SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
HUMAN BEHAVIOUR
SACCADES
ATTENTION

Received
10 September 2014

Accepted
3 December 2014

Published
19 January 2015

Correspondence and requests for materials should be addressed to S.K. (kitazawa@fbs.osaka-u.ac.jp)

A model of face selection in viewing video stories

Yuki Suda[1] & Shigeru Kitazawa[1,2,3,4]

[1]Department of Neurophysiology, Graduate School of Medicine, Juntendo University, Bunkyo, Tokyo, 113-8421, JAPAN, [2]Dynamic Brain Network Laboratory, Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka, 565-0871, JAPAN, [3]Department of Brain Physiology, Graduate School of Medicine, Osaka University, Suita, Osaka, 565-0871, JAPAN, [4]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, and Osaka University, Suita, Osaka, 565-0871, JAPAN.

When typical adults watch TV programs, they show surprisingly stereo-typed gaze behaviours, as indicated by the almost simultaneous shifts of their gazes from one face to another. However, a standard saliency model based on low-level physical features alone failed to explain such typical gaze behaviours. To find rules that explain the typical gaze behaviours, we examined temporo-spatial gaze patterns in adults while they viewed video clips with human characters that were played with or without sound, and in the forward or reverse direction. We here show the following: 1) the "peak" face scanpath, which followed the face that attracted the largest number of views but ignored other objects in the scene, still retained the key features of actual scanpaths, 2) gaze behaviours remained unchanged whether the sound was provided or not, 3) the gaze behaviours were sensitive to time reversal, and 4) nearly 60% of the variance of gaze behaviours was explained by the face saliency that was defined as a function of its size, novelty, head movements, and mouth movements. These results suggest that humans share a face-oriented network that integrates several visual features of multiple faces, and directs our eyes to the most salient face at each moment.

When typical adults watch TV programs or movie scenes, they show surprisingly stereo-typed gaze behaviours in time and space[1–3], as typically indicated by almost simultaneous shifts of their gazes from one face to another[4]. Recent studies have shown with the aid of multidimensional scaling that such typical temporo-spatial gaze behaviours were not shared by adults with autism[4] or by monkeys[3] and were quite different from those generated artificially based on low-level physical features[3]. These previous findings suggest that temporo-spatial gaze behaviours of typical adults are driven by some human specific social "saliency", which must be different from the standard saliency model that depends solely on the low-level physical features[5,6]. In the present study, we propose a rule that determines the dynamic choice among the faces while viewing motion pictures.

First, we placed a face detection mechanism along a cascade of other factors in our model for face selection (Figure 1a) because typical adults spend most of their time viewing a face, whether they viewed motion pictures[4,7,8] or still pictures[9,10]. However, we also looked at other items in a scene; for example, text typically attracts attention[4,9,11]. Thus, in the first part of this study, we tested whether the essence of actual gaze behaviours was retained after disregarding non-face objects while they viewed short video clips that featured two or more human characters.

We then examined gaze patterns when the video clips were played without sound, to evaluate if verbal semantics plays a critical role in determining our gaze behaviour. If verbal semantics in conversation is critical, our stereo-typed gaze behaviours should be significantly altered by not providing sound. We also presented the video clips in reverse, to further test whether the normal context other than the verbal semantics is critical for the generation of gaze behaviours. If the gaze behaviours are depending solely on the low-level physical features, the behaviours should be symmetric in time reversal.

We then hypothesized that the saliency of each face is determined by a combination of its size, novelty, head movement, and mouth movement (Figure 1a). We chose size because a recent neuroimaging study has shown that responses in the fusiform face area were clearly modulated by the size of the face stimuli[12]. We added the novelty component because the amygdala, which is assumed to be a member of face detection circuits, is reported to respond strongly to novel faces[13–15]. We chose head and mouth movements because a number of face areas over the occipitotemporal cortex are responsive both to biological motion and face

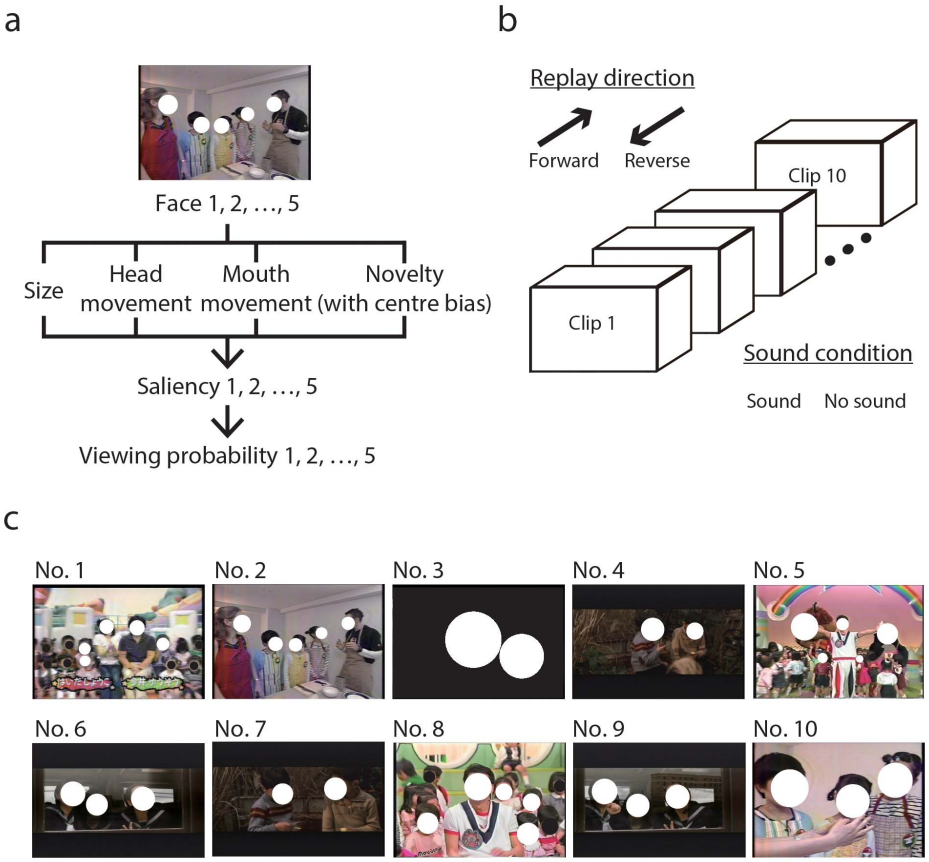

Figure 1 | Designs of the model and experiments. (a) A schematic diagram of the face saliency model. Faces were detected first, and a saliency was assigned to each as a combination of its size, head motion, mouth movement (for speech) and novelty (weighted by the distance from the screen centre). Then, the face saliency was used to predict the viewing probability of each face. (b) Replay conditions. The video stimuli (77 s long) were replayed in four different conditions in a two-by-two factorial manner: replay direction (forward/reverse) by sound condition (on/off). (c) Example frames taken from 10 video clips used for analysis. Up to five faces (circled) were chosen for calculating the viewing proportions. The other faces are covered by black circles. Note that faces were not covered by these circles in the actual experiments. Five clips (no. 1, 2, 5, 8, 10) were taken from TV programmes for young children "Okaasan to Issho" (NHK, Japan Broadcasting Corporation). Four (no. 4, 6, 7, 9) were taken from a film "Always: Sunset on Third Street" (Toho Co., Ltd). Clip no. 3 was taken from a film "A.I.-Artifical Intelligence" (Warner bros., not shown).

stimuli[16]. In addition, movements in not only the mouth but also the head are reported to provide information on the timing of speech and turn-takings in conversation[17,18].

We show here that our gaze behaviours did not much depend on the availability of sound, but were sensitive to time reversal. We further show that our face saliency model explained nearly 60% of the variance of gaze behaviours, and that the weights for the four components were dynamically adjusted depending on the direction of replay.

Results

Comparison of actual and artificial scanpaths. To characterize actual scanpaths recorded from individual participants (a time series of 1736 gaze positions for each participant), we compared them with several artificially generated scanpaths: "peak-face", "random-face", and "physical" saliency scanpaths. In viewing a frame shown in Figure 2a, most participants looked at the teachers in the centre and at the left female teacher in particular. As a result, the "peak-face" scanpath, defined as an artificial scanpath that followed the face with the largest number of gazes, fell on the face of the female teacher (orange dot). In contrast, a "random-face" scanpath, defined as an artificial scanpath that followed a face that was chosen randomly in each frame, fell on a boy in the right (red diamond), and a "physical" saliency scanpath (intensity channel), defined as an artificial scanpath that followed the peak of saliency in terms of low-level physical features defined by intensity, fell on the border of the black hair of a girl in the left against a white background (green square). The "peak" face scanpath seemed to represent the actual gazes of the 24 participants, but other artificial scanpaths did not.

