Perception of differences in naturalistic dynamic scenes, and a V1-based model

Michelle P. S. To Department of Psychology, Lancaster University, Lancaster, UK

Iain D. Gilchrist School of Experimental Psychology, University of Bristol, Bristol, UK

David J. Tolhurst Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK

We investigate whether a computational model of V1 can predict how observers rate perceptual differences between paired movie clips of natural scenes. Observers viewed 198 pairs of movies clips, rating how different the two clips appeared to them on a magnitude scale. Sixty-six of the movie pairs were naturalistic and those remaining were low-pass or high-pass spatially filtered versions of those originals. We examined three ways of comparing a movie pair. The *Spatial Model* compared corresponding frames *between* each movie pairwise, combining those differences using Minkowski summation. The *Temporal Model* compared successive frames *within* each movie, summed those differences for each movie, and then compared the overall differences between the paired movies. The *Ordered-Temporal Model* combined elements from both models, and yielded the single strongest predictions of observers' ratings. We modeled naturalistic sustained and transient impulse functions and compared frames directly with no temporal filtering. Overall, modeling naturalistic temporal filtering improved the models' performance; in particular, the predictions of the ratings for low-pass spatially filtered movies were much improved by employing a transient impulse function. The correlations between model predictions and observers' ratings rose from 0.507 without temporal filtering to 0.759 ($p = 0.01\%$) when realistic impulses were included. The sustained impulse function and the Spatial Model carried more weight in ratings for normal and high-pass movies, whereas the transient impulse function with the Ordered-Temporal Model was most important for spatially low-pass movies. This is consistent with models in which high spatial frequency channels with sustained responses primarily code for spatial details in movies, while low spatial frequency channels with transient responses code for dynamic events.

Introduction

Our perception of the world is a product of the patterns of neural activation across the sensory system. In vision, there has been extensive research examining the pattern of activity of V1 neurons to different types of visual stimuli, from simplistic gratings to more complex natural images. Several computational models have been proposed to simulate how the visual system responds to static natural-image stimuli. Although the coding of motion and temporal change by single neurons has been well studied, it is only relatively recently that the perception of naturalistic movies has been the subject of psychophysical investigation (e.g., Hasson, Nir, Levy, Fuhrmann, & Malach, 2004; Hirose, Kennedy, & Tatler, 2010; Itti, Dhayale, & Pighin, 2003; Smith & Mital, 2013; Troscianko, Meese, & Hinde, 2012; Wang & Li, 2007; Watson, 1998).

Previously, we have shown that a V1-based Visual Difference Predictor (VDP) model, derived from one used to explain the contrast detection thresholds of pairs of sine-wave gratings (Watson & Solomon, 1997), can generate adequate predictions of how observers perceive differences between pairs of static natural scenes (To, Baddeley, Troscianko, & Tolhurst, 2011; To, Lovell, Troscianko, & Tolhurst, 2010; Tolhurst et al., 2010). We presented observers with pairs of photographically derived still images and asked them to rate the perceived difference between the two images. We then built different models that compared the paired images in a head-to-head fashion, and found that they were moderately successful in predicting observers' ratings for a large variety of forms of image

Citation: To, M. P. S., Gilchrist, I. D., & Tolhurst, D. J. (2015). Perception of differences in naturalistic dynamic scenes, and a V1-based model. *Journal of Vision, 15*(1):19, 1–13. http://www.journalofvision.org/content/15/1/19, doi: 10.1167/15.1.19.

doi: 10.1167/15.1.19 Received August 18, 2014; published January 16, 2015 ISSN 1534-7362

difference. The models were most successful when they included physiologically realistic mechanisms, such as nonspecific suppression or contrast normalization (Heeger, 1992), surround suppression (Blakemore & Tobin, 1972), and Gabor receptive fields that are elongated and whose bandwidth changes with best spatial frequency (Tolhurst & Thompson, 1981). A model based on complex cells was more successful than one based on simple cells. Here we ask whether such models can be extended to the perception of natural-image movie stimuli.

These visual discrimination models can generate decent predictions of how well the visual system can discriminate between static images. However, the visual environment is obviously in constant flux and natural scenes are rarely still. It is therefore important to investigate whether such modeling can be extended to dynamic movie clip pairs. Modeling the task of discriminating between two naturalistic movie clips needs extra consideration over the task of discriminating two static images.

