Integrating audiovisual information for the control of overt attention

Selim Onat
Institute of Cognitive Science, University of Osnabrück, Germany

Klaus Libertus
Institute of Cognitive Science, University of Osnabrück, Germany

Peter König
Institute of Cognitive Science, University of Osnabrück, Germany

In everyday life, our brains decide about the relevance of huge amounts of sensory input. Further complicating this situation, this input is distributed over different modalities. This raises the question of how different sources of information interact for the control of overt attention during free exploration of the environment under natural conditions. Different modalities may work independently or interact to determine the consequent overt behavior. To answer this question, we presented natural images and lateralized natural sounds in a variety of conditions and we measured the eye movements of human subjects. We show that, in multimodal conditions, fixation probabilities increase on the side of the image where the sound originates showing that, at a coarser scale, lateralized auditory stimulation topographically increases the salience of the visual field. However, this shift of attention is specific because the probability of fixation of a given location on the side of the sound scales with the saliency of the visual stimulus, meaning that the selection of fixation points during multimodal conditions is dependent on the saliences of both auditory and visual stimuli. Further analysis shows that a linear combination of both unimodal saliencies provides a good model for this integration process, which is optimal according to information-theoretical criteria. Our results support a functional joint saliency map, which integrates different unimodal saliencies before any decision is taken about the subsequent fixation point. These results provide guidelines for the performance and architecture of any model of overt attention that deals with more than one modality.

Keywords: overt attention, crossmodal integration, natural stimuli, linearity, saliency map, eye movements


Introduction

How are different sources of information integrated in the brain while we overtly explore natural multimodal scenes? It is well established that the speed and accuracy of eye movements in performance tasks improve significantly with congruent multimodal stimulation (Arndt & Colonius, 2003; Corneil & Munoz, 1996; Corneil, Van Wanrooij, Munoz, & Van Opstal, 2002). This supports the claim that sensory evidence is integrated before a motor response. Indeed, recent findings indicate that areas in the brain may interact in many different ways (Driver & Spence, 2000; Macaluso & Driver, 2005). The convergence of unimodal information creates multimodal functionality (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Meredith & Stein, 1986) even at low-level areas traditionally conceived as unimodal (Calvert et al., 1997; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Macaluso, Frith, & Driver, 2000); evidence is also currently mounting for early feedforward convergence of unimodal signals (Foxe & Schroeder, 2005; Fu et al., 2003; Kayser, Petkov, Augath, & Logothetis, 2005; Molholm et al., 2002).

Little is known, however, about integration processes under the relevant operational—that is, natural—conditions. Most importantly, we lack a formal description of the integration process during overt attention. It is important to develop formalizations using behavioral data as it reflects the final outcome of the processes within the CNS.

Current models of overt attention are based on the concept of a saliency map: A given stimulus is first separated into different feature channels; after the local contrast within each feature space is computed, these channels are then combined, possibly incorporating task-specific biases (Itti & Koch, 2001; Koch & Ullman, 1985, Parkhurst, Law, & Niebur, 2002, Peters, Iyer, Itti, & Koch, 2005). The selection of the next fixation point from a saliency map involves a strong nonlinearity. This is typically implemented as a winner-takes-all mechanism. The timing of this nonlinearity with respect to the integration of multiple feature channels crucially
influences both the performance and the structure of the resulting system.

In principle, three idealized multimodal integration schemes can be considered.

Early interaction

The information from different modalities could be integrated early, before the computation of a saliency measure and the selection of a fixation point. Indeed, many studies provide evidence for an interaction between signals in early sensory cortices (Calvert, Campbell, & Brammer, 2000; Kayser et al., 2005). An integration in the form of interaction may be the result of modulatory effects of extramodal signals during computation of saliency maps within a given modality. Or alternatively such an interaction may be caused by multiplicative integration of unimodal saliencies after the saliency maps for each modality are computed. The selection of the most salient point after cross-modal interaction has taken place imposes an expansive nonlinearity. Consequently, sensory signals from different modalities should interact supra-linearly for the control of gaze movements.

Linear integration

Alternatively, saliency could be computed separately for different modalities and subsequently be combined linearly before fixation selection. Recent research shows that the bulk of neuronal operation underlying multisensory integration in superior colliculus can be well described by a summation of unimodal channels (Stanford, Quessy, & Stein, 2005). Within this scheme multimodal saliency would be the result of the linear combination of unimodal saliencies. From an information-theoretical perspective, this type of linear summation is optimal, as is the resulting multimodal map, in the sense that the final information gain is equal to the sums of the unimodal information gains.

Late combination

No true integration between modalities occurs. Instead, overt behavior results from competition between candidate fixation points in the independent unimodal saliency maps. The implementation of such a max operator results in a sublinear integration of the unimodal saliency maps. Although improvements in saccade latencies and fixation accuracies in (nonnatural) multimodal stimulus conditions have been used to support the counterclaim that cross-modal integration does take place (Arndt & Colonius, 2003; Corneil et al., 2002), this hypothesis still warrants investigation under natural free-viewing conditions.

We presented human subjects with lateralized natural sounds and natural images in a variety of conditions and tracked their eye movements as a measure of overt attentional allocation. These measurements were used to compute empirically determined saliency maps, thus allowing us to investigate the above hypotheses.

Materials and methods

Participants and recording

Forty-two subjects (19 males, mean age 23) participated in the experiment. All subjects gave informed written consent but were naive to the purpose of the experiment. All experimental procedures were in compliance with guidelines described in Declaration of Helsinki. Each subject performed only one session.