This was generally true over the entire duration of the video stimuli, as indicated by the plots on the MDS plane (Figure 2b). Actual scanpaths (cyan and magenta symbols) clustered together, and the "peak-face" scanpaths fell near the centre of the cluster (orange symbols). In contrast, the "random-face" scanpaths (diamonds) and the "physical" saliency scanpaths (squares) fell in the periphery. Of particular importance, forward replay scanpaths (cyan circles) and reverse replay scanpaths (magenta triangles) formed two distinct clusters on the MDS plane, irrespective of whether the sound was available (filled symbols) or not (open symbols).

These observations were directly confirmed statistically by comparing the mean of within- or across-group distances (Figure 2c; one-way ANOVA, $F_{6,\ 6377} = 2885$, $p < 0.0001$). First, the mean distance between the "peak" face scanpaths and the actual scanpaths (Peak/Actual, 49 ± 20 pixels; mean ± *s.d.*) was significantly smaller than those calculated for the "random-face" (Random/Actual, 158 ± 9 pixels, $p < 0.0001$), and the "physical" saliency scanpaths (Physical/Actual, 157 ± 23 pixels, $p < 0.0001$). In addition, the mean distance between the actual scanpaths and the "peak" face scanpaths (Peak/Actual, 49 ± 20 pixels) was significantly "smaller" than the mean

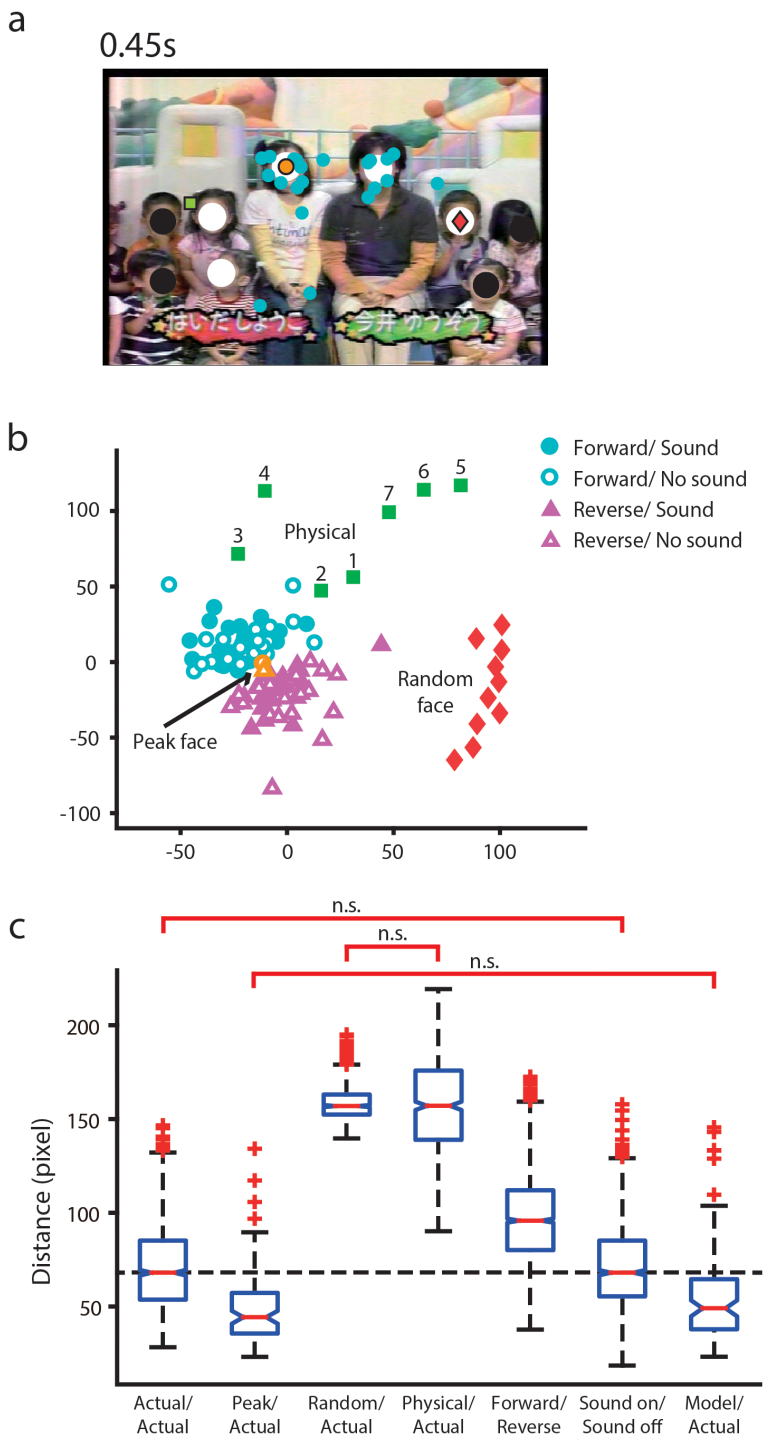

Figure 2 | Comparisons of actual and artificial scanpaths. (a) Actual gaze positions (cyan circles) and model predictions superimposed on a typical frame in Clip No. 1 ("Okaasan to Issho", NHK). Model predictions are represented by an orange dot ("peak" face scanpath), a green square ("physical" saliency scanpath, intensity channel), and a red diamond ("random" face scanpath). Note that faces were not covered by the circles in the actual experiments. (b) Distribution of temporo-spatial gaze patterns in the MDS plane. Each symbol represents a full temporo-spatial scanpath from a single participant (n = 24) under four different replay conditions (cyan: forward replay, magenta: reverse replay, filled: with sound, open: without sound) or an artificially generated scanpath (orange symbols: "peak" face scanpaths, green squares: "physical" saliency scanpaths, red diamonds: "random" face scanpaths). Numbers near green squares discriminate between six low-level feature channels (1: colour, 2: intensity, 3: orientation, 4: contrast, 5: flicker, 6: motion) and all feature channels (7) that were used for generating "physical" saliency scanpaths. (c) Distances between the actual and artificial scanpaths. Box plots show the distances between the actual scanpaths (Actual/Actual), between the "peak" face and the actual scanpaths (Peak/Actual), between the "random" face and the actual scanpaths (Random/Actual), between the "physical" saliency and the actual scanpaths (Physical/Actual), between the actual scanpaths in the forward and those in the reverse replay conditions (Forward/Reverse), between the actual scanpaths in the two replay conditions with and without sound (Sound on/Sound off), and between the actual and predicted scanpaths from the face saliency model (Model/Actual). The one-way analysis of variance showed that the means of the 7 groups were significantly different ($F_{6,\ 6377} = 2885$, $p < 0.0001$). Post-hoc analyses (Ryan's method) showed that the mean was significantly different in all pairs ($p < 0.0001$) except in the three pairs shown in brackets (*n.s.*). One pixel corresponds to 0.05 degrees.

distance between actual scanpaths (Actual/Actual, 71 ± 22 pixels, Figure 2c, $p < 0.0001$). Second, the mean distance between actual scanpaths (Actual/Actual, 71 ± 22 pixels) was not affected by the availability of the sound (Sound on/Sound off, 72 ± 22 pixels, $p = 0.63$). Third, the mean distance between the actual scanpaths (Actual/Actual, 71 ± 22 pixels) became significantly larger when measured between scanpaths in the forward and reverse replay conditions (Forward/Reverse, 97 ± 23 pixels, $p < 0.0001$).

We are able to draw several important implications from these results. First, the "peak-face" scanpath, after disregarding all objects other than faces, still represented the essence of actual gaze behaviours. Second, actual gaze behaviours little depended on verbal semantics that was completely lost when the sound was not provided. Third, gaze behaviours were not symmetric about the time reversal. Our proposed model should take account of all these points.

Application of the face saliency model. To test the extent to which our proposed face saliency model was able to explain the temporal profiles of the face viewing proportions, we fitted the model to each of 40 data sets (10 video clips × 4 replay conditions). While viewing video clip No. 4 that featured two boys taking turns in their conversation (forward replay condition with sound), most participants viewed the left boy at 1.83 s with a peak viewing proportion of 0.8 (Figure 3g, blue dotted line). Then, most of them shifted their gazes to the right boy with a peak viewing proportion of 0.8 at 4.53 s (red dotted line). More than 80% of this typical gaze behaviour (dotted lines) was explained by the face saliency model (solid lines) in that 83% of the variance of the viewing proportion was explained by the model (*d.c.* = 0.83, Figure 3g). In this particular example, the head motion contributed the basic temporal profiles, mouth movement (with ordinary speech sound) added more transients, size provided basic biases, but the novelty component contributed little (Figure 3f, leftmost column, Forward/Sound). The features of temporal profiles and the relative contributions of the four components in the forward replay condition with sound were carried over to the forward replay condition without sound (*d.c.* = 0.78, Forward/No sound, Figure 3f). However, temporal profiles in the reverse replay condition with sound were quite different from those in the forward replay conditions. The contribution from the mouth movement component (with speech sound in reverse) became much larger, and the size component ceased to contribute (*d.c.* = 0.76, Reverse/Sound, Figure 3f). When the video clip was replayed in reverse without sound, temporal profiles were similar to those in the reverse replay condition with sound, but the dynamic range of the viewing proportions had shrunk. As a result, the determination coefficient in this condition was the worst of the four conditions but was still as large as 0.62 (Reverse/No sound, Figure 3f). Taken together, the four-component face saliency model, through adjustments of the relative weights of the four components, captured the key features of the gaze behaviours while viewing Clip 4.

