What has been missed for predicting human attention in viewing driving clips?

Jiawei Xu [1] , **Shigang Yue** [1] , **Federica Menchinelli** [2] , **Kun Guo** Corresp. [2]

[1] School of Computer Science, University of Lincoln, Lincoln, United Kingdom

[2] School of Psychology, University of Lincoln, Lincoln, United Kingdom

Corresponding Author: Kun Guo
Email address: kguo@lincoln.ac.uk

Recent research progress on the topic of human visual attention allocation in scene perception and its simulation is based mainly on studies with static images. However, natural vision requires us to extract visual information that constantly changes due to egocentric movements or dynamics of the world. It is unclear to what extent spatio-temporal regularity, an inherent regularity in dynamic vision, affects human gaze distribution and saliency computation in visual attention models. In this free-viewing eye-tracking study we manipulated the spatio-temporal regularity of traffic videos by presenting them in normal video sequence, reversed video sequence, normal frame sequence, and randomised frame sequence. The recorded human gaze allocation was then used as the 'ground truth' to examine the predictive ability of a number of state-of-the-art visual attention models. The analysis revealed high inter-observer agreement across individual human observers, but all the tested attention models performed significantly worse than humans. The inferior predictability of the models was evident from indistinguishable gaze prediction irrespective of stimuli presentation sequence, and weak central fixation bias. Our findings suggest that a realistic visual attention model for the processing of dynamic scenes should incorporate human visual sensitivity to spatio-temporal regularity and central fixation bias.

1   **What has been missed for predicting human attention in viewing driving clips?**

3   Jiawei Xu[1], Shigang Yue[1], Federica Menchinell[2], Kun Guo[2,*]

4   [1]School of Computer Science, [2]School of Psychology, University of Lincoln,

5   Lincoln LN6 7TS, UK

7   [*] Corresponding Author:

8   Email: kguo@lincoln.ac.uk

10   **Abstract**

11   Recent research progress on the topic of human visual attention allocation in scene

12   perception and its simulation is based mainly on studies with static images. However, natural

13   vision requires us to extract visual information that constantly changes due to egocentric

14   movements or dynamics of the world. It is unclear to what extent spatio-temporal regularity, an

15   inherent regularity in dynamic vision, affects human gaze distribution and saliency computation

16   in visual attention models. In this free-viewing eye-tracking study we manipulated the spatio-

17   temporal regularity of traffic videos by presenting them in normal video sequence, reversed

18   video sequence, normal frame sequence, and randomised frame sequence. The recorded human

19   gaze allocation was then used as the 'ground truth' to examine the predictive ability of a number

20   of state-of-the-art visual attention models. The analysis revealed high inter-observer agreement

21   across individual human observers, but all the tested attention models performed significantly

22   worse than humans. The inferior predictability of the models was evident from indistinguishable

23   gaze prediction irrespective of stimuli presentation sequence, and weak central fixation bias. Our

24   findings suggest that a realistic visual attention model for the processing of dynamic scenes

25   should incorporate human visual sensitivity to spatio-temporal regularity and central fixation

26   bias.

28   **Introduction**

29   The amount of available visual information in our surroundings is often beyond our

30   brain's processing capability. To effectively interact with our natural and social world, we

31   selectively gaze at and process a limited number of local scene regions or visual items that are

32   informative or interesting to us. As our gaze allocation is a sensitive index of attention,

33   motivation and preference (Henderson, 2007), the fixated regions tend to have a distinct

34   subjective perceptual quality which enables them to stand out from the neighbours, and the

35   choice of these salient targets reflects our internal representation of the external world.

36   The central research question in this active visual exploration process is to understand

37   how we choose the fixated regions in the scene. Psychological studies have suggested that both

38   bottom-up local saliency computation and top-down cognitive processes are actively involved in

39   determining our fixations in scene exploration (Henderson, 2007). The bottom-up signals include

40   image immanent features that could transiently attract fixation irrespective of a particular task

41   demand. For instance, we tend to avoid low-contrast and homogeneous 'predictable' regions in

42   natural scenes, and bias our fixation to local features with high-contrast, high spatial frequency,

43   high edge density, and complex local structure (e.g., curved lines, edges and corners, as well as

44   occlusions or isolated spots) (Mannan, Ruddock, & Wooding 1996; Reinagel & Zador, 1999;

45   Parkhurst & Niebur, 2003; Acik, Onat, Schumann, Einhauser, & Konig, 2009), or to local

46   regions deviated from surrounding image statistics (Einhäuser et al., 2006). The top-down

47   signals, such as an individual's expectation, experience, memory, semantic and task-related

48   knowledge, are task-dependent and could significantly modulate our gaze allocation in scene

49   exploration (e.g., Guo, Smith, Powell, & Nicholls, 2012; Pollux, Hall, & Guo, 2014; Gavin,

50   Houghton, & Guo, 2015). Recent studies have further proposed that fixation selection is uniquely

51   driven by learned associations between stimuli and rewards (Anderson, 2013), and is influenced

52   by aspects of innate human bias, such as the tendency to fixate human/animal face and body in

53   the scene and to look more often at the central part of the scene (Tatler, Hayhoe, Land, & Ballard,

54   2011).

55   Motivated by these empirical studies, computer vision scientists have tried to develop

56   human-vision-inspired foveated active artificial vision systems capable of working in real-time,

57   and have proposed various computational models for predicting and/or mimicking human gaze

58   allocation in natural vision (Borji, Sihite, & Itti, 2013). Closely resembling our knowledge about

59   neural processing in the early visual system, the widely cited bottom-up saliency model (Itti &

60   Koch, 2000) compares local image intensity, colour and orientation through centre-surround

61   filtering at eight spatial scales, combines them into a single salience (conspicuity) map with a

62   winner-take-all network and inhibition-of-return, and then produces a sequence of predicted

63   fixations that scan the scene in order of decreasing salience. To improve its relatively low level

64   of predictive power (e.g., 57% − 68% correct fixation prediction in some scene free-viewing

65   tasks; Betz, Kietzmann, Wilming, & Konig, 2010), aspects of innate human bias and top-down

66   processing, such as face or object detection, scene context, natural statistics and learned

67   associations between stimuli and rewards, are later incorporated into the model (e.g., Torralba,

68   Oliva, Castelhano, & Henderson, 2006; Judd, Ehinger, Durand, & Torralba, 2009; Kanan, Tong,

69   Zhang, & Cottrell, 2009; Goferman, Zelnik-Manor, & Tal, 2012). This combination of bottom-

70   up saliency-driven information and top-down scene understanding has greatly improved gaze

71   prediction in real-word visual search and free-viewing tasks.