The experiments were conducted in a small dimly lit room. The subjects sat in a stable chair with back support, facing a monitor. A video-based head-mounted eye tracker (Eye Link 2, SR Research, Ontario, Canada) with sampling rate of 250 Hz and nominal spatial resolution of 0.01° was used for recording eye movements.

For calibration purposes, subjects were asked to fixate points appearing in a random sequence on a 3 × 3 grid by using built-in programs provided with Eye Link (similar to Tatler, Baddeley, & Vincent, 2006). This procedure was repeated several times to obtain optimal accuracy of calibration. It lasted for several minutes, thus allowing subjects to adapt to the conditions in the experimental room. All the data analyzed in the present article were obtained from recordings with an average absolute global error of less than 0.3°.

During the experiment, a fixation point appeared in the center of the screen before each stimulus presentation. The experimenter triggered the stimulus display only after the subject had fixated this point. The data obtained during this control fixation were used to correct for slow drifts of the eye tracker; that is, if drift errors were high a new calibration protocol was started again. Subjects could take a break and remove the headset at any time. In those instances, which occurred rarely, the continuation of the experiment began with the calibration procedure described above.

Stimuli

The multimodal stimuli consisted of images and sounds. Images depicted natural scenes like forests, bushes, branches, hills, open landscapes, close ups of grasses and stones, but also to a limited extent human artifacts, for example, roads, house parts. Some of these stimuli were used in a previous study (Einhauser & König, 2003). The photographs were taken using a 3.3 megapixel digital camera (Nikon Coolpix 995, Tokyo, Japan), down-
sampled to a resolution of 1024 × 768 pixels, and converted to grayscale. The images were displayed on a 21-in. CRT monitor (SyncMaster 1100DF, Samsung Electronics, Suwon, South Korea), at a resolution of 1024 × 768 pixels and with a refresh rate of 120 Hz. The distance of the monitor from the subject’s eyes was 80 cm. The stimuli covered 28° × 21° of visual angle on the horizontal and vertical axes, respectively. The screen was calibrated for optimal contrast and brightness using a commercial colormeter (EyeOne Display, GretagMacBeth, Regensburg, Switzerland).

The natural auditory stimuli were taken from free samples of a commercial Internet Source (Askland Technologies Inc., Victorville, Canada). Overall, 32 different sound tracks were used. All of these were songs of different birds, thus in accordance with the semantic content of the images presented. The auditory stimuli were generated using a sound card (Sound Blaster Audigy EX ZE Platinum Pro, Creative Labs, Singapore) and loudspeakers (Logitech 2.1 Z-3, CA, USA). Loudspeakers flanked both sides of the monitor at a distance of 20 cm, at the same depth plane as the screen. In order to avoid the speakers attracting the attention of the subjects, they were hidden behind black curtains. Sounds were played from the left or right speaker, depending on the experimental condition. The auditory signal amplitude was in the range of 50–70 and 55–75 dB for left and right conditions, respectively. The slight increase was due to the acoustic conditions in the experimental room.

Experimental paradigm

We employed two unimodal conditions (visual and auditory) and one multimodal condition (audiovisual) (Figure 1A). During the visual condition (V), 32 natural images were presented. The auditory condition used 16 × 2 presentations of natural sounds, originating from left (AL) and right (AR) side relative to the subject’s visual field.

During auditory conditions, we presented auditory stimuli jointly with white noise images. These were constructed by shuffling the pixel coordinates of the original images. They lack any spatial structure and as a result do not bias the fixation behavior of the subjects. Obtaining a truly unimodal saliency map for auditory conditions adds some undesirable technical issues. Firstly, due to the operation of the eye tracker and monitor truly zero-light conditions are hard to achieve. Secondly, presenting a dark stimulus leads to more fixations outside the dynamics range of the monitor and eye tracker. Finally, a sudden drastic change in mean luminance of the visual stimulus would introduce nonstationarities in the form of dark adaptation and create a potential confound. To avoid these problems, we presented white noise images of identical mean luminance as the natural pictures.

The multimodal conditions (AVL and AVR) each comprised the simultaneous presentation of 32 auditory and visual stimuli pairs, without any onset or offset asynchrony. In order to balance the stimulus set, a new pairing of audiovisual stimuli was presented on each side.

Figure 1.
Experimental paradigm and subject selection. (A) Five different experimental conditions are used to analyze the effect of unimodal (visual and auditory) and cross-modal (audiovisual) stimuli on overt behavior. Forty-two subjects studied 32 natural images in three different conditions. In the visual condition (V), natural images were presented. In multimodal conditions, the images were paired with localized natural sounds originating from either the left (AVL) or right (AVR) side of the monitor. In the remaining conditions (AL and AR), the sole effect of the localized auditory stimulus was characterized. The duration of stimulus presentation in all conditions was 6 s. (B) Distribution of values of each subject’s σ parameter, which characterizes the horizontal spread of the pdf $p_{xy}(x, y)$ of a given subject s at condition V. $p_{xy}(x, y)$ from two subjects are shown in the inset with corresponding horizontal marginal distributions shown above. Arrows mark the histogram bins to which these subjects belong. The vertical dashed line indicates the threshold σ value required for a given subject to be included in further analysis.
to each subject. Stimuli were shown in a pseudorandom order, with a different permutation used for each subject. Each stimulus was presented for 6 s. A given session contained 128 stimulus presentations and lasted in total for up to 50 min. The only instruction given to the subjects was to watch and listen carefully to the images and sounds. No information about the presence of the speakers at both sides of the monitor or the lateralization of the auditory stimuli was provided to the subjects.