Validity of the four component model. The determination coefficient (*d.c.*) ranged from 0.46 to 0.99 with the mean of 0.79 (Figure 4a, n = 40), and was as large as 0.87 when the *d.c.* was calculated over the entire 10 video clips (Table 1, top row, *d.f.* = 40, *d.c.* = 0.87). The face saliency model was successful in that the model explained 87% of the dynamic changes in the face viewing proportions. When we dropped one of 9 parameters in the four-component face saliency model, the decrease in *d.c.* ranged from 0.01 to 0.17 (2nd to 10th rows in Table 1). However, Akaike's information criterion increased (became worse) in all cases ($\Delta AIC > 1500$). Thus, we judge that all four components and 9 parameters were indispensable, although we think that our model has much room for further improvement, as discussed later.

To further confirm the validity of the four component model, we made a "model-peak-face" scanpath that followed the face with the largest value of viewing proportion predicted by the face saliency model. The mean distance between the "model" peak face scanpath and the actual scanpaths (Model/Actual, 57 ± 27 pixels, Figure 2c) was as small as that between the original "peak" face scanpath and the actual scanpaths (Peak/Actual, 49 ± 20 pixels). These results also indicate that the "model-peak-face" scanpaths were significantly nearer to the actual scanpaths than the actual scanpaths were between themselves.

We further tested whether the estimated parameters can be generalized across different video clips. For this purpose, we used each video clip as test data while using the other 9 as data for model fitting (Fig. 4c). The *d.c.* decreased but was still as large as 0.57 when the gaze positions over the entire 10 video clips were evaluated as a whole.

It is worth noting here that two of the four components, "size" and "head motion", correlates with the zooming of the scene, whereas the other two, "mouth movement" and "novelty", was assigned a value of zero or one irrespective of whether the face was zoomed in or out. Therefore, the model parameters are not necessarily expected to generalize across different scenes with different magnifications. This was further confirmed by fitting the full four-component model to the entire 10 video clips in each of four conditions (the 11th row in Table 1, degrees of freedom = 36). The *d.c.* was still as large as 0.66 but the AIC became much larger ($\Delta AIC = 21209$) than when the model was fitted to each video clip. Thus, the model parameters should be estimated for each scene, and we may not compare the values of estimated parameters across different video clips.

For the purpose of comparison, we introduced the "relative contribution" from each of the four components to the face saliency, as defined by Eq. 4 in Methods. The sum of the relative contributions across the four components was normalized to one (Eq. 5), so that the value can be compared across different conditions and video clips.

Effects of replay and sound conditions on the relative contributions. The relative contributions from the four components were similar on average (Figure 5, $W_s = 0.31 \pm 0.037$; $W_h = 0.25 \pm 0.033$; $W_m = 0.27 \pm 0.030$; $W_n = 0.18 \pm 0.026$; means ± *s.e.m.*). To clarify how the weights across the four components were adjusted depending on the replay conditions with and without sound, we applied the three-way ANOVA to the relative contributions of the four components. As a result, the three main effects (component, replay direction, and the sound condition) were not significant, but an interaction term between the components (size, head motion, mouth movement, and novelty) and the replay direction (forward and reverse) was found to be significant ($F_{3,27} = 3.6, p = 0.026$). Post-hoc tests showed that the relative contribution of novelty was significantly larger in the reverse replay condition than in the forward replay condition ($p = 0.016$, Figure 5). In contrast, the mean relative contribution of the head movement was significantly larger in the forward condition than in the reverse condition ($p = 0.047$, Figure 5).

To summarize, the sound condition did not alter weights across the four components. By contrast, the replay direction dynamically adjusted weights on the novelty and the head motion: the weight on the head motion in the forward replay condition was shifted to the weight on the novelty in the reverse replay condition.

Effects of gestures. As previously shown in Figure 3, we were able to explain nearly 60% of gaze behaviours on average by our face saliency model. The model is by no means perfect with only four components in the model. To look for directions in which to improve, we searched for frames in which there was a large discrepancy between the actual viewing proportion and the value predicted by the present model. The worst three cases are shown in Figure 6. Of importance, in two of the three scenes, the main characters were extending the arms (Figure 6a, d) or raising the right hand (Figure 6b, e). At these occasions, gazes of the participants moved onto the faces of the characters that raised or extended their arms but never onto the