72   However, the experimental findings discussed above and computational visual attention

73   models in scene perception are derived mainly from studies using images in laboratory settings.

74   In the real world we need to select, extract and process visual information that constantly

75   changes due to egocentric movements or dynamics of the world. The motion cues are likely to

76   attract visual attention and play a crucial role in subsequent scene understanding. Indeed, recent

77   psychological studies have shown that motion and continuous temporal change are stronger

78 predictors for gaze allocation than local image intensity or colour (Carmi & Itti, 2006; Le Meur,

79 Le Callet, & Barba, 2007; Dorr, Martinetz, Gegenfurtner, & Barth, 2010), and can enhance

80 perceptual performance in understanding ambiguous visual cues, such as degraded or subtle

81 facial expressions (Cunningham & Wallraven, 2009).

82 Currently, most of the visual attention models focus on spatial information processing

83 with little consideration of temporal cues (Borji, Sihite, & Itti, 2014), especially for the evolved

84 spatial-temporal coherence over a period of time. For example, in the saliency-based visual

85 attention system (Itti, Koch, & Niebur, 1998), three feature channels – colour, intensity, and

86 orientation, and multi-scale image features are combined into a single topographical saliency

87 map, in order to break down the complex scene understanding problem by rapidly selecting

88 conspicuous locations for the detailed analysis. Attention based on information maximization

89 (AIM) model uses Shannon's self-information measure to calculate the saliency of image regions

90 (Bruce & Tsotsos, 2006), while the incremental coding length (ICL) approach measures the

91 respective entropy gain of each feature to maximize the entropy of the sampled visual features in

92 both dynamic and static scenes (Hou & Zhang, 2008). Recently a spatiotemporal saliency

93 algorithm based on a centre-surround framework has been proposed in which the salience map

94 was calculated using dynamic textures (Mahadevan & Vasconcelos, 2010). Marat's model

95 (Marat et al., 2009) extracts two signals from the video stream corresponding to the two main

96 outputs of the retina: parvocellular and magnocellular pathways, both signals are split into

97 elementary feature maps by cortical-like filters. These maps are then fused into a spatio-temporal

98 saliency map.

99 To further enhance the model's predictive power, we proposed a dynamic visual attention

100 model (Xu, Yue, & Tang, 2013) based on three channels. The first channel is motion intensity to

101   reveal the faster moving objects. Then the spatial cues indicate the different moving objects in

102   the spatial location, while the temporal cues denote the variability of one object in the temporal

103   dimension. The rarity factor is imported to describe the unexpected or novel moving object that

104   appeared in the visual field, which is likely to attract visual attention and should be captured or

105   reflected in the saliency map, but to ignore the receding objects. The model has been developed

106   further to analyze the attention shifts by fusing the top-down bias and bottom-up cues (Xu & Yue,

107   2014). The proposed model has five modules: the pre-learning process, top-down bias, bottom-

108   up mechanism, multi-layer neural network and attention.   The motion attention is estimated

109   based on a motion vector field (MVF) comprising of three inductors: intensity inductor, spatial

110   coherence inductor, and temporal coherence inductor (Ma, Hua, Lu, & Zhang, 2005). Ben et al.

111   (2008) also proposed a biologically motivated dynamic bottom-up selective attention model,

112   which employs the maximum entropy algorithm to analyze the dynamics of the successive static

113   saliency maps.

114   Although some of these models work reasonably well in simulated laboratory conditions,

115   it is unclear how well they could predict human attention allocation in naturalistic conditions.

116   For instance, these models tend to assign a high saliency rating to a moving object regardless of

117   the predictability of motion trajectory, but we are unlikely to attend to a predictable movement

118   (e.g., a travelling car with constant speed and predictable direction) for a prolonged period in

119   natural environments. Given that our visual system is evolved to process dynamic spatiotemporal

120   information, it is reasonable to assume that natural scene dynamics could elicit an optimal gaze

121   strategy to facilitate scene understanding. Among various scene-dependent spatial and temporal

122   cues in natural vision, spatio-temporal regularity or predictability, in which objects around us

123   often occur and move in statistically predictable ways to create a stream of visual inputs which

124   are spatially and temporally coherent (e.g., the trajectory of a car moving on the motorway or an

125   apple falling from a tree), is an inherent common regularity in the natural environment. Recent

126   psychological and brain imaging studies have shown that the human visual system exploits this

127   spatio-temporal regularity to effectively process dynamic visual inputs, such as detecting and

128   discriminating the moving target (e.g., Guo et al., 2004; Pollux & Guo, 2009; Hall, Pollux, &

129   Guo, 2010; Pollux, Hall, Roebuck, & Guo, 2011; Roebuck, Bourke, & Guo, 2014).

130   To the best of our knowledge, we are not aware of any visual attention model that has

131   examined the effect of movement predictability on its predictive power or has incorporated this

132   spatio-temporal regularity into saliency computation. In this study, we installed a camera in the

133   car and recorded typical driving scenes with different levels of traffic complexity. To

134   systematically manipulate the videos' spatio-temporal regularity, each clip was presented in four

135   different formats, such as normal video sequence, reversed video sequence, normal image

136   sequence, and randomised image sequence. Such manipulation of the presentation sequence of

137   the videos could selectively disrupt the temporal frame order (i.e. reversed video sequence), the

138   spatial frame coherence (i.e. normal image sequence) or the spatio-temporal frame coherence (i.e.

139   randomised image sequence) of the visual stream inputs. The gaze allocation of human observers

140   viewing these clips was recorded and then used as the 'ground truth' to examine the predictive

141   ability of the state-of-the-art visual attention models.

142

143   **Methods**

144   **Ethics Statement**

145   The Ethical Committee in School of Psychology, University of Lincoln, approved this

146   study. Written informed consent was obtained from each participant prior to the study, and all

147   procedures complied with the British Psychological Society Code of Ethics and Conduct. All the

148   relevant measures, conditions and data exclusions have been reported, and the sample size was

149   determined based on our previous research (Röhrbein et al., 2015).