**Data analysis**

We defined fixation points and intervening saccades using a set of heuristics. A saccade was characterized by an acceleration exceeding 8000 deg/s², a velocity above 30 deg/s, a motion threshold of 0.1°, and a duration of more than 4 ms. The intervening episodes were defined as fixation events. The result of applying these parameters was plotted and was visually assessed to check that they produce reasonable results.

**Probability distributions**

From the fixation points of the individual subjects, we built probability density functions (pdf). The first fixation on each stimulus was discarded, as it was always located at the position of the preceding fixation cross in the center of the image. A pdf, \( p_{s,i,c}(x, y) \), for a given subject \( s \), image \( i \), and condition \( c \) was calculated as in Equation 1,

\[
p_{s,i,c}(x, y, t) = \frac{1}{F} \sum_{f=1}^{F} \delta(x - x_f)\delta(y - y_f)\delta(t - t_f),
\]

with \( \delta(x) \) as the discrete Dirac function (the Dirac function is equal to zero unless its argument is zero, and it has a unit integral). \( x_f, y_f, \) and \( t_f \) are the coordinates and time of the \( f \)th fixation. \( F \) is the total number of fixations. We distinguish three different pdfs for a given condition with respect to how these individual pdfs were averaged: subject, image, and spatiotemporal pdfs. Subject pdfs \( p_{s,c}(x, y) \) for a given subject \( s \) and condition \( c \) were built by averaging all the pdfs obtained from a given subject over the images, without mixing the conditions, according to Equation 2,

\[
p_{s,c}(x, y) = \frac{1}{I} \sum_{i} p_{s,i,c}(x, y).
\]

Image pdfs \( (p(x, y)) \) and spatiotemporal pdfs \( (p(x, t)) \) were similarly computed by averaging over the appropriate dimensions. Image pdfs inform us about consistent biases that influence subjects’ gaze and are therefore empirically determined saliency maps specific to a given image. Raw pdfs are matrices of the same size as an image and store fixation counts in their entries. In all cases, these raw pdfs were smoothed for further analysis by convolution. A circular two-dimensional Gaussian kernel was used, which had a unit integral and a width parameter \( \sigma \) with a value of 0.6° unless otherwise stated. This spatial scale is twice as large as the maximal calibration error and maintains sufficient spatial structure for data analysis.

**PDF parameterization**

Several parameters were extracted from the pdfs,

\[
\mu_c = \frac{\sum_x p_c(x) x}{\sum_x p_c(x)},
\]

The center of gravity is measured according to Equation 3 in order to quantify the global shift of subject and image pdfs. \( \mu_c \) is the center of gravity along the \( X \)-axis for condition \( c \), \( p_c(x) \) is the marginal distribution of a pdf along the \( X \)-axis. In condition \( V \), fixation probabilities were usually distributed symmetrically on both sides of the visual field with centralized center of gravity values. This simple statistic successfully quantifies any bias toward the sound location,

\[
\sigma = \left[ \sum_x (x - \mu_V)^2 p_V(x) \right]^{1/2}.
\]

The spread of a given pdf was measured from a subject pdf under condition \( V \) using Equation 4 in order to quantify how explorative that subject’s scanning behavior was. \( \sigma \) is the spread along the \( X \)-axis, \( p_V(x) \) is the marginal distribution along the \( X \)-axis of the subject pdf, and \( \mu_V \) is the center of gravity computed as in Equation 3. The marginal distributions arising from condition \( V \) were well-behaved (Figure 1B inset), thus allowing us to examine the explorative behavior of subjects. Seven of 42 subjects did not engage in explorative viewing during the analysis of the images, resulting in small spread values. These subjects were excluded from further analysis.

The spatiotemporal pdfs, \( p_c(x, t) \), contain information about how the fixation density along the horizontal axis varies over time. We assessed the statistical differences of pdfs from different conditions in a time-localized manner; that is, we obtained a \( p \) value as a function of time for three pairs of conditions (AVL and V; AVR and V; AL and AR). A two-sided Kolmogorov–Smirnov goodness-of-fit hypothesis test in corresponding temporal portions of the pdfs was used with significance level \( \alpha \) set to .001. This was done after binning the probability distribution \( p_c(x, t) \) over the time axis with a bin size of 240 ms, yielding 25 time intervals for comparison per pair of conditions. A temporal interval over which the null hypothesis was rejected in at least in two of the condition pairs was considered as a temporal interval of interest.

The similarity between image pdfs obtained from the same image under different conditions—for example, \( p_c \),
\( v(x, y) \) of image \( i \) and condition \( V \) and \( p_{i,AV}(x, y) \) of the same image and AV conditions (i.e., either AVL or AVR)—was evaluated by computing \( r_{V,AV}^2 \). Before the coefficients were calculated, the AV image pdfs were first normalized to \( p_{i,AV}^N \) according to Equation 5.

\[
p_{i,AV} = \frac{\langle p_V(x) \rangle}{\langle p_{AV}(x) \rangle} \cdot p_{i,AV}(x,y).
\]

This normalization corrects the image pdfs of a given bimodal condition for the global effect of the sound location, so that the expected number of fixations over the horizontal axis is the same over different conditions. The resulting distribution of \( r_{V,AV}^2 \) values was then compared to a control distribution of \( r^2 \) values, which measure the baseline correlation between image pdfs coming from the same condition pairs (i.e., V and AVL, or V and AVR) but differing images.