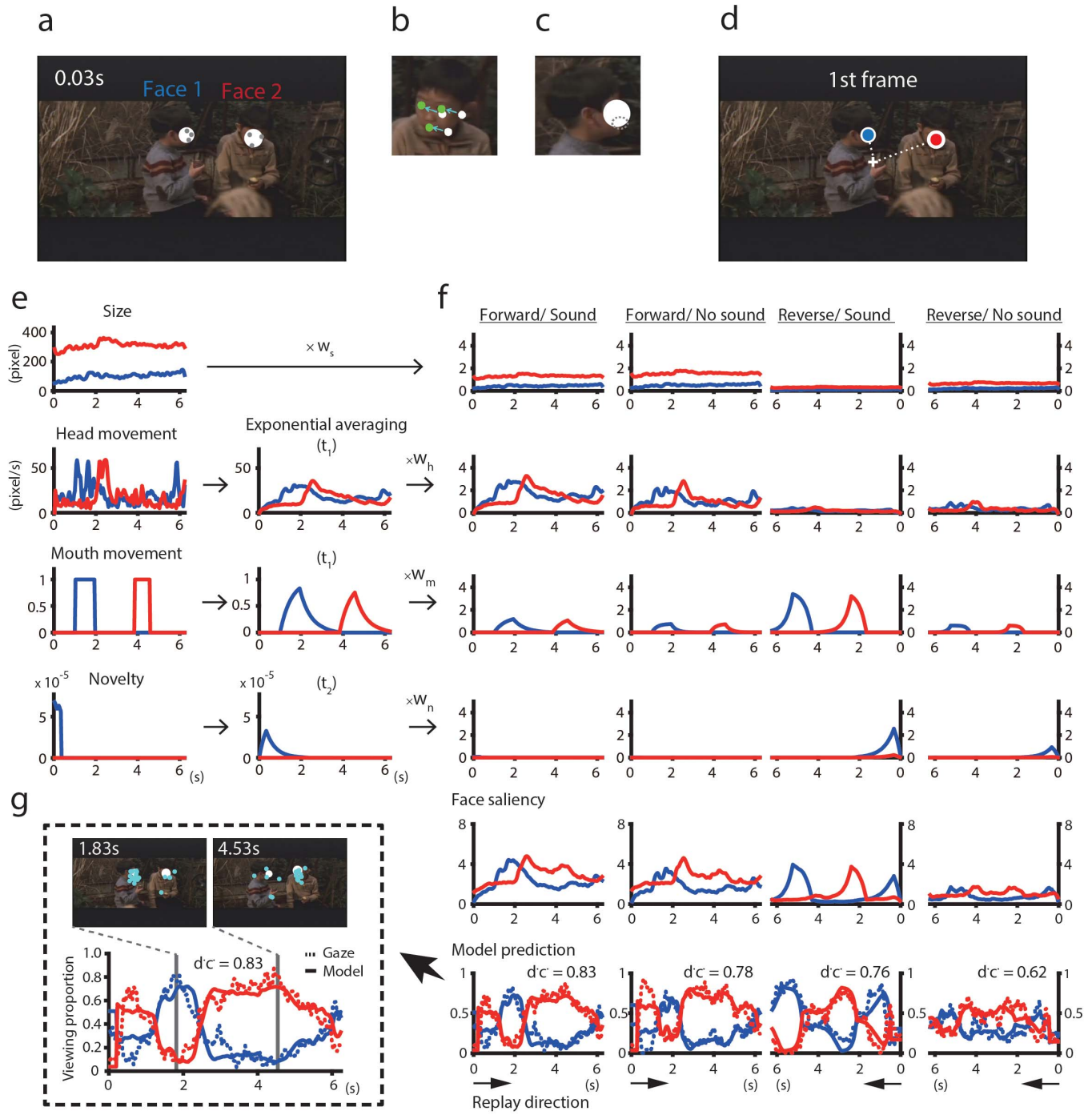

Figure 3 | The face saliency model applied to Clip 4 ("Always: Sunset on Third Street", Toho Co., Ltd). (a–d) The definition of the four components that constituted face saliency. Size (a), head motion (b), mouth movement (c), and novelty (d). The novelty was weighted by the distance from the screen centre (cross in (d)). The eyes, nose, and mouth were registered manually in each frame for each face (dots in (a)). The positions of these face parts were used for calculating the size (a), head motion (b), and distance between the screen centre and the face (d). (e) Examples of the four components were calculated for a boy on the left (blue) and another on the right (red). The exponential average was further applied to the head motion, mouth movement, and novelty components. (f) Fitting of the face saliency model to actual face viewing proportions (broken lines in the bottom panels) obtained under four different replay conditions (one for each column). Face saliency (fifth row) was defined as the weighted sum of the four components (top four rows). Then, the face saliency was transformed into face viewing probability (solid lines in the bottom row) using a logistic function. Determination coefficients (*d.c.*) are shown for each condition. Note that the time axis is inverted in the reverse replay conditions (arrows near the abscissae). (g) Magnification of the results of model fitting in the forward replay conditions with sound. Distributions of the actual gaze positions are shown at two time points when the viewing proportions reached a peak for the left (1.83 s) and the right (4.53 s) boys.

hands or the arms that were actually moved. Body gestures in general are likely to attract our attention to the face of the owner of the body parts in motion.

In the third worst case (Figure 6c, f), the face saliency model assigned a viewing proportion of 25% to a girl (second from the right) when most participants were looking at the girl (> 70%, magenta

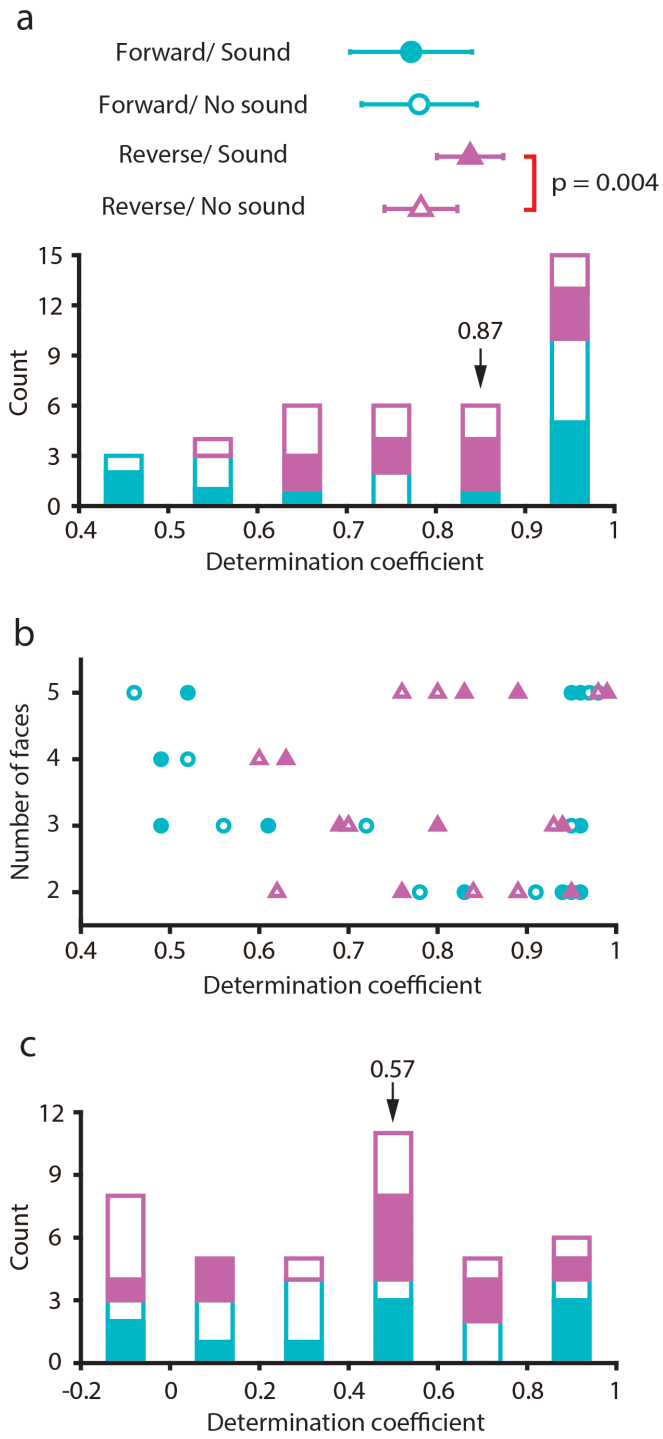

Figure 4 | Determination coefficients compared across replay conditions (a, c) and the number of faces in each video clip (b). Histograms in (a) and (c) show the distribution of the determination coefficient under different replay conditions discriminated by colour (cyan: forward, magenta: reverse) and filling of the box (filled: sound on, open: sound off). In (a) the face saliency model was fitted to data in each of 10 video clips, but in (c) the model was fitted to data in 9 of them and yielded parameters were used for testing data in the other one for cross-validation. Arrows in (a) and (c) show the determination coefficient calculated over the data in 40 video clips. The mean in (a) was compared using a two-way ANOVA (replay × sound) that yielded a significant interaction between the replay direction and the availability of sound ($F_{1,\ 9} = 17.9$, $p = 0.0022$). The bracket shows a pair with a significant difference revealed by post-hoc comparisons ($p = 0.004$, Ryan's method). Error bars show the standard error of the mean. (b) The determination coefficient (abscissa) plotted against the number of faces (ordinate). Note that there was no significant correlation.

dotted line). The scene followed the end of a speech of a man who was standing on the right and was intensely looking at the girl while he was speaking. We speculate that the attention of the man was shared by the participants, and the joint attention increased the saliency of the girl who was supposed to be the next speaker.

Discussion

In the present study, we proposed a model of face selection in viewing video stories to explain stereo-typed gaze behaviours of typical adults as indicated by almost simultaneous shifts of their gazes from one face to another[4]. We have here demonstrated the following: 1) the "peak" face scanpath, which followed the face that attracted the largest number of views but ignored all other objects in the scene, still retained the key features of actual gaze behaviours, 2) gaze behaviours remained unchanged whether the sound was provided or not, 3) the gaze behaviours in the reverse replay condition were distinct from the simple time-reversal of the gaze behaviours in the forward replay condition, and 4) nearly 60% of the variance of gaze behaviours was explained by the face saliency that was defined as a function of its size, novelty, head movements, and mouth movements. We will examine the implications of these key findings.

Strong preference to face. The "peak" face scanpaths were nearer to the actual scanpaths on average than the actual scanpaths were between themselves (Figure 2c). This clearly shows that the "peak" face scanpaths still retained the key features of actual gaze behaviours, though the scanpaths ignored all other objects in the scene, like text as exemplified in Figure 2a. In fact, text, in addition to face, is reported to attract attention of participants who viewed still pictures[9,11]. However, most of the first fixation in viewing a still picture fell on a face, not on text[9]. By using the same video clips that were used in the present study, we previously showed that the viewing rate of text was half as large as that of face even at the peak viewing time of text after presenting a caption[4]. It is likely that the strong preference to face was continuously updated in viewing video clips, in which any one of the faces in the scene never stayed still like a face in a still picture. The strong preference to face in adults with typical development explains why the peak face scanpath retained key features of the actual scanpaths.

It could then be argued that the success of the "peak" face scanpath just relied on the strong preference toward face in general. However, the failure of the "random" face scanpaths rejects this possibility. A random selection of a face among two to five faces yielded scanpaths that were four times farther from the actual scanpaths than the "peak" face scanpath. It is the choice of the right face at the right timing that is essential for reproducing the typical gaze behaviours.

Lack of contribution from verbal semantics. To our surprise, lack of sound did not make much difference in the scanpaths (Figure 2b, c). The most striking example is shown in Figure 3: gaze behaviours in viewing a video clip where two boys took turns in conversation was amazingly similar whether the sound was presented or not (Figure 3f, left two panels in the row at the bottom). In both conditions, approximately 80% of the almost simultaneous shifts of the gazes from one boy to another was reproduced with similar weights on the four components. These results strongly suggest that face selection in our typical gaze behaviours does not primarily depend on verbal information. This agrees with a recent report that participants look more at faces, and especially at talking faces, regardless of auditory conditions[19].