150   **Eye-tracking study**

151   Thirty-five undergraduate students (27 female, 8 male), age ranging from 18 to 28 years

152   old with a mean of 19.4±0.37 (Mean±SEM), volunteered to participate in this study. All

153   participants had normal or corrected-to-normal visual acuity, and normal colour vision (checked

154   with Ishihara's Tests for Colour Deficiency, 24 Plates Edition).

155   Visual stimuli were presented on a non-interlaced gamma-corrected colour monitor (30

156   cd/m^2 background luminance, 100 Hz frame rate, Mitsubishi Diamond Pro 2070SB) with a

157   resolution of 1024×768 pixels. At a viewing distance of 57 cm, the monitor subtended a visual

158   angle of 40 × 30°.

159   Thirty high-definition colour video clips were sampled from the authors' collection of

160   traffic videos, taken by attaching a digital camcorder to the car windscreen. These video clips

161   represent a variety of daily traffic conditions, such as approaching traffic lights, road junction

162   and roundabout; low-speed driving at car park and residential area or on inner-city road; high-

163   speed driving on single carriageway, dual carriageway and motorway (see Fig.1 for example

164   frames from typical video scenes). Each video clip lasted 3 seconds (30 frames per second) and

165   all clips were identical in size (720×404 pixels).

166   A   B

167
168   C   D

169
170   E   F

171
172   G   H

173
174   Figure 1. Example frames from traffic videos used in this study. The videos included clips

175   representing approaching traffic lights (A) and a road junction (B), low-speed driving at a car

176   park (C) and residential area (D), driving in an inner-city road (E and F), and high-speed driving

177   in a dual carriageway (G) and motorway (H).

178

179   To examine to what extent the spatio-temporal regularity of the videos would affect our

180   gaze allocation, each video clip was manipulated and later presented in four conditions: (1)

181   *normal video sequence*: the clip was presented in its original, predictable sequence (from 1st

182   frame to 90th frame); (2) *reversed video sequence*: the same clip was presented in reverse motion

183   or time-reversed sequence (from 90th frame to 1st frame); (3) *normal image sequence*: six still

184   shots were sampled at the beginning, middle and end of the video sequence (i.e. 1st, 16th, 31st,

185   46th, 61st and 76th frame), and then presented in their original order, each still image was

186   displayed for 500 ms; (4) *randomised image sequence*: the same six still images were presented

187   in a randomised sequence, each lasting 500 ms. As a result, 120 video clips were generated for

188   the testing session (30 original clips × 4 manipulations). These clips were displayed once in a

189   random order during the testing.

190   A free-viewing task was used to mimic natural viewing condition. During the

191   experiments the participants sat in a chair with their head restrained by a chin-rest, and viewed

192   the display binocularly. To calibrate eye movement signals, a small red fixation point (FP, 0.3°

193   diameter, 15 cd/m^2 luminance) was displayed randomly at one of 9 positions (3 × 3 matrix)

194   across the monitor. The distance between adjacent FP positions was 10°. The participant was

195   instructed to follow the FP and maintain fixation for 1 second. After the calibration procedure,

196   the trial was started with four FPs presented at the corners of the screen area in which the videos

197   were displayed. To minimize initial central fixation bias in scene viewing (Tatler, 2007), the

198   participant was instructed to randomly choose one FP to fixate. If the participant maintained

199   fixation for 1 second, the FP disappeared and a testing video clip was presented at the centre of

200   the screen. During the free-viewing presentation, the participant was instructed to "view the

201   video clip as you normally do". No reinforcement was given during this procedure. All

202   participants have viewed all of the 120 video clips.

203   Horizontal and vertical eye positions from the self-reported dominant eye (determined

204   through the Hole-in-Card test or the Dolman method if necessary) were measured using a Video

205   Eyetracker Toolbox (a camera-based system tracking pupil centre and corneal reflection) with 50

206   Hz sampling frequency and up to 0.25° accuracy (Cambridge Research Systems, UK.

207   http://crsltd.com/tools-for-vision-science/eye-tracking/high-speed-video-eye-tracker-toolbox/).

208   The software developed in Matlab computed horizontal and vertical eye displacement signals as

209   a function of time to determine eye velocity and position (Guo, Mahmoodi, Robertson, & Young,

210   2006). The location of gaze point within each video frame was extracted from the raw data.

211

212   **Data analysis and model comparison**

213   Human gaze allocation in viewing of the normal video sequences was treated as the

214   'ground truth' or 'bench mark' to compare the participants' gaze behaviour on the four

215   presentation conditions and the predictive power of the chosen computational models. The bench

216   mark gaze location at each frame was averaged from 35 participants after checking inter-

217   observer agreement (or individual differences in gaze allocation) through the "all-except–one

218   observers" protocol (i.e. how well the fixation distribution of each participant can be predicted

219   by that of other participants; Röhrbein et al., 2015). We then adopted a frame-based similarity

220   measurement to quantify the differences in comparison. For each frame within a given video

221   sequence, we first calculated the Euclidean Distance (D_f) between bench mark gaze location

222   (x_0, y_0) and predicted gaze location (x_1, y_1) from a chosen computational model.

223        $$D_f = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

224   Then for the same video frame, we used the same equation to calculate the Euclidean

225   Distance between bench mark gaze location (x_0, y_0) and actual gaze location (x_1, y_1) from each

226   participant $(\check{D}_i)$, and computed the averaged Euclidean Distance across all the participants $(\check{D})$.

227
$$\check{D} = \frac{1}{35} \left(\sum_{i=1,2,\ldots,35} \check{D}_i \right)$$

228   The similarity score (S) for this video frame was then computed as

229
$$S = \frac{\check{D}}{\check{D} + D_f}$$

230
231   After repeating the same protocol for all the frames within this given video sequence, the

232   final similarity score for the clip was then averaged across all the similarity measures for each

233   frame in the video, and was ranged from 0 to 1, in which 0 represents no similarity at all and 1

234   means identical spatial-temporal gaze distribution between human observers and model

235   prediction. This normalized approach for similarity measurement could minimize the impact of

236   stimulus size or content on similarity calculation, and facilitate data comparison across different

237   stimulus conditions, different participant groups or different studies.