The Kullback–Leibler divergence, which does not assume any a priori relationship between two distributions, was used to quantify the similarity of the different image pdfs obtained from different conditions. This measure was evaluated according to the following formula, where \( D_{KL}(p_{i,c}, p_{i,c}) \) denotes the Kullback–Leibler divergence measure between two pdfs, \( p_{i,c}(x, y) \), \( p_{i,c}(x, y) \) in bits,

\[
D_{KL}[p_{i,c}(x,y), p_{i,c}(x,y)] = H[p_{i,c}(x,y), p_{i,c}(x,y)] - H[p_{i,c}(x,y)].
\]

The Kullback–Leibler divergence measures the difference between the cross-entropy of two probability distributions and the entropy of one of them. The cross-entropy is always greater than or equal to zero, which allows its usage as a distance measure between two different pdfs. However, unlike other distance measurements, it is not symmetric. Therefore, it is used to measure the distance between the prior and posterior distributions. In comparing different image pdfs, we used the condition V as the prior distribution. One problem we encountered was zero entries in the probability distributions. As the logarithm of zero is not defined, a small constant (\( c = 10^{-9} \)) was added to all entries in the pdf. The precise choice of this constant did not make a difference to the results of our analysis.

**Modeling the cross-modal interaction**

In order to quantify the cross-modal interaction, we carried out a multiple regression analysis. We devised a model (Equation 7) with a cross-product interaction term using smoothed unimodal and multimodal pdfs as independent and dependent variables, respectively,

\[
p_{i,AV} = \beta_{i,0} + \beta_{i,1} \cdot p_{i,V} + \beta_{i,2} \cdot p_{i,A} + \beta_{i,3} \cdot p_{i,VA}.
\]

In Equation 7, \( p_{i,AV}, p_{i,V}, \) and \( p_{i,A} \) are the image pdfs of image \( i \) at audiovisual, visual, and auditory conditions, respectively. The interaction term \( p_{i,VA} \) is supposed to approximate the image pdf that would arise from a multiplicative cross-modal interaction. It is created by the element-wise multiplication of both unimodal image pdfs and renormalized to a unit integral.

The integrative process was further characterized by constructing integration plots. The probability of fixation at each \( x-y \) location was extracted from the 32 image pdfs of visual, auditory, and bimodal conditions, yielding a triplet of values representing the saliency. A given triplet defines the saliency of a given image location in the multimodal condition as a function of the saliency of the same location in both unimodal conditions, represented by the point \( (p_V(x, y), p_A(x, y), p_{AV}(x, y)) \) in the integration plot. These points were irregularly distributed and filled the three-dimensional space unevenly. We discarded the 15% of the values which lay in sparsely distributed regions, and we concentrated instead on the region where most of the data were located. The data points inside this region of interest were then binned, yielding an expected value and variance for each bin. Weighted least square analysis was carried out to approximate the distribution by estimating the coefficients of the following equation:

\[
\frac{p_{AV}}{g(p_{AV})} = \beta_0 + \beta_1 \cdot \frac{p_V}{g(p_V)} + \beta_2 \cdot \frac{p_A}{g(p_A)} + \beta_3 \cdot \frac{p_{VA}}{g(p_{VA})}.
\]

The difference between the above equation and Equation 8 does not take different images into consideration and pools the data over images and space. Additionally, each individual probability value is normalized by its geometric mean \( g(p_i)) \), which normalizes for its individual range thus allowing a direct comparison of the regression coefficients.

**Luminance contrast measurements**

Luminance contrast was computed as the standard deviation of the luminance values of the pixels inside a square patch of about 1° centered at fixation positions. The luminance contrast computed at the fixation points over images and subjects yielded the actual contrast distribution. This distribution was compared to a control distribution to evaluate a potential bias at the fixation points. An unbiased pool of fixation coordinates served as the control distribution—for a given image, this was constructed by taking all fixations from all images other than the image under consideration. This control distribution takes the center bias of the subjects’ fixations into account, as well as any potential systematic effect in our stimulus database (Baddeley & Tatler, 2006; Tatler, Baddeley, & Gilchrist, 2005). The contrast effect at fixation points was computed by taking the ratio of the
average contrast values at control and actual fixations. In order to evaluate the luminance contrast effect over time, the actual and control fixation points were separated according to their occurrences in time. The analysis was carried out using different temporal bin sizes ranging from 100 to 1000 ms.

All analysis was carried out using MatLab (Mathworks, Natick, MA, USA).

Results

First, we analyze the effect of the lateralized auditory stimuli on the fixation behavior of subjects during the study of natural images (Figure 1A). Second, we characterize the temporal interval during which the influence of auditory stimuli is strongest. Third, we demonstrate the specific interaction of visual and auditory information. And finally, we address the predictions derived from the three hypotheses of cross-modal interaction stated above.

Subjects’ gaze is biased toward the sound location

In Figure 2, the spatial distribution of fixation points averaged over images in all 5 conditions is shown for 2 subjects. In the visual condition (V), the fixation density covers a large area and is equally distributed over the left and right half of the screen, with neither of the subjects showing a consistent horizontal bias. The center of gravity ($\mu_V$) is located at 505 and 541 pixels, respectively (white crosses), for these two subjects, in the close vicinity of the center of the screen (located at 512 pixels). In the multimodal conditions (AVL and AVR), both subjects show a change in their fixation behavior. The horizontal distance between $\mu_{AVL}$ and $\mu_{AVR}$ is 221 and 90 pixels for the two subjects, respectively. Thus, in these two subjects, combined visual and auditory stimulation introduces a robust bias of fixation toward the side of sound presentation.