However, we are not insisting that our gaze behaviour primarily depends on low-level physical features in general. On the contrary,

Table 1 | Model comparisons using Akaike's Information Criterion (AIC)

Model			# of model parameters	# of fitting	Degrees of freedom	Determination coefficient	ΔAIC
Face saliency model		Full	9	40	360	0.869	0
	Excluded parameter	Size	8	40	320	0.698	19351
		Head	8	40	320	0.854	2428
		Mouth	8	40	320	0.756	14361
		Novelty	6	40	240	0.811	8258
		Centre bias in Novelty	8	40	320	0.811	8418
		Max prob	8	40	320	0.813	8285
		τ_1	8	40	320	0.852	2786
		τ_2	8	40	320	0.860	1572
		τ_3	8	40	320	0.826	6491
Face saliency model		Full	9	4	36	0.665	21209
Alternative model 1	Added parameter	Centre bias	10	40	400	0.852	2930
Alternative model 2	Added parameter (Physical saliency)	Color	10	40	400	0.871	−310
		Intensity	10	40	400	0.871	−233
		Orientation	10	40	400	0.872	−435
		Contrast	10	40	400	0.869	130
		Flicker	10	40	400	0.869	246
		Motion	10	40	400	0.869	94
		6 channels	10	40	400	0.870	−133

Each model was fitted to each of 10 clips in each of 4 conditions (# of fitting = 40), or to 10 clips as a whole in each of 4 conditions (# of fitting = 4). ΔAIC represents the difference of AIC between each model and the full face saliency model in the top row. Note that the smaller AIC is the better.

we have shown that the typical gaze behaviour cannot be explained by low-level physical features alone: the "physical" saliency scanpaths was markedly different from actual scanpaths, as has been repeatedly shown in recent studies[3,20,21]. It is the size and motion of the "face" (and face parts), or the novel appearance of the "face" that explained our typical gaze behaviours.

Implications of asymmetry in terms of time-reversal. The full-gaze pattern analysis using multidimensional scaling has demonstrated that scanpaths in the forward replay condition were distinct from those in the reverse replay condition (Figure 2b). This indicates that typical gaze behaviours were asymmetric in terms of time-reversal. This asymmetry may not be surprising, assuming that the face saliency model is true. First, the novelty component was asymmetric. What was novel in the forward replay condition at some point in time was never novel at the same time point when the stimuli were played in reverse. Second, the kernels for exponential averaging, which were applied to the head motion, mouth movement, and the novelty component, were asymmetric in time.

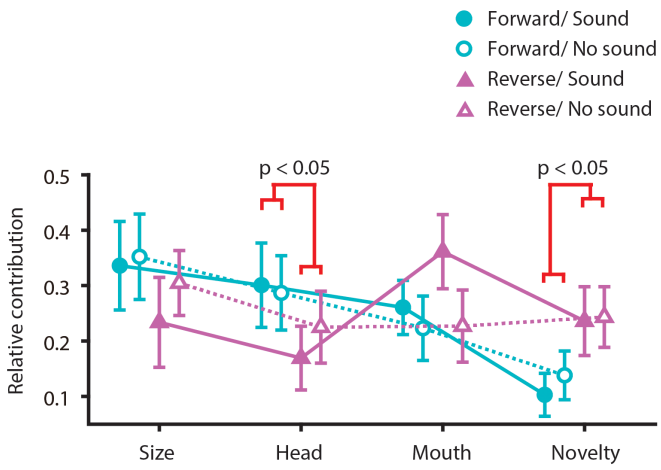

Figure 5 | Effects of replay direction and sound conditions on the relative contributions. The mean relative contributions are plotted with error bars (*s.e.m.*). The sum of four relative contributions within each replay condition was normalized to one. Replay directions are discriminated by colours (cyan: forward, magenta: reverse), and the sound conditions are discriminated by the filling of the symbols (filled: sound on, open: off). Brackets with p-values show results of post-hoc tests after the three-way ANOVA. Note that the relative contribution of head movement was significantly larger in the forward replay conditions than in the reverse replay conditions (head), the relative contribution of novelty was significantly larger in the reverse replay conditions than in the forward replay conditions (novelty).

However, we further found that participants depended significantly more on the head movement when the stimuli were played in the forward direction, but depended significantly more on the novelty component when the stimuli were played in reverse (Figure 5). We speculate that the participants were able to read into certain contexts, such as a timing of turn-taking in speech (e.g. Figure 3), from the head movement when the video was played in the normal direction. It was indeed reported that a linear head movement, termed as the postural shift, occurs just prior to turn-taking[18]. The significant decrease in the weight of head motion in the reverse replay condition suggests that this type of "visual" motion information that predicted timings of turn-takings in conversation was lost to some extent in the reverse replay condition (e.g., see changes in weights for head movements in Figure 3f).

It may then be questioned how the "context" can be read by the brain. One possibility is that a higher region in the cerebral cortex processes the semantics of visual information and provides predicted timings of turn-takings. However, such kind of top-down controls are often slow, and may not be advantageous for the quick and amazingly simultaneous shifts of gazes from one face to another. By contrast, Giese and Possio[22] proposed a simple feed-forward network with some lateral connections, and showed by simulation that the correct sequences of motion can be learned robustly by "motion pattern neurons" with a simple time-dependent hebbian learning rule. Thus, the drop of weight on the head motion in the reverse replay condition can be explained without assuming a higher cognitive mechanism, but just by assuming such sequence selective "motion pattern neurons".

Candidates of other components. The four component facial saliency model failed to reproduce gaze behaviours when a teacher raised her hand (Figure 6b, e) and when a gymnastics instructor

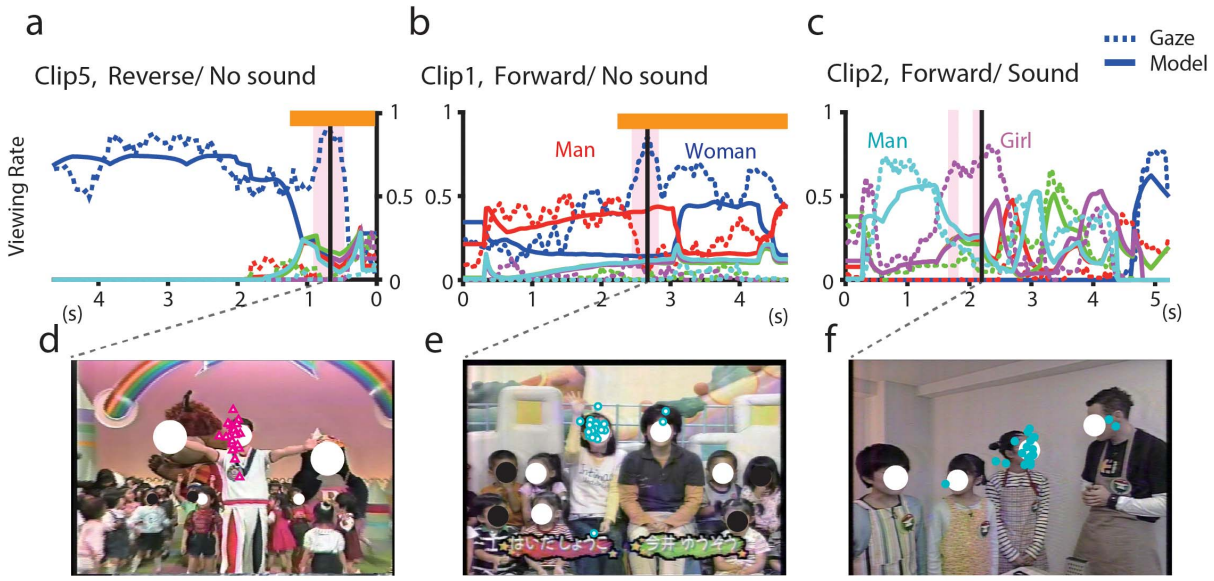

Figure 6 | Effects of gestures on face saliency. Three frames, where the face saliency model (solid lines) yielded the first (a), second (b), and third (c) largest errors from the actual viewing proportion (dotted line), are shown, respectively, in the bottom (d, e, and f; taken from Clips No. 5, 1, and 2, respectively; "Okaasan to Issho", NHK). Shadings show a period when the error exceeded 0.4, and the vertical line shows the timing of peak errors. Note that a gymnastics instructor was opening or closing his arms (d) and a teacher was raising her hand (e), when most participants were looking at their faces. Orange bars show periods when they moved their arms. Note that in (f), most participants were looking at a girl (second from the right) who was intensely looked at by the previous speaker on the right (a tall man). Note that faces were not covered by the white circles in the actual experiments.

opened his arms (Figure 6a, d). Surprisingly, in both cases, participants did not look at the hand or the arms but looked directly at the face of the actors in response to their gestures. We may therefore improve the face saliency model by adding another component that detects body gestures and attributes the gesture to the face of the actor (gesture component).