238   Normalized Scanpath Saliency (NSS) and Area under Curve (AUC) or Receiver

239   Operating Characteristics (ROC) are other commonly used measures in the literature to compare

240   the predictive power of visual attention models (Röhrbein, Goddard, Schneider, James, & Guo,

241   2015). However, NSS is the average of the response values of human gaze allocation without

242   considering spatial phase (Rothenstein & Tsotsos, 2006), and the computation of AUC or ROC is

243   more suitable for static image rather than dynamic video analysis (e.g., unable to capture frame-

244   by-frame dynamic changes in a model's predictive performance; Green & Swets, 1966; Zou,

245   O'Malley, & Mauri, 2007). In this study we therefore focused on comparing similarity score

246   calculated from different presentation conditions with different computational models. As

247   discussed in the Introduction, very few visual attention models have incorporated motion cues in

248   the saliency computation. Based on their significance in the literature, we chose to compare the

249   predicitive power of six models from Itti et al. (1998), Bruce and Tsotsos (2006), Hou and

250   Zhang (2008), Mahadevan and Vasconcelos (2010), Marat et al. (2009), and Xu et al. (2013).

251   These models have all been constructed for dynamic displays and contain a 'motion' channel.

252   For each tested video clip, its saliency map (or predicted fixation allocation) was computed by

253   these models separately using Matlab programmes with default settings obtained from the model

254   developers.

255

256   **Results**

257   We first calculated the baseline of human performance using the similarity score, which

258   measures how well the fixation distribution of each participant can be predicted by that of other

259   participants (all-except-one observers). Specifically, we selected one participant's fixation

260   allocations as actual fixations that we wanted to predict, and compared them with the averaged

261   fixation allocations from the remaining participants. By repeating this procedure for all

262   participants and averaging the resulting similarity scores, we obtained a measure for the

263   variability (or inter-observer agreement) within all human gaze patterns that can serve as an

264   upper bound to the performance of a given computational model. As shown in Fig. 2,

265   irrespective of the stimuli presentation sequence, the gaze patterns from individual human

266   participants were highly similar, with a similarity score distributed around 0.83, suggesting small

267   human individual differences and a reliable 'ground truth' baseline or benchmark for model

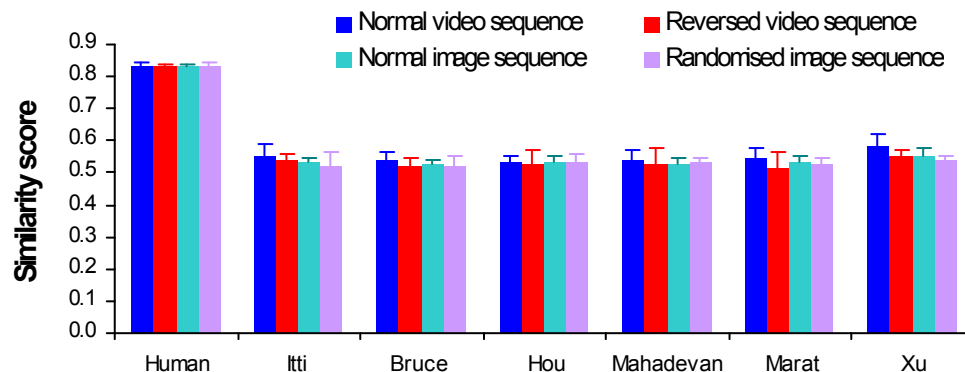
268   comparisons. The chance or random baseline was calculated by measuring the similarity score

269   between actual human gaze allocation and randomly selected 'gaze' allocation at each frame for

270   each video clip and presentation sequence. After repeating the same procedure 100 times, the

271   averaged similarity score varied between 0.46 and 0.49 across different presentation sequences

272   which was significantly below human baseline.

273

274   Figure 2. Similarity scores from human observers and different visual attention models in

275   computing saliency map of traffic videos and images with different presentation sequences. Error

276   bars represent standard deviation.

277

278   When compared with the averaged human gaze pattern and chance baseline, the

279   predictive power of the tested models was clearly below the human baseline but was above the

280   chance baseline regardless of presentation sequence (Fig. 2). To systematically compare the

281   models' performance under different stimuli presentation sequences, a 7 (models: human

282   baseline plus 6 visual attention models) × 4 (presentation sequences) repeated-measures analysis

283   of variance (ANOVA) was conducted with similarity score as the dependent variable.

284   Greenhouse–Geisser correction was applied where sphericity was violated. The analysis revealed

285   a significant main effect of model ($F(6,174)=1614.88$, $p<0.001$, $\eta_p^2=0.98$) and presentation

286   sequence ($F(2.03,58.77)=10.67$, $p<0.001$, $\eta_p^2=0.27$), and a significant interaction between model

287   and presentation sequence ($F(6.23,180.67)=4.2$, $p<0.001$, $\eta_p^2=0.13$). Detailed post-hoc

288   comparisons with Bonferroni correction further showed that all the visual attention models

289   performed significantly worse than human baseline (all *ps*<0.01). Among the tested models, the

290   Xu model (Xu et al., 2013) performed slightly better than other models (all *ps*<0.01), and the

291   performance from the rest of models was indistinguishable (all *ps*>0.05). Regarding stimulus

292   presentation sequence, the normal video sequence induced better model performance for the

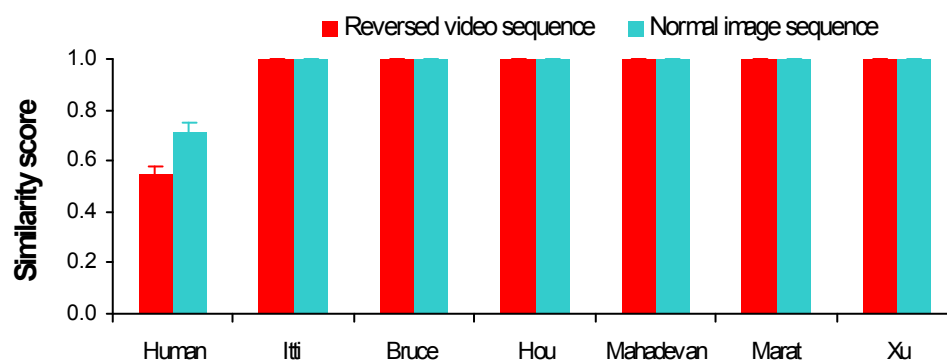
293   Bruce (Bruce & Tsotsos, 2006), Marat (Marat et al., 2009), and Xu (Xu et al., 2013) models (all

294   *ps*<0.01). All the models demonstrated the same predictive performance for the presentation of

295   reversed video sequence, normal image sequence and randomised image sequence (all *ps*>0.01,

296   Fig. 2).

297

298

299   Figure 3. Changes in similarity score over different stimulus presentation sequence. For each

300   human observer and each visual attention model, we computed the similarity score between

301   fixation allocations (actual fixation allocations from human observers, and predicted fixation

302   allocations from the models) in the normal video sequence and in the reversed video sequence or

303   in the normal image sequence. Error bars represent standard deviation.