Auditory unimodal conditions (AL and AR) induce different patterns of fixation (Figure 2, right-hand columns, note different scales of color bars). First, despite lateralized sound presentation, a nonnegligible proportion of fixations are located close to the center. Nevertheless, the lateralized sound stimulus induces a shift toward the sound location, even in the absence of any structured, meaningful visual stimulation. In most cases, the shift had an upward component. Furthermore, the off-center fixations are less homogenously distributed and their distribution does not qualitatively resemble the distribution of fixations in the visual condition.

The complete statistics of all subjects and images are shown in Figure 3. The distribution of center of gravity shifts ($\mu_{AVL} - \mu_V$ and $\mu_{AVR} - \mu_V$) for all subjects over all images is skewed (Figure 3A). In most subjects, we observe a moderate effect of lateralized sound presentation. A small group of subjects showed only a small influence; one subject had an extreme effect. The medians of the two distributions are both significantly different from zero (sign test, $p < 10^{-5}$). Crosses indicate the positions of the two example subjects described above. They represent the 70th and 90th percentiles of the distributions. Hence, the complete statistics support the observations reported for the two subjects above.

An analysis of the influence of auditory stimuli on the selection of fixation points in individual images (over all subjects) is shown in Figure 3B. We see that for all visual stimuli the shift in average horizontal position of fixation points is toward the sound location ($t$ test, $p < 10^{-5}$). In both of the panels A and B, the distributions flank both sides of zero, with mean values of $-37$ and 36 pixels for $\mu_{AVL} - \mu_V$ and $\mu_{AVR} - \mu_V$, respectively. Thus, auditory stimulation introduces a robust bias of fixation toward the
Effect is more prominent in the first half of presentation

Next we analyze the temporal evolution of the above-described horizontal effect of the auditory stimulus. Figure 4 depicts the difference between the spatiotemporal fixation pdf for AVR and V, AVL and V, and AL and AR conditions. Comparing visual and multimodal conditions shortly after stimulus onset (Figures 4A and B, lower plots), we observe a rapid shift in fixation density on the side of the auditory stimulus, which is sustained for the presentation period. This increase in fixation density is realized at the expense of fixations in the central region and less so at the expense of fixations on the contralateral side. This can be seen by comparing the marginal distributions (averaged over time) originating from the pdfs used to calculate the difference (Figures 4A and B, upper plots). The intervals of time for which the differences reach significance level (two-sided KS test, $p = .001$) are indicated by vertical black bars. Comparing conditions AL and AR (Figure 4, right panel), we observe an increase in fixation probability on the half of the horizontal axis corresponding to the side of the sound location. This difference decays only slowly over time.

In the light of these results, we define the temporal interval of interest for further analysis as the region where at least
two of the three comparisons were significantly different. This is the time when the gaze is most biased toward the side of auditory stimulation and lasts from 240 to 2640 ms.

**Specific interaction of visual and auditory information**

As a next step, we investigate the integration of auditory and visual information in the temporal interval of interest. The first column of Figure 5 depicts examples of the natural images used in the experiment. The other columns show the image pdfs computed over all subjects under conditions V, AVL, and AVR. As these pdfs were computed over many subjects' fixations, they constitute empirically determined saliency maps. For each image, we observe a characteristic distribution of salient regions, that is, regions with high fixation probabilities. It is important to note that the fixation densities are highly unevenly distributed, suggesting a similarity between

![Figure 5](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933518/)
subjects’ behaviors. We computed the correlation coefficients between image pdfs generated from two subsets of 5 randomly selected subjects (repeating the same analysis 300 times). For all three conditions, the distribution of coefficients over all images and repetitions peaked at around .6. This suggests that different subjects had similar behaviors for scrutinizing the image during the temporal interval of interest (240–2640 ms). It is not clear whether this was the result of the specific image content or shared search strategies between subjects.

The two right-hand columns show the respective distributions obtained in multimodal conditions. First, as noted earlier, the lateralized stimulus causes the center of gravity to shift along the horizontal axis (Figure 5, white crosses). Importantly, the spatial distributions of the spots are alike in different conditions but differ across images. The regions with high fixation probability in one multimodal condition (Figure 5) still effectively attract gaze when the side of auditory stimulation is switched, as well as in the unimodal condition.

This observation is quantified by measuring the Kullback–Leibler divergence (KL divergence) and $r^2$ statistic between saliency maps (such as those in Figure 5) belonging to unimodal and cross-modal conditions. For the examples shown in Figure 5, we obtain a KL divergence of 0.88, 1.32, 0.94, 1.02, and 1.09 bits between V and AVL conditions and 1.38, 1.79, 0.87, 1.09, and 0.79 bits between V and AVR. $r^2$ statistics range from .4 to .97, indicating that a substantial part of the total variance of the multimodal conditions is explained by the distribution of fixation points in the unimodal visual condition.

The distribution of KL divergence values obtained from the complete data set (32 images times 2 auditory stimulus locations) is presented in Figure 6A. The more similar the two pdfs are, the closer the KL divergence values get to zero; zero being the lower limit in the case of identity. This distribution is centered at $1.08 \pm 0.03$ bits (mean $\pm$ SEM). This is significantly different to the mean of the control distribution ($3.45 \pm 0.02$ bits), which was created using 3200 randomly selected nonmatched V–AV pairs. The control distribution provides the upper limit for KL divergence values given our data set. Hence, given the distribution of fixation points on an image in the visual condition, the amount of information necessary to describe the distribution of fixation points on this image in the multimodal conditions is about one third of the information necessary to describe the difference in fixation points on different images in these conditions.