In the third worst occasion, gazes were clustered on a girl who did not move at all but was apparently making eye contact with a prior speaker and was most likely to be the next speaker (Figure 6c, f). This particular example showed that gazes of the participants were greatly affected by the direction of the gaze of a current speaker to whom participants were attending. The result agrees with the well-known effect of joint attention on gaze behaviours[23]. We may therefore add another component that detects a different target to whom the current target is paying attention (joint attention component).

We may be able to improve the model by adding a centre bias[24], a general tendency to view the centre of the screen, as the fifth component (Alternative model 1, Fig. 7b; 10 parameters). However, the model was no better than the 9 parameter model in terms of d.c. or AIC (Table 1). This may be because the centre bias was already implemented in the novelty component in the 9 parameter model (Eq. 1). In addition, it is probable that the centre bias is apparent only at the beginning of each scene[24] but disappears thereafter.

Another line of improvement may be achieved by adding a physical saliency as a parallel component[25,26] (Alternative model 2, Fig. 7b, 10 parameters). In fact, the parallel model was slightly better than the original model in terms of both d.c. (improvement from 0.869 to 0.872 at best) and AIC, when we used colour (d.c. = 0.871, ΔAIC = -310), intensity (0.871, -233), orientation (0.872, -435), or all channels (0.870, -133). However, the model was slightly worse in terms of the AIC, when we used contrast (0.869, +130), flicker (0.869, +246), or motion (0.869, +94).

Te variance that was left unexplained by the current four component model could be explained by adding these and other components to the definition of face saliency.

Neural correlates of the face saliency. It is often assumed that the subcortical neural circuits, including the superior colliculus, the pulvinar and the amygdala, contribute to face detection[27–29]. The amygdala is also implicated for detection of novelty[13,15] and is thus a good candidate that may contribute to the novelty component.

Of course, a number of face areas in the cerebral cortex should be contributing to the four components. For example, most of the occipitotemporal face areas are responsive both to biological motion and face stimuli (for review, see Ref. 16) and are candidate neural correlates of the head and mouth movement components. Neurons in the superior temporal sulcus in monkeys are reported to respond to articulated head motions[30]. Human homologues of such neurons are candidate neural correlates of the head movement component. As for the size, most studies to date emphasize that neural responses in the cortical face areas are not altered by size[31]. However, a recent neuroimaging study reported that responses in the fusiform face area increased with the size of the face stimuli[12]. A previous study has also reported that the size of the face did affect the responses of some face responsive neurons in the superior temporal sulcus of monkeys[32]. Thus, these temporal face areas are among candidates that would represent the size component.

However, we still do not know anything regarding how multiple faces are represented in these areas, how the four components that seem to be represented over several areas, including subcortical and cortical areas, are combined, or how the face saliency assigned to each face is to be compared among a number of faces. Neural correlates of face saliency warrant further investigations.

Methods

Participants. Twenty-four healthy adults (12 men and 12 women) with a mean age of 23 years (s.d.: 5.2 years) participated. They were normal or corrected to normal in their visual acuity. All experimental protocols were approved by the Ethical Review Board of Juntendo University School of Medicine, and was performed in accordance with the approved guidelines and the Declaration of Helsinki. All participants gave written informed consent before the experiments.

Apparatus, stimuli, and general task designs. Participants were seated on a chair facing a 17-inch TFT flat-screen monitor (Eizo, FlexScan S1701; full screen was used at a resolution of 640 × 480 pixels), which was placed 60 cm away from their eyes. The gaze positions of both eyes were recorded at 50 Hz with a remote eye tracker (Tobii, ×50, Tobii Technology AB) set below the monitor.

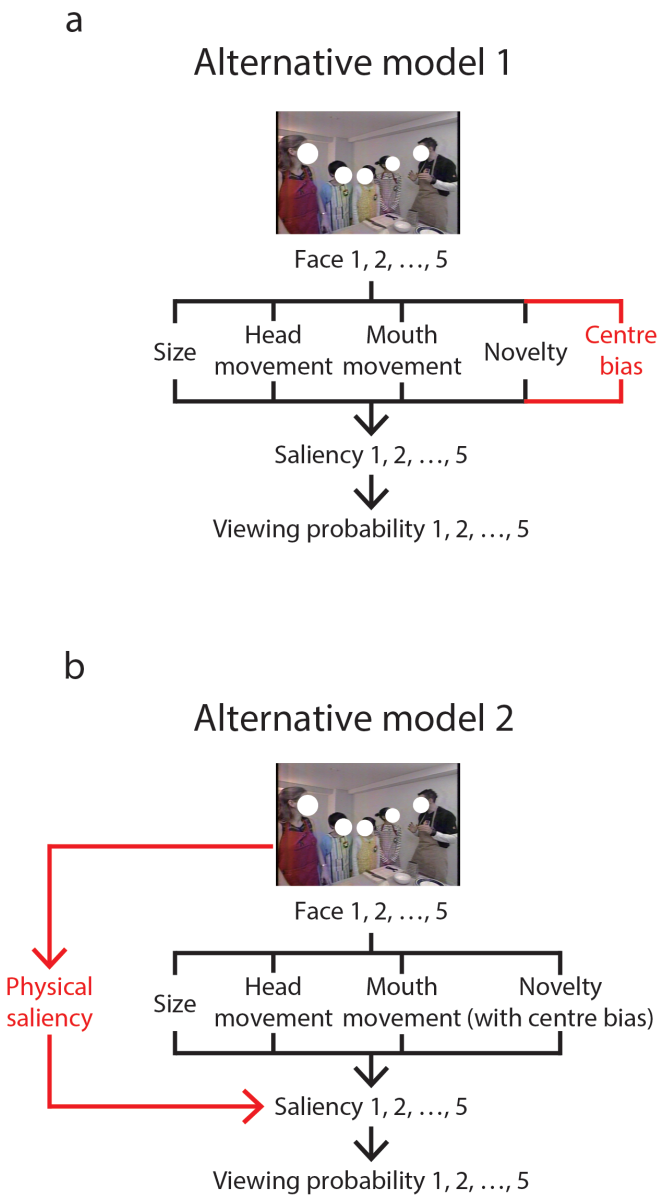

Figure 7 | Alternative models. In (a), the centre bias, a general tendency to view the face in the centre of the screen, was added as the fifth component of the face saliency. In (b), a parallel component representing physical saliency was added to the face saliency. The physical saliency was averaged around each face and added as the fifth component of the face saliency.

Participants were instructed to view the video stimulus presented on the monitor with their heads rested on a chin rest. The video stimulus was prepared in a previous study[4]. It was 77 s long (2237 frames) and consisted of 12 short video clips, each of which lasted for approximately 6 s. A blank of 0.5 s was inserted after each clip.

The stimulus was presented to each participant four times under four different replay conditions: forward replay with or without sound and reverse replay with or without sound (Figure 1b). The order of the four replay conditions was counter-balanced across the 24 participants. Six additional participants were tested but were excluded from the analysis because the percentage of valid data points did not reach the cutoff of 70%.

Data analysis. The gaze positions of the right and the left eyes were averaged to yield a single gaze position for each time point. Data points were included for further analysis only when both eye positions were available. We examined the gaze positions of the participants while they viewed 10 video clips that featured two or more human characters (1739 of 2237 frames, Figure 1c).

Calculation of the viewing proportion for each face. To quantify how many participants viewed a particular face at each frame, we calculated the viewing proportion for each face as described previously[4]. We first identified all faces for each of the 1739 frames. When the number of faces was more than five, we chose the five faces that were nearest to the centre of the screen. Then, we identified face parts, including eyes, nose and mouth, manually for each face. The distance between a gaze position and the registered face parts was then calculated, and a value between zero and one was assigned using a Gaussian function that yielded one at a distance of zero and 0.6 at a distance of 30 pixels (1.5 degrees, one S.D. of the Gaussian function, and three times the spatial resolution of the eye tracker). When the sum of the assigned values over the registered targets exceeded one, each assigned value was divided by the sum. The normalized viewing index of face parts was summed for each face, and this value served as a viewing index of the face. We calculated the viewing proportion of each face by averaging the viewing index of the face over the 24 participants. The viewing proportion for each registered face, which took the value of one when all participants were looking at the particular face, was plotted against the time. The temporal profiles of the face viewing proportions served as dependent variables to be explained by our proposed face saliency model.