304

305

306   As clearly shown in Fig. 2, the tested computational visual attention models performed

307   poorly across different video presentation sequences when using actual human gaze allocation as

308   the baseline (or bench mark). To understand the potential causes for the models' inferior

309 performance, we directly examined the impact of presentation sequence manipulation on actual

310 and predicted gaze allocation by using gaze allocation in normal video sequence as the baseline.

311 Specifically, for each human observer and each visual attention model, we computed the

312 similarity score between fixation allocation (actual fixation allocations for human observers, and

313 predicted fixation allocations for the models) in the normal video sequence and in the reversed

314 video sequence, and between fixation allocation in the normal video sequence and in the normal

315 image sequence (Fig. 3). Such analysis could directly inform the extent to which the disruption

316 of temporal frame order (i.e. reversed video sequence) or spatial frame coherence (i.e. normal

317 image sequence) of the visual stream inputs would affect actual and predicted gaze allocation.

318 A 7 (models) × 2 (presentation sequences) ANOVA revealed a significant main effect of

319 model $(F(6,174)=7139.36,\ p<0.001,\ \eta_p^2=0.99)$ and presentation sequence $(F(1,29)=273.28,$

320 $p<0.001,\ \eta_p^2=0.9)$, and a significant interaction between model and presentation sequence

321 $(F(6,174)=273.29,\ p<0.001,\ \eta_p^2=0.9)$. Specifically, manipulating presentation sequence has

322 clearly decreased similarity between actual human fixation distributions in the normal video

323 sequence and in the reversed video sequence or in the normal image sequence, and such effect

324 was more evident in the reversed video than in the normal image sequence (all $ps<0.01$, Fig. 3),

325 suggesting that temporal frame order might play a more important role in determining human

326 gaze allocation in video-watching than spatial frame coherence. All the visual attention models,

327 on the other hand, have predicted the same location of salient region for each video frame

328 regardless of presentation sequence (i.e. indistinguishable high similarity scores when comparing

329 between normal video sequence and reversed video sequence, or between normal video sequence

330 and normal image sequence), suggesting that the crucial role of temporal frame order has been

331 ignored in these computational models.

332

333

334

335

336

337

338

339

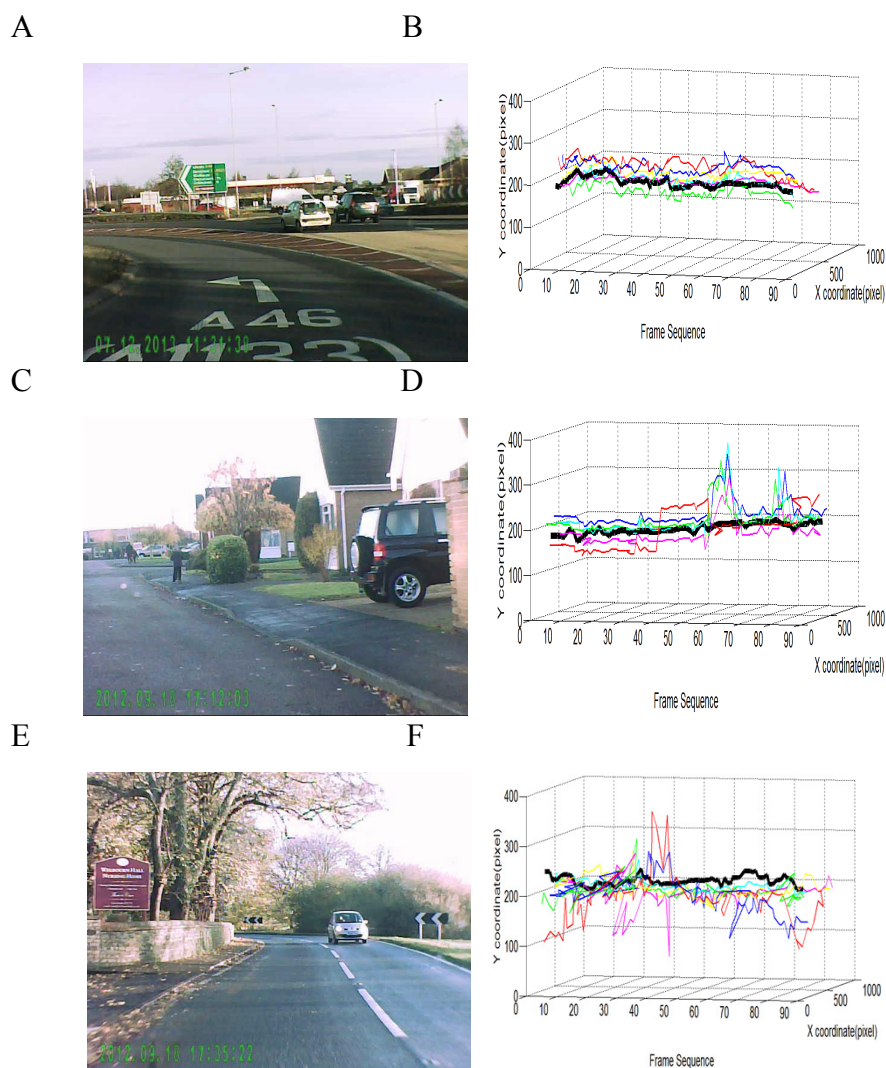



Figure 4. Distribution of actual and predicted gaze allocation for each frame of three typical videos (A and B, C and D, E and F) presented in normal video sequence. The pictures are

348 example frames extracted from videos representing junction approach (A), low-speed (C) and

349 high-speed driving (E). The black lines in B, D, and F represent the averaged co-ordinates of

350 human gaze allocation for each of the 90 frames. The coloured curves represent predicted gaze

351 allocation from the six tested visual attention models.