These results are supported by a more conventional linear measure. Figure 6B shows the distribution of actual $r^2$ values calculated between image pdfs from multimodal and unimodal conditions originating from the same image. The distribution is centered at $.71 \pm .13$ and the difference between this measure and a control $r^2$ measure calculated from shuffled image pairs is highly significant ($t$ test, $p < 10^{-5}$). This implies that for most images, the unimodal fixation pdfs account for more than half of the variance in the observed distribution of fixation points in multimodal conditions. Hence, the bias of gaze movements toward the side of the auditory stimulus largely conserves the characteristics of the visual saliency distribution. Therefore, the behavior of the subjects under the simultaneous presence of auditory and visual stimuli is an integration of both modalities.

As a complementary approach, we investigate the effect of multimodal stimuli on the relationship between visual stimulus properties and the selection of fixation points.

![Figure 6](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933518/)
The integration is linear

To quantitatively characterize the integration of unimodal saliencies, we perform a multiple regression analysis. The experiments involved two unimodal auditory conditions (AL and AR) and two corresponding multimodal conditions (AVL and AVR). In order to simplify subsequent discussion, we will use a more general notation of A and AV for unimodal auditory and multimodal conditions, respectively. We model the multimodal (AV) distributions of fixation points by means of a linear combination of unimodal (A and V) distributions and the normalized (to unit integral) product of unimodal distributions (A × V) as indicated in Equation 7 (see Materials and methods). Here we assume that the effect of a multiplicative integration of unimodal saliencies is well approximated by a simple multiplication of their probability distributions. The fits were computed separately for left and right conditions, and the results were pooled for visualization purposes.

The distribution of 64 coefficients for each of the regressors is shown in Figure 7. On average, over all images, the contribution of unimodal visual saliency is largest, with a mean at 0.75 ± 0.15 (mean ± SD; Figure 7, circles). The contribution of unimodal auditory saliency is smaller (0.16 ± 0.12). The coefficient of the cross-product interaction term is, however, slightly negative with mean −0.05 ± 0.10. We repeated the same analysis for a subset of subjects (n = 14, 40%) for whom the lateralized auditory stimulus had the strongest effect on fixations in terms of gravity center shift in the unimodal auditory conditions. In these subjects, the contribution of auditory coefficients was increased (.32 ± .17) at the expense of the visual ones (.53 ± .20), without any apparent effect on the interaction term (−.06 ± .11, t test, p = .59; Figure 7, crosses). In both cases, the intercept was very close but still significantly different from zero. These results suggest that biggest contributions to the multimodal pdfs originate from the linear combinations of the unimodal pdfs.

In a subsequent analysis, we carried out the regression analysis using a different combination of dependent variables and evaluated how the introduction of an additional dependent variable increased the explained variance. Using only unimodal visual pdfs as one regressor, we obtained $r^2$ values having a median value of .72 over all images—as expected from the previous

Several previous studies investigating human eye movements under natural conditions describe a systematic increase of luminance contrast at fixation points (Rainagel & Zador, 1999; Tatler et al., 2005). If the auditory stimuli cause an orientation behavior independent of the visual stimuli, then we can expect the luminance contrast at fixation points to be reduced. If a true integration occurs, we expect this correlation between luminance contrast and probability of fixation to be maintained under multimodal stimulus conditions. Figure 6C shows the ratio of luminance contrast at actual fixations and control locations for the unimodal and both multimodal conditions. Nearly all values (log(actual/control)) are greater than zero, indicating a positive correlation between fixation points and luminance contrast. Moreover, the effect of contrast is constant over the entire presentation, with no systematic difference during the temporal interval of interest (Figure 6C, gray area). Furthermore, the three conditions do not differ significantly in the size of effect. This holds for temporal bin sizes ranging from 100 to 1000 ms (data not shown). Therefore, we can conclude that the additional presence of a lateralized auditory stimulus does not reduce the correlation between the subjects’ fixation points and luminance contrast.
section. Additionally including unimodal auditory pdfs increased the median $r^2$ only slightly by 3%. Repeating this analysis using only the subset of subjects showing the strongest auditory lateralization effect, we obtained a median value of 0.36 with the sole visual regressor. The subsequent introduction of the unimodal auditory pdfs as second dependent variable increased the goodness of fit by 21% over all images. Further including the cross-interaction term as the third dependent variable, the goodness of fit increased slightly, by 5%. Therefore, we can argue that mechanism linearly combining the unimodal saliencies can well account for the observed behavior in the multimodal conditions.

As a model-free approach, we compute integration plots using saliencies obtained from different conditions. It should be noted that no assumptions are made regarding the calculation of the saliency; that is, these saliency values are empirically determined by the gaze locations of many subjects. Integration plots are constructed by plotting the saliency of a given spatial location in the multimodal pdfs as a function of unimodal saliencies of the same spatial location. The specific distribution within this three-dimensional space describes the integration process. In Figures 8A, 8B, and 8C, the height of each surface depicts the corresponding salience of the same location during the multimodal condition as a function of the saliency of the same location in unimodal conditions. The three hypotheses about the integration of auditory and visual information make different predictions (see Introduction): Early interaction leads to a facilitatory effect and an expansive nonlinearity (Figure 8A). The landscape predicted in the case of linear integration is planar and is shown in Figure 8B. Late combination gives rise to a compressive nonlinearity (Figure 8C).