Comparison of actual and artificial scanpaths. To test whether the face viewing proportions alone captured the essence of the actual gaze behaviours of the participants, we generated an artificial scanpath that followed the "peak" face that attracted the largest viewing proportion at each frame of the motion pictures. Four "peak-face" scanpaths were prepared, one for each of the four replay conditions. For comparison, we also generated ten "random" face scanpaths that followed a face that was chosen at random. We further generated artificial scanpaths based on saliency in terms of low-level physical features defined by colour, intensity, orientation, contrast, flicker, or motion ("physical" saliency scanpaths) using the Harel method, termed GBVS[6,33]. We used a MATLAB source code that was openly available on the web (http://www.vision.caltech.edu/~harel/share/gbvs.php) and prepared seven "physical" saliency scanpaths based on seven saliency maps, one for each of the six channels and another by combining all channels.

To quantify differences and similarities in the temporo-spatial patterns of the actual (24 participants x 4 replay conditions = 96 scanpaths) and artificial scanpaths (four "peak" face, 10 "random" face, and seven "physical" saliency scanpaths), we used a multi-dimensional scaling technique described previously[4]. Briefly, we directly calculated the absolute distance between a pair of "gaze" points from every pair of scanpaths at each of the 1739 time points. Then, the median was taken as the distance between the two scanpaths to define a between-scanpath distance matrix (117×117; $96 + 10 + 7 + 4 = 117$). With this matrix, we applied multidimensional scaling (MDS) to plot each scanpath in a two-dimensional plane (MDS plane). If the temporo-spatial trajectories were similar for a pair of scanpaths, they would be plotted very near each other and vice versa. We used the MATLAB statistics toolbox (MathWorks) for calculations.

Furthermore, we directly compared the mean distance between the actual scanpaths and each group of artificial scanpaths ("peak" face, "random" face, and "physical" saliency), and the mean distance between actual scanpaths by using one-way analysis of variance and a post-hoc analysis (Ryan's method, Day & Quinn 1989). The comparison also included the mean distances between actual scanpaths in the forward replay conditions and those in the reverse replay conditions, the mean distances between actual scanpaths in the two replay conditions with sound and those without sound, and, finally, the mean distances between the actual scanpaths and those that followed the "peak" face predicted from the face saliency model ("model" peak scanpaths).

Face saliency model. We hypothesized that the saliency of each face was determined by a combination of its size, head motion, mouth movement (for speech) and novelty (weighted by the distance from the screen centre). The size of each face was defined as an area of a triangle formed by two eyes and a mouth (Face 2, Figure 3a); when two eyes were not available in a profile view, one eye, nose and mouth were used (Face 1, Figure 3a). The head motion was defined as the mean of the absolute transition across two neighbouring frames calculated for each of the three points (Figure 3b). The mouth movement was assigned a value of zero or one, according to whether the character was speaking (1) or not (0) (Figure 3c). The novelty was assigned a value of one for 10 frames (0.33 s) when a face appeared for the first time and then was dropped to zero. Most human characters appeared at the beginning of each video clip, and participants generally viewed the area around the centre of the screen during the preceding 0.5-s blank period. Thus, the novelty was further weighted according to the distance from the centre using a Gaussian function that yielded a value of one at the centre and decayed with a standard deviation (σ_d) (Figure 3d). Three of the four time series, the motion, mouth movement, and novelty, were further subjected to exponential averaging with two time constants, one (τ_1) for the motion and the mouth movement, both of which were assumed to be indicators of speech[34], and another (τ_2) for the novelty component that was expected to decay much faster than the indicators of speech (Figure 3e). The size was used without averaging because it was already stable before introducing a time constant. It is worth noting that the novelty depended on the direction of replay: a face was judged as novel when it first appeared in the particular direction of replay. In addition, each face was assigned the same novelty value of one when it appeared for the first time in each of four presentations. That is, the number of repeated presentations (1 to 4) was disregarded. This is not unreasonable because novelty of a visual stimulus is reported to survive even after a number of repeated presentations[35].

We then defined a face saliency (f) for each face as a linear summation of the four components as follows:

$$f_i(t) = w_s size_i\ (t) + w_h head_i\ (t;\tau_1) + w_m mouth_i\ (t;\tau_1) + w_n novelty_i\ (t;\sigma_d,\tau_2), \quad (1)$$

where the subscript i denotes parameters for the i-th face and t denotes the time measured from the beginning of each video stimulus.

We further normalized the face saliency of each face (s_i) by subtracting half of the saliency summed over all faces:

$$s_i(\mathrm{t}) = f_i(t) - \frac{1}{2}\sum\nolimits_i f_i(t). \quad (2)$$

We further assumed that the viewing proportion of the particular face was determined by a logistic function of the normalized saliency. The normalized saliency (s_i) took a value of zero when the saliency of the particular face balanced with the saliency summed over the other faces, in which case the logistic function yielded a value of 0.5. The final formula that predicted a viewing proportion (p) was defined as follows:

$$\mathrm{P}_i(\mathrm{t}) = \mathrm{P}_{max}\frac{1}{1+e^{-s_i(t-\tau_s)}}, \quad (3)$$

where p_{max} denotes the upper limit of the viewing proportion and τ_3 denotes the delay that is required to move the eyes according to the normalized saliency. The MATLAB global optimization toolbox was used to minimise the squared error between the model prediction and the actual temporal profiles of the viewing proportions; a total of 9 parameters (w_s, w_h, w_m, w_n, σ_d, τ_1, τ_2, τ_3, and p_{max}) were adjusted. We repeated this optimization procedure for each video clip (n = 10) for each of the four replay conditions (n = 4). The determination coefficient was calculated from residual errors between the data and the model predictions after each optimization process. We further tested whether the estimated parameters can be generalized across different video clips. For this purpose, we used each video clip as test data while using the other 9 as data for model fitting.

To quantify the relative contributions of the four components (size, head motion, mouth movement, novelty) to the face saliency, we further integrated each term in the right-hand side of Eq. 1 over time and divided each value by the face saliency (left-hand side of Eq. 1) integrated over time. For example, the relative contribution of the size (W_s) was defined as

$$W_s = \sum\nolimits_i \int_0^T w_s size_i(t)dt \Big/ \sum\nolimits_i \int_0^T f_i(t)dt, \quad (4)$$

where T denotes the duration of a video clip. Thus the sum of the relative contributions was normalized to one:

$$W_s + W_h + W_m + W_n = 1. \quad (5)$$

Comparison with other models. We examined whether the four-component model with 9 parameters can be improved by excluding one of them (9 cases), or by adding another component: the centre bias component (Alternative model 1, Fig. 7), or the physical saliency component (Alternative model 2).

The centre bias component was defined as a Gaussian function with the peak in the centre of the screen that decayed with a standard deviation of σ_d. The centre bias consisted the fifth component in the face saliency model as follows:

$$\begin{aligned} f_i(t) = {} & w_s size_i\ (t) + w_h head_i\ (t;\tau_1) + w_m mouth_i\ (t;\tau_1) \\ & + w_n novelty_i\ (t;\tau_2) + w_c centre_i\ (d_i(t),\sigma_d), \end{aligned} \quad (6)$$

where $d_i(t)$ denotes the distance between the i-th face and the centre of the screen at time t. It should be noted that the centre bias that was implemented in the novelty component in the original model (Eq. 4) was removed.

In another model (Alternative model 2, Fig. 7b), the mean physical saliency around the face (within a circular area with the diameter of 30 pixels) was added as the fifth component:

$$\begin{aligned} f_i(t) = {} & w_s size_i\ (t) + w_h head_i\ (t;\tau_1) + w_m mouth_i\ (t;\tau_1) \\ & + w_n novelty_i\ (t;\sigma_d,\tau_2) + w_p phys\_saliency_i\ (t). \end{aligned} \quad (7)$$

We actually tested 7 models by substituting the fifth term with physical saliency defined by colour, intensity, orientation, contrast, flicker, motion, or their combination (Table 1).

To compare these models for their relative goodness of fit, the determination coefficient (*d.c.*) and Akaike's Information Criterion (*AIC*) were calculated for each model. The *d.c.* and *AIC* were calculated using the formula as follows:

$$d.c. = 1 - \mathrm{var}(residual\ error)/\mathrm{var}(horizontal\ error), \quad (8)$$

and

$$AIC = n\ \ln(\mathrm{var}(residual\ error)) + 2k, \quad (9)$$

where n denotes the number of data points (n = 23196) and k denotes the degrees of freedom in each model (k, Table 1). The *d.c.* represents the proportion of the variance of the data explained by the model: it takes the maximum value of one when there are no residual errors. However, the *d.c.* is not suitable for choosing the best model, because the *d.c.* increases with the degrees of freedom. AIC takes not only the residual error but also the degrees of freedom (k) into account: the degrees of freedom are added as a penalty to the natural logarithm of the variance of the residual error. Thus, the model that yields the smallest AIC can be judged as the best model. For the sake of comparison, the AIC difference (ΔAIC) was calculated by subtracting the AIC of the standard face saliency model with 9 parameters (Table 1)[36].