352

353   When exploring static images, humans tend to look more frequently at the central part of

354 the image, especially at the initial stage of image viewing (Tatler, 2007; Röhrbein et al., 2015).

355 This central fixation bias is mainly due to the image centre providing a convenient starting point

356 to ensure rapid access to every point of interest in the image (Tatler, 2007). As demonstrated in

357 Fig. 4, human observers seemed to show similar central location bias in video watching.

358 Irrespective of video content, the spatial location of their gaze point was less varied across each

359 frame and was more often close to the frame image centre in comparison with the predicted gaze

360 location from the tested visual attention models. To quantify the difference between the spread of

361 the fixations made from humans and models, for each video we averaged both actual and

362 predicted fixation distance from each frame's image centre (Fig. 5). The one-way ANOVA

363 revealed a significant difference between the actual and predicted fixation distributions

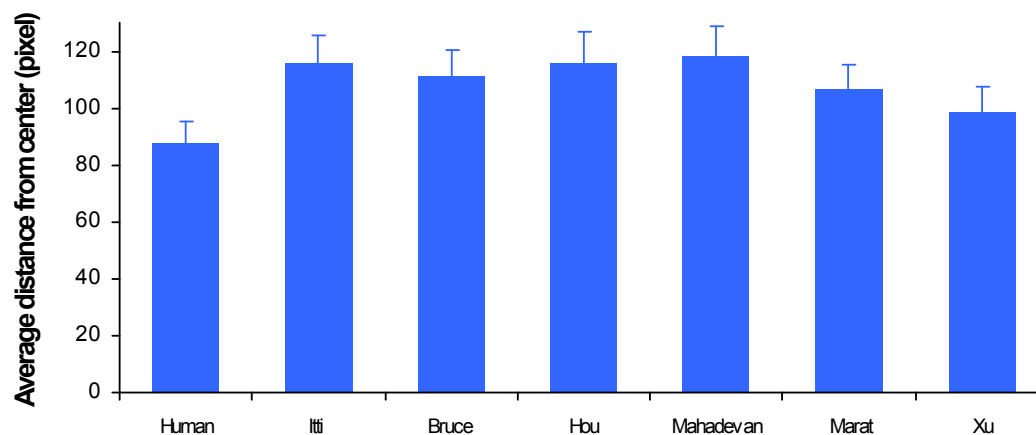
364 ($F(6,174)=3.35$, $p=0.004$, $\eta_p^2=0.11$). Specifically, humans demonstrated a stronger central

365 fixation bias than the visual attention models (all $ps<0.05$), but there was no clear difference in

366 the predicted fixation distance from the image centre across the six tested models (all $ps>0.05$).

367 Clearly, this property of human central fixation bias could (at least) partly account for the

368 inferior predictive performance of current computational models.

369

370 Figure 5. Comparison of actual and predicted fixation distribution in viewing of traffic videos,

371 using metrics of averaged fixation distance from the video frame centre. Error bars represent

372 standard error of the mean.

373

374

375 **Discussion**

376   In this eye-tracking study, we selectively disrupted the spatio-temporal regularity of

377 typical traffic videos, and compared actual human gaze allocation with predicted gaze allocation

378 by state-of-the-art visual attention models. Our analysis clearly demonstrated a high human inter-

379 observer agreement for viewing the original and manipulated video clips (Fig. 2), suggesting a

380 very similar attention allocation strategy across individuals in driving conditions. The tested

381 computational visual attention models, however, performed poorly in comparison with human

382 gaze behaviour. Specifically, the models showed indistinguishable gaze prediction irrespective of

383 stimuli presentation sequence (Fig. 3), suggesting that the spatio-temporal regularity has not been

384 appropriately implemented by these models. Furthermore, human gaze allocation demonstrated a

385 clear central fixation bias that was much stronger than in the gaze distribution predicted by the

386 models (Fig. 5). It seems that the poor performance of visual attention models in processing of

387   dynamic scenes was at least due to the lack of consideration of spatio-temporal regularity and

388   central fixation bias.

389   As an inherent property of dynamic natural scenes, spatio-temporal regularity is routinely

390   exploited by the human visual system to effectively process visual inputs. For instance, in

391   driving situation we frequently expect that a particular object (e.g. a car in the motorway) will be

392   presented at a particular location and time because of the spatial and temporal structure of the

393   current scene, and prior knowledge of the visual world regularities (Guo et al., 2004).

394   Consequently, we can easily direct our attention to the expected object and/or location to

395   perform visual tracking or target analysis with increased perceptual sensitivity (Cunningham &

396   Wallraven, 2009; Guo et al., 2004; Hall et al., 2010), and ignore those objects that have high

397   physical saliency but are not part of spatio-temporal regularity in dynamic scenes. For example,

398   while watching driving videos or actually driving a car, the experienced drivers allocate more

399   fixations to the front and centre view than to the dashboard, right side or left side view

400   (Nabatilan et al., 2012). They also tend to make predictive fixations by looking far ahead and

401   visually targeting objects or locations on the desired trajectory or future path (Kandil, Rotter, &

402   Lappe, 2010; Lappi, Lehtonen, Pekkanen, & Itkonen, 2013).

403   In this study, we further observed that in comparison with normal video sequence,

404   manipulating spatio-temporal regularity by selectively disrupting temporal (i.e. reversed video

405   sequence) or spatial frame order (i.e. normal image sequence) significantly altered human

406   observers' attention allocation (Fig. 3). This effect was more evident in the presentation of

407   reversed video sequences, suggesting a high sensitivity of the human visual attention system to

408   the temporal sequence of predictable dynamic scenes. Such sensitivity could optimize visual

409   processing efficiency and accuracy. Indeed, psychological studies have observed that the

410   judgment of face gender, identity and expression in degraded or noisy presentation conditions is

411   significantly impaired when reversing the facial movement video or altering its natural motion

412   speed (Lander & Bruce, 2000; Hill & Johnston, 2001; Cunningham & Wallraven, 2009).