Applying this approach to our complete data set leads to a highly nonuniform distribution of observed unimodal saliencies, when considered in terms of their frequency of occurrence. In Figure 8D, the count matrix of joint occurrences of unimodal saliencies is presented; the surface is very sparsely filled at regions with high values of salience and practically no point is present at regions where both unimodal saliencies are high (top right region of Figure 8D). Statistically reliable statements about the data within these regions are thus not possible. Within this space, we defined a region of interest depicted as a rectangle in Figure 8D; this portion of the space contains 85% of the total number of samples. Inside this region, we bin the data using a 10 x 10 grid and calculate the expected value (Figure 8E) and an error estimate (variance) for each bin.

The relationship between these unimodal and multimodal saliencies is further analyzed using a weighted regression analysis with unimodal saliencies as dependent variables. This yielded $0.54 \pm 0.04$ (95% CI) and $0.59 \pm 0.09$ for the linear contribution of visual and auditory saliency, respectively. Both coefficients were highly significant ($t$ test, $p < 10^{-6}$) except for the intercept coefficient ($t$ test, $p = .23$). $r^2$ is equal to .89 suggesting a good fit. We repeated the same analysis with the interaction term included after normalizing each regressor with its geometric mean in order to have the same exponent range, thus permitting an evaluation of the contribution of different regressors to the multimodal saliency. This yielded $0.57 \pm 0.08$, $0.29 \pm 0.08$, and $0.29 \pm 0.05$ for visual, auditory, and interaction terms, respectively. The linear contributions of unimodal saliencies were highly significant whereas the intercept and the interaction terms were not statistically different to zero.

Using such large bins increases the statistical power within each bin at the expense of detailed structure. We therefore conducted the same analysis using up to 50 bins covering the same region of interest. The $r^2$ of the fitted data at this resolution was .87, ensuring that the fit was still reasonably good. The values of coefficients were practically the same, and the only noticeable change during the incremental increase of the resolution was that the interaction term reached the significance level ($p < .05$) at the resolution of $20 \times 20$, thus demonstrating a slight facilitatory effect. These results support the conclusion that linear integration is the dominating factor in cross-modal integration during overt attention, with an additional small facilitatory component.

The above analysis is influenced by a particular property of the auditory saliency maps. Many fixation densities in the AL and AR conditions are located at the center of the screen (see Figure 4C). We tried to avoid this problem in two different ways. In the first method, we performed the same analysis on the subset of subjects mentioned earlier who were most influenced by the lateralized auditory stimulus, thus minimizing the central bias. Restricting the analysis allowed us to define a new region of interest, which included 90% of the total data points (Figure 8E) and discarded only those points that were very sparsely distributed in high saliency regions. $r^2$ values varied within the range of .81 and .9, decreasing with higher binning resolutions. As above, increasing the number of bins revealed a slight but significant facilitatory effect. Within this subset of subjects, the contribution of auditory saliency ($0.36 \pm 0.08$) was again shown to increase at the expense of the visual contribution ($0.50 \pm 0.08$). Removing the interaction term from the regression analysis caused a maximum drop of only 2.5% in the goodness of fit for all tested bin resolutions within this subset of subjects.

In the second method used to remove the central bias artifact of unimodal auditory pdfs, we took the differences between the left and right auditory conditions; that is, we subtracted the two empirically determined auditory saliency maps to yield difference maps. In each case, the saliency map for the condition where the sound was presented contralaterally was subtracted from the saliency map of the congruent side (i.e., AL–AR for auditory stimulus presented on the left, AR–AL for auditory stimulation from the right). These newly generated maps...
are well behaved and allow the analysis of a larger region of saliency space (90% of total samples). The above analysis was repeated using the difference maps and is shown in Figure 8G. It should be noted that the positive values on the Y-axis are the data points originating from the region of the screen from which the sound emanates, during the temporal interval of interest. We performed separate regression analyses for these halves of the resulting interaction map. In the lower part ($p(A) < 0$), the best predictor was the visual saliencies, as can be seen from the contour lines. In the upper part ($p(A) > 0$), a linear additive model incorporating auditory and visual...
saliencies well approximates the surface. The results derived in an analysis of the effects of different bin sizes were comparable to the above results; that is, a model combining linearly unimodal saliencies along with a slight facilitatory component was sufficient to explain a major extent of the observed data.

We repeated the last analysis with image pdfs obtained with varying degrees of smoothing. Decreasing the width of the convolution kernel systematically reduced the explained variance of the fits on integration plots built within probability ranges containing comparable amount of points. In a large interval of tested parameters (0.4°–0.8°), the main result was conserved; that is, the saliency surface was, to a large extent, captured by a linear combination of unimodal saliencies, with a slight multiplicative effect also evident.

**Discussion**

In this study, we investigated the nature of the multimodal integration during overt attention under natural conditions. We first showed that humans do orient their overt attention toward the part of the scene where the sound originates. This effect lasted for the entire period of the presentation of the stimuli but had a stronger bias during the first half of presentation. More interestingly, this shift was far from a simple orientation behavior—overt behavior during multimodal stimuli was found to be dependent on the saliency of both visual and auditory unimodal stimuli. Although subjects’ fixation points were biased toward the localized auditory stimuli, this bias was found to be dependent on visual information. Our analysis suggests that a predominantly linear combination of unimodal saliencies accounts for the cross-modal integration process.