Comparison across replay conditions. To evaluate whether typical gaze behaviours depended on the content of speech and the existence of sound, we compared results of the model fitting across the four different replay conditions. The determination coefficient was compared using a two-way repeated measure ANOVA with the two main factors of the replay condition (forward/reverse) and the existence of the sound (on/off). Then, a three-way ANOVA (component × replay condition × sound) was applied to the relative contribution to test whether there is any dynamic adjustment of the relative contributions (W_s, W_h, W_m, W_n) depending on the availability of the speech context and/or sound.

Analysis of discrepancy. The face saliency model proposed in the present study is by no means perfect and requires improvement. To determine the direction of improvement, we picked the worst three frames of video stimuli, at which the residual error between the model prediction and the actual viewing proportion was the first, second, and the third largest of all frames.

1. Dorr, M., Martinetz, T., Gegenfurtner, K. R. & Barth, E. Variability of eye movements when viewing dynamic natural scenes. *J Vis* **10**, 28, doi:10.1167/10.10.28 (2010).
2. Goldstein, R. B., Woods, R. L. & Peli, E. Where people look when watching movies: do all viewers look at the same place? *Comput Biol Med* **37**, 957–964, doi:10.1016/j.compbiomed.2006.08.018 (2007).
3. Shepherd, S. V., Steckenfinger, S. A., Hasson, U. & Ghazanfar, A. A. Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Curr Biol* **20**, 649–656, doi:10.1016/j.cub.2010.02.032 (2010).
4. Nakano, T. *et al.* Atypical gaze patterns in children and adults with autism spectrum disorders dissociated from developmental changes in gaze behaviour. *Proc R soc B* **277**, 2935–2943, doi:10.1098/rspb.2010.0587 (2010).
5. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat Rev Neurosci* **2**, 194–203, doi:10.1038/35058500 (2001).
6. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **20**, 1254–1259 (1998).
7. Klin, A., Jones, W., Schultz, R., Volkmar, F. & Cohen, D. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch Gen Psychiatry* **59**, 809–816 (2002).
8. Rahman, A., Pellerin, D. & Houzet, D. Influence of number, location and size of faces on gaze in video. *J Eye Mov Res* **7**, 1–11 (2014).
9. Cerf, M., Frady, E. P. & Koch, C. Faces and text attract gaze independent of the task: Experimental data and computer model. *J Vis* **9**, 10 11–15 doi:10.1167/9.12.10 (2009).
10. Cerf, M., Harel, J., Einhäuser, W. & Koch, C. Predicting human gaze using low-level saliency combined with face detection. Paper presented at *Neural Information Processing Systems 2007. Vancouver, Canada* (eds Platt, J. C., Koller, D., Singer, Y. & Roweis, S. T.) in press (Neural Infomation Processing Systems Foundation, Inc.
11. Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S. & Zhao, Q. Predicting human gaze beyond pixels. *J Vis* **14** doi:10.1167/14.1.28 (2014).
12. Yue, X. M., Cassidy, B. S., Devaney, K. J., Holt, D. J. & Tootell, R. B. H. Lower-Level Stimulus Features Strongly Influence Responses in the Fusiform Face Area. *Cerebral Cortex* **21**, 35–47, doi:10.1093/cercor/bhq050 (2011).
13. Blackford, J. U., Buckholtz, J. W., Avery, S. N. & Zald, D. H. A unique role for the human amygdala in novelty detection. *Neuroimage* **50**, 1188–1193, doi:10.1016/j.neuroimage.2009.12.083 (2010).
14. Schwartz, C. E. *et al.* Differential amygdalar response to novel versus newly familiar neutral faces: a functional MRI probe developed for studying inhibited temperament. *Biological Psychiatry* **53**, 854–862, doi:10.1016/s0006-3223(02)01906-6 (2003).
15. Wright, C. I. *et al.* Novelty responses and differential effects of order in the amygdala, substantia innominata, and inferior temporal cortex. *Neuroimage* **18**, 660–669, doi:10.1016/s1053-8119(02)00037-x (2003).
16. Engell, A. D. & McCarthy, G. Probabilistic atlases for face and biological motion perception: an analysis of their reliability and overlap. *Neuroimage* **74**, 140–151, doi:10.1016/j.neuroimage.2013.02.025 (2013).
17. McClave, E. Z. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* **32**, 855–878, doi:10.1016/s0378-2166(99)00079-x (2000).
18. Hadar, U. Gestural modulation of speech production - the role of head movement. *Language & Communication* **9**, 245–257, doi:10.1016/0271-5309(89)90022-0 (1989).
19. Coutrot, A. & Guyader, N. How saliency, faces, and sound influence gaze in dynamic social scenes. *J Vis* **14**, 5, doi:10.1167/14.8.5 (2014).
20. Kano, F. & Tomonaga, M. Perceptual mechanism underlying gaze guidance in chimpanzees and humans. *Anim Cogn* **14**, 377–386, doi:10.1007/s10071-010-0372-3 (2011).

21. Tatler, B. W., Hayhoe, M. M., Land, M. F. & Ballard, D. H. Eye guidance in natural vision: reinterpreting salience. *J Vis* **11**, 5, doi:10.1167/11.5.5 (2011).
22. Giese, M. A. & Poggio, T. Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci* **4**, 179–192, doi:10.1038/nrn1057 (2003).
23. Frischen, A., Bayliss, A. P. & Tipper, S. P. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol Bull* **133**, 694–724, doi:10.1037/0033-2909.133.4.694 (2007).
24. Marat, S., Rahman, A., Pellerin, D., Guyader, N. & Houzet, D. Improving visual saliency by adding 'face feature map' and 'center bias'. *Cogn Comput* **5**, 63–75 (2013).
25. Sugano, Y., Matsushita, Y. & Sato, Y. Appearance-based gaze estimation using visual saliency. *IEEE Trans Pattern Anal Mach Intell* **35**, 329–341, doi:10.1109/TPAMI.2012.101 (2013).
26. Schauerte, B. & Stiefelhagen, R. Predicting human gaze using quaternion DCT image signature saliency and face detection. Paper presented at *2012 IEEE Workshop on Applications of Computer Vision (WACV) Breckenridge, Colorado* 137–144 doi:10.1109/WACV.2012.6163035 (Piscataway, New Jersey, IEEE, 2012, 9 Jan).
27. de Gelder, B., Frissen, I., Barton, J. & Hadjikhani, N. A modulatory role for facial expressions in prosopagnosia. *Proc Natl Acad Sci U S A* **100**, 13105–13110, doi:10.1073/pnas.1735530100 (2003).
28. Johnson, M. H. Subcortical face processing. *Nat Rev Neurosci* **6**, 766–774, doi:10.1038/nrn1766 (2005).
29. Nakano, T., Higashida, N. & Kitazawa, S. Facilitation of face recognition through the retino-tectal pathway. *Neuropsychologia* **51**, 2043–2049, doi:10.1016/j.neuropsychologia.2013.06.018 (2013).
30. Jellema, T. & Perrett, D. I. Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? *Neuropsychologia* **41**, 1728–1737 (2003).
31. Rolls, E. T. The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia* **45**, 124–143, doi:10.1016/j.neuropsychologia.2006.04.019 (2007).
32. Rolls, E. T. & Baylis, G. C. Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp Brain Res* **65**, 38–48 (1986).
33. Harel, J., Koch, C. & Perona, P. Graph-based visual saliency. Paper presented at *20th Annual Conference on Neural Information Processing Systems 2006. Vancouver, British Columbia, Canada* 545–552 (New York, Neural Information Processing Systems (NIPS), 2006, 4 Dec).
34. Hadar, U., Steiner, T. J., Grant, E. C. & Rose, F. C. Head movement correlates of juncture and stress at sentence level. *Lang Speech* **26**, 117–129 (1983).
35. Foley, N. C., Jangraw, D. C., Peck, C. & Gottlieb, J. Novelty enhances visual salience independently of reward in the parietal lobe. *J Neurosci* **34**, 7947–7957, doi:10.1523/JNEUROSCI.4171-13.2014 (2014).
36. Burnham, K. P. & Anderson, D. R. *Model selection and multimodel inference: a practical information -theoretic approach.* 2nd edn, (Springer, 2002).

Acknowledgments

The study was partly supported by Grants-in-Aid for Scientific Research on Innovative Areas (#25119002) and the Health Labour Sciences Research Grant from the Ministry of Health Labour and Welfare to S.K.

Author contributions

Y.S. and S.K. designed the study, collected and analysed the data, and wrote the paper.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Suda, Y. & Kitazawa, S. A model of face selection in viewing video stories. *Sci. Rep.* **4**, 7666; DOI:10.1038/srep07666 (2014).