eye movements in drawing simple lines. *Perception, 36*, 1152–1167.

Tchalenko, J., Dempere-Marco, R., Hu, X., & Yang, G. (2003). Eye movement and voluntary control in portrait drawing. In *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 705–727). Amsterdam: Elsevier.

Tchalenko, J., & Miall, C. (2008). Eye–hand strategies in copying complex lines. *Cortex*, 28.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision, 53*, 153–167.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766–786.

Viviani, P., & Flash, T. (1995). Minimum-jerk model, two-thirds power law, and isochrony: Converging approaches to the movement planning. *Journal of Experimental Psychology, 21*, 32–53.

Wooding, D. (2002). Fixation maps: Quantifying eye-movement traces. In *Eye tracking research and applications (ETRA) symposium* (pp. 31–36).