413   Similarly, reversing the direction of movement impairs object recognition in dynamic settings

414   (Vuong & Tarr, 2004; Wang & Zhang, 2010). Clearly, spatio-temporal regularity plays a crucial

415   role in attention allocation and associated scene perception in natural vision.

416   When exploring static scene images, human natural gaze behaviour often shows a central

417   fixation bias, which can be ascribed to several factors. First, in photography, objects of interest

418   are most often placed in the centre of the image. Second, the image centre as point of fixation

419   could ensure rapid access to every point of interest in the image. It might therefore be

420   advantageous to start scene viewing in the centre. The motor bias, the tendency to make shorter

421   and more horizontal saccades, does not play a big role in central fixation bias (Tatler, 2007). This

422   central fixation bias is often incorporated in the visual attention models for processing static

423   scenes (Borji et al., 2014). When viewing dynamic driving videos, our observers also

424   demonstrated a similar 'central' fixation bias. They tended to allocate attention to the horizontal

425   plane close to eye level, and paid little attention to the model-predicted salient region above or

426   below the eye level. This finding is in agreement with the recent studies demonstrating a

427   tendency to maintain gaze near the centre of the display when watching a wide range of video

428   clips, such as natural or urban scenes, sports, cartoon and film clips (Le Meur et al., 2007; Berg

429   et al., 2009; Tseng et al., 2009; Dorr et al., 2010). It seems that central fixation bias is a

430   spontaneous human gaze behavior that is independent of image/video content and cognitive task.

431   Although using computationally different methods to compute local saliency, all the

432   tested visual attention models performed far below the human baseline, suggesting that they

433   cannot realistically predict human attention allocation in viewing of (at least) dynamic driving

434   clips. Furthermore, for each video frame these models predicted the same local saliency region or

435   object irrespective of stimuli presentation sequence (i.e. normal video sequence vs reversed

436   video sequence vs normal image sequence), indicating that the video's spatio-temporal regularity

437   has not been appropriately   implemented in the computational models. Considering that the

438   human visual system applies spatial-temporal regularity as part of top-down guidance for

439   determining gaze allocation, it seems that this factor, like other top-down factors such as

440   knowledge of scene context and image statistics distribution (Torralba et al., 2006; Judd et al.,

441   2009; Kanan et al., 2009; Goferman et al., 2012), should also be incorporated in a realistic

442   manner into the visual attention model to cope with predictable dynamic visual inputs, possibly

443   through visual motion sensitive neuron models (e.g. directional selective visual neuron models;

444   Rind, 1990) and/or depth sensitive neuron models (Rind & Bramwell, 1996; Gabbiani & Krapp,

445   2006; Yue & Rind, 2006, 2013). It should be noted, however, that the driving video clips just

446   represent part of the visual motion signals in our natural environment. The nature of human

447   driving behaviour may create a strong central fixation bias and subsequently amplify its role in

448   visual saliency computation. The global optic flow associated with driving clips may also disrupt

449   those attention models designed to detect local featural motion. The extent to which we can

450   generalize current findings to other types of motion signals remains to be seen.

451

452   **References**

453   Acik A, Onat S, Schumann F, Einhauser W, Konig P (2009) Effects of luminance contrast and its

454   modifications on fixation behaviour during free viewing of images from different categories.

455   Vision Res 49: 1541–1553.

456   Anderson BA (2013) A value-driven mechanism of attentional selection. Journal of Vision

457   13(3):7, 1-16.

458   Ban S, Lee I, Lee M (2008) Dynamic visual selective attention model. Neurocomputing 71: 853–
459   856.

460   Betz T, Kietzmann TC, Wilming N, Konig P (2010) Investigating task-dependent top-down
461   effects on overt visual attention. J Vis 19(3):15, 1-14.

462   Berg DJ, Boehnke SE, Marino RA, Munoz DP, Itti L (2009) Free viewing of dynamic stimuli by
463   humans and monkeys. J Vis 9(5):19, 1–15.

464   Borji A, Sihite DN, Itti L (2013) Quantitative analysis of human-model agreement in visual
465   saliency modelling: a comparative study. IEEE Trans Image Process 22: 55-69.

466   Borji A, Sihite DN, Itti L (2014) What/where to look next? modeling top-down visual attention
467   in complex interactive environments. IEEE Trans Syst Man Cybern 44: 523-538.

468   Bruce NDB, Tsotsos JK (2006) Saliency based on information maximization. Adv Neural Inf
469   Process Syst 18: 155-162.

470   Carmi R, Itti L (2006) Visual causes versus correlates of attentional selection in dynamic scenes.
471   Vision Res 46: 4333–4345.

472   Cunningham DW, Wallraven C (2009) Dynamic information for the recognition of
473   conversational expressions. J Vis 9(13):7, 1–17.

474   Dorr M, Martinetz T, Gegenfurtner KR, Barth E (2010) Variability of eye movements when
475   viewing dynamic natural scenes. J Vis 10(10):28, 1–17.

476   Einhäuser W, Rutishauser U, Frady EP, Nadler S, König P, Koch C (2006) The relation of phase
477   noise and luminance contrast to overt attention in complex visual stimuli. J Vis 6: 1148–
478   1158.

479   Gabbiani F, Krapp HG (2006) Spike-frequency adaptation and intrinsic properties of an
480   identified looming sensitive neuron. J Neurophysiol 96: 2951-2962.

481   Gavin CJ, Houghton S, Guo K (2015) Dog owners show experience-based viewing behaviour in
482   judging dog face approachability. Psychol Res doi:10.1007/s00426-015-0718-1

483   Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. IEEE Trans
484   Pattern Anal Mach Intell 34:1915-1926.

485   Green D, Swets J (1966) Signal detection theory and psychophysics. New York: John Wiley.

486   Guo K, Mahmoodi S, Robertson RG, Young MP (2006) Longer fixation duration while viewing
487   face images. Exp Brain Res 171: 91–98.

488 Guo K, Nevado A, Robertson RG, Pulgarin M, Thiele A, Young MP (2004) Effects on
489   orientation perception of manipulating the spatiotemporal prior probability of stimuli. Vision
490   Res 44: 2349-2358.