We quantified the saliency associated with a given image region by analysis of the measured overt eye movements of a large amount of subjects. Subjects’ behavior was similar within the temporal interval where the effect of the lateralized sound was strongest. However, we do not know whether this was the result of a search strategy shared between subjects or whether it originates purely from the bottom-up content present in the image. The results presented here do not depend on the precise determinants of the saliency associated to different image regions. Similarly, we evaluated the saliency of different parts of the visual field associated with the lateralized auditory stimulation. In many subjects, this resulted in a shift of fixation toward the side of the sound, in accord with previous studies showing that sound source location is an important parameter.

Prior studies (Arndt & Colonius, 2003; Corneil & Munoz, 1996; Corneil et al., 2002) have shown that congruent multimodal stimulation during tasks where subjects were required to move their gaze to targets as fast as possible results in faster saccadic reaction times together with an increase in the accuracy of saccades. Here we are extending these results to more operationally relevant conditions by using natural stimuli under free-viewing conditions where the subjects are not constrained in their behavior. Moreover, we are formally describing the cross-modal behavior in terms of unimodal behavior.

Concerning the temporal dynamics of the integration process, we found that the localized sound stimuli attract subjects’ attention more strongly during the first half of presentation, corresponding to an interval of approximately 2.5 s. Although it is observed that the lateralization of fixation density continues throughout the whole presentation time (Figure 4), the effects are much weaker and do not reach the significance level. This effect can be understood as a consequence of inhibition of return, the subject losing interest in that side of the image, or alternatively due to an increasing efficiency of top-down signals over time, resulting in a superior efficiency of the sensory signals to attract attention during early periods of exposure only.

One interesting point is to know whether the present results—a linear integration of auditory and visual saliencies—generalize to situations with a combination of complex visual and complex auditory scenes. In the proposed computational scheme, the origin of visual and auditory saliency maps is not constrained, but measured experimentally. The spatial structure of the auditory salience map is more complex but presumably does not much the spatial acuity of the visual system. As a consequence, in the case several auditory stimuli would contribute no fundamental property in the integration process needs to be changed, and we expect the same integration scheme to hold.

In our study, majority of the natural images we have presented to the subjects were devoid of human artifacts. It could be argued that our auditory stimuli were semantically more congruent with natural scenes where there was no human artifact visible and therefore the cross-modal integration would be stronger. Although some arbitrary decisions has to be taken, we separated our visual stimuli into two classes depending on whether human artifacts were present or not, and we conducted the regression analysis with these two sets separately. We have not found stronger integration in the case of natural images without human artifacts compared to the case where human artifacts were visible.

How do these results fit with current neurophysiological knowledge? One of the most studied structures in the context of cross-modal integration is the superior colliculus (SC), a deep brain structure (Meredith & Stein, 1986; Stein, Jiang, & Stanford, 2004). It has long been known that SC contains neurons that receive inputs from different modalities. Neurons fire more strongly with simultaneous congruent spatial stimulation in different modalities, compared to unimodal firing rates. A recent report
computations involved in the control of overt attention. At the behavioral level, information converges before motor decisions are taken and that this integration is mostly multiplicative. What we have shown is that at the behavioral level the information converges before motor decisions are taken and that this integration is mostly linear. We thus provide boundary constraints on the computations involved in the control of overt attention. The integration of multiple sources of information is also a central issue in models of attention operating in unimodal conditions. As already mentioned, modality-specific information is separated into different feature channels, and the subsequent integration of these different sources is usually subject to arbitrary decisions on the part of the modeler due to the lack of biologically relevant data arising from natural conditions. Whether unimodal feature channels are also linearly integrated is a testable hypothesis and needs further experimental research. One problem we encountered was the centralized fixation density present in the fixation probability distributions of unimodal auditory conditions. Although subjects were effectively oriented by the lateralized sound source, most of their fixations were concentrated at the center of the monitor. We avoided this problem in our analysis by taking the difference of the probability distributions obtained in unimodal auditory left and right conditions, and also by constraining analysis to the subset of subjects whose behavior was most influenced by auditory stimulation. However, we believe that this problem may be alleviated by using multiple sounds simulated to originate from different parts of the image. It is common for complex systems, composed of nonlinear units, to function in a linear way. Neurons are the basic functional constituents of nervous systems and may express highly nonlinear behaviors; for example, the Hodgkin–Huxley equations describing the relation between membrane potential and ionic currents are highly nonlinear. Furthermore, the excitatory and inhibitory recurrent connections within and between cortical areas allow for complex nonlinear interactions. However, irrespective of these underlying nonlinear aspects, many neuronal functions are still well described by linear models. Neurons in early sensory cortices (Schnupp, Mrsic-Flogel, & King, 2001) such as simple cells (Carandini, Heeger, & Movshon, 1997), for example, are well approximated when considered as linear filters operating on input signals. A recent study involving microstimulation in motor cortex showed that signals for movement direction and muscle activation also combine linearly (Ethier, Brizzi, Darling, & Capaday, 2006). We have shown that the cross-modal integration during overt attention process is best described as a linear integration.
of sensory information, possibly originating from different brain areas. In doing so, we have provided an important constraint for any model of cross-modal interaction. This raises an important challenge for any biologically plausible model of human overt attention operating in environments with multiple source of information.

Acknowledgments

We thank Adnan Ghori for his assistance with data acquisition, Cliodhna Quigley for her extensive comments on a previous version of the manuscript, and Alper Acik and Frank Schumann for helpful discussions.

Commercial relationships: none.
Corresponding author: Selim Onat.
Email: sonat@uos.de.
Address: Albrechtstr 28, Osnabru¨ck University, IKW/ NBP, 49069 Osnabru¨ck, Germany.

References