491 Guo K, Smith C, Powell K, Nicholls K (2012) Consistent left gaze bias in processing different
492   facial cues. Psychol Res 76: 263–269.

493 Hall S, Pollux PMJ, Guo K (2010) Exploitation of natural geometrical regularities facilitates
494   target detection. Vision Res 50: 2411-2420.

495 Henderson JM (2007) Regarding scenes. Curr Dir Psychol Sci 16: 219-222.

496 Hou X, Zhang L (2008) Dynamic visual attention: Searching for coding length increments.
497   In *Advances in neural information processing systems* (pp. 681–688). MIT Press.

498 Hill H, Johnston A (2001) Categorizing sex and identity from the biological motion of faces.
499   Curr Biol 11: 880–885.

500 Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual
501   attention. Vision Res 40: 1489–1506.

502 Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene
503   analysis. IEEE Trans Pattern Anal Mach Intell 20: 1254-1259.

504 Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. IEEE
505   12th International Conference on Computer Vision 2106-2113.

506 Kanan C, Tong MH, Zhang L, Cottrell GW (2009) SUN: Top-down saliency using natural
507   statistics. Vis Cogn 17: 979-1003.

508 Kandil F, Rotter A, Lappe M (2010) Car drivers attend to different gaze targets when negotiating
509   closed vs. open bends. J Vis 10(4):24, 1–11.

510 Lander K, Bruce V (2000) Recognizing famous faces: Exploring the benefits of facial motion.
511   Ecol Psychol 12: 259-272.

512 Lappi O, Lehtonen E, Pekkanen J, Itkonen T (2013) Beyond the tangent point: Gaze targets in
513   naturalistic driving. J Vis 13(13):11, 1–18.

514 Le Meur O, Le Callet P, Barba D (2007) Predicting visual fixations on video based on low level
515   visual features. Vision Res 47: 2483–2498.

516 Ma Y, Hua X, Lu L, Zhang H (2005) A generic framework of user attention model and its
517   application in video summarization. IEEE Trans Multimedia l7: 907-920.

518   Mahadevan V, Vasconcelos N (2010) Spatiotemporal saliency in dynamic scenes. IEEE Trans
519   Pattern Anal Mach Intell 32: 171-177.

520   Mannan SK, Ruddock KH, Wooding DS (1996) The relationship between the locations of spatial
521   features and those of fixations made during visual examination of briefly presented images.
522   Spat Vis 10: 165–188.

523   Marat S, Phuoc TH, Granjon L, Guyader N, Pellerin D, Guèrin-Duguè A (2009) Modelling
524   spatio-temporal saliency to predict gaze direction for short videos. Int J Comput Vis 82: 231-
525   243.

526   Nabatilan LB, Aghazadeh F, Nimbarte AD, Harvey CC, Chowdhury SK (2012) Effect of driving
527   experience on visual behavior and driving performance under different driving
528   conditions. Cogn Technol Work 14: 355-363.

529   Parkhurst DJ, Niebur E (2003) Scene content selected by active vision. Spat Vis 16: 125–154.

530   Pollux PMJ, Guo K (2009) Event-related potential correlates of spatiotemporal regularities in
531   vision. NeuroReport 20: 525-530.

532   Pollux PMJ, Hall S, Guo K (2014) Facial expression training optimises viewing strategy in
533   children and adults. PLoS One 9(8), e105418.

534   Pollux PMJ, Hall S, Roebuck H, Guo K (2011) ERP correlates of the interaction between
535   attention and spatiotemporal context regularity in vision. Neurosci 190: 258-269.

536   Reinagel P, Zador AM (1999) Natural scene statistics at the centre of gaze. Network–Comp
537   Neural 10: 341–350.

538   Rind FC (1990) A directionally selective motion-detecting neurone in the brain of the locust:
539   physiological and morphological characterization. J Exp Biol 149: 1–19.

540   Rind FC, Bramwell DI (1996) Neural network based on the input organization of an identified
541   neuron signalling impending collision. J Neurophysiol 75: 967-985.

542   Roebuck H, Bourke P, Guo K (2014) Role of lateral and feedback connections in primary visual
543   cortex in the processing of spatiotemporal regularity? a TMS study. Neurosci 263: 231-239.

544   Röhrbein F, Goddard P, Schneider M, James G, Guo K (2015) How does image noise affect
545   actual and predicted human gaze allocation in assessing image quality? Vision Res 112: 11-
546   25.

547   Rothenstein A, Tsotsos J (2006) Attention links sensing to recognition. Image Vision Comput 26:
548   114-126.

549 Tatler BW (2007) The central fixation bias in scene viewing: Selecting an optimal viewing
550   position independently of motor biases and image feature distributions. J Vis 7(14):4, 1–17.

551 Tatler BW, Hayhoe MM, Land MF, Ballard DH (2011) Eye guidance in natural vision:
552   Reinterpreting salience. J Vis 11(5):5, 1–23.

553 Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye
554   movements and attention in real-world scenes: The role of global features in object search.
555   Psychol Rev 113: 766–786.

556 Tseng PH, Carmi R, Cameron IG, Munoz DP, Itti L (2009) Quantifying center bias of observers
557   in free viewing of dynamic natural scenes. J Vis 9(7):4.

558 Vuong QC, Tarr MJ (2004) Rotation direction affects object recognition. Vision Res 44: 1717–
559   1730.

560 Wang Y, Zhang K (2010) Decomposing the spatiotemporal signature in dynamic 3D object
561   recognition. J Vis 10(10):23, 1–16.

562 Xu J, Yue S, Tang Y (2013) A motion attention model based on rarity weighting and motion
563   cues in dynamic scenes. Int J Pattern Recog 27(6): 1355009.

564 Xu J, Yue S (2014) Mimicking visual searching with integrated top down cues and low-level
565   features. Neurocomputing 133: 1-17.

566 Yue S, Rind FC (2006) Collision detection in complex dynamic scenes using a LGMD based
567   visual neural network with feature enhancement. IEEE Trans. Neural Networks 17: 705–716.

568 Yue S, Rind FC (2013) Redundant neural vision systems - competing for collision recognition
569   roles. IEEE Trans Auton Ment Dev 5: 173-186.

570 Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating
571   diagnostic tests and predictive models. Circulation 115: 654–657.

572
573
574
575