First, since movie clips are composed of multiple frames, observers must first process all the frames within each movie separately and must then compare some percept or memory of the frames between the two movies. While comparing static natural scenes requires only one head-to-head computational comparison, in the case of comparing movies that are composed of numerous frames, multiple comparisons need to be made and many comparison rules are possible. It is possible that observers might compare the frames between movies frame by frame; this is suggested by Watson (1998) and Watson, Hu, and McGowan (2001) in their biologically inspired model for evaluating the degree of perceived distortion caused by lossy compression of video streams. Observers might also compare the frames within each movie clip separately before comparing the two clips, perhaps by coding object movement within the movies, as has been suggested for enhanced measures of video distortion (e.g., Seshadrinathan & Bovik, 2010; Vu & Chandler, 2014; Wang & Li, 2007). It is important to identify the best rule to describe how movie frames across movies are compared.

Second, when constructing a V1-based computational model of visual discrimination of dynamic images, one must also obviously consider the temporal sensitivity of the visual system. One hypothesis has it that, in the early stages of the visual system, there are two parallel pathways. The transient or magnocellular (M-cell) pathway acts primarily on low spatial frequencies while the sustained or parvocellular (P-cell) pathway acts primarily on middle and high spatial frequencies (Derrington & Lennie, 1984; Gouras, 1968; Hess & Snowden, 1992; Horiguchi, Nakadomari, Misaki, & Wandell, 2009; Keesey, 1972; Kulikowski & Tolhurst, 1973; Tolhurst, 1973, 1975). The two pathways have been proposed to convey different perceptual information about temporal and spatial structure (Keesey, 1972; Kulikowski & Tolhurst, 1973; Tolhurst, 1973).

In the experiment described here, we presented observers with pairs of stimuli derived from achromatic movies of natural scenes and have asked them to rate how different these clips appear to them. We have examined different derivations of our static V1-based VDP model for static images to accommodate observers' magnitude estimation ratings of dynamic natural scenes. We examine the different contributions of the transient low spatial-frequency channels compared to the sustained high spatial-frequency channels. We also examine different possible strategies for breaking the comparison of two movies into multiple head-to-head comparisons between pairs of frames. Some of the results have been reported briefly (To, Gilchrist, & Tolhurst, 2014; To et al., 2012).

Methods

Observers viewed monochrome movie clips presented on a 19-in. Sony CRT display at 57 cm. The movie frames were 240 pixels square, subtending 12° square in the center of the screen. Apart from that central square, the rest of the 800 × 600 pixels of the display were mid-gray (88 cd.m^{-2}). In the times between movie clips, the whole screen was maintained at that same gray. Presentation was through a ViSaGe system (Cambridge Research Systems, Rochester, UK) so that nonlinearities in the display's luminance output were corrected without loss of bit resolution. The display frame rate was 120 fps (frames per s).

In a single trial, the observers viewed two slightly different movie clips successively, and they were asked to give a numerical magnitude rating of how large they perceived the differences between the clips to be. Each clip usually lasted 1.25 s and there was a gap of 80 ms between the two clips in a trial.

Experimental procedure

There were 198 movie clip pairs (see below), and they were presented once each to each of seven observers. In addition, a standard movie pair, described below, was presented after every ten test-trials; the perceived magnitude difference of this pair was deemed to be "20." Observers were instructed to rate how different pairs of movies appeared to them. They were asked to base their ratings on their overall general visual experience and were not given any specific instructions

on what aspects of the videos (e.g., spatial, temporal, and/or spatio-temporal) they should rely on. They were told that if they perceived the difference between a test movie pair to be less, equal, or greater than the standard pair, their rating should obviously be less, equal, or greater than 20, respectively. They were to use a ratio scale so that, if for instance a given movie pair seemed to have a difference twice as great as that of the standard pair, they would assign a value twice as large to that pair (i.e., 40). No upper limit was set, so that observers could rate large differences as highly as they saw fit. Observers were also told that sometimes movie pairs might be identical, in which case, they should set the rating to zero (in fact, all the movie pairs did differ to some extent).

Before an experiment, each observer underwent a training session when they were asked to rate 50 pairs of movie clips containing various types of differences that could be presented to them later on in the experiment. All movie clips (apart from the standard movie pair) used in demonstration or training phases were made from three original movies that were different from the seven used in the testing phase proper.

The testing phase was divided into three blocks of 66 image pairs. Each block lasted about 15 min and started with the presentation of the standard movie pair, which was subsequently presented after every 10 trials to remind the observers of the standard difference of 20. The image presentation sequence was randomized differently for each observer.

Data collation and analysis

For each observer, their 198 magnitude ratings were normalized by dividing by that observer's average rating. Then, for each stimulus pair, the normalized ratings of the seven observers were averaged together. These 198 averages of normalized ratings were then multiplied by a single scalar value to bring their grand average up to the same value as the grand average of all the ratings given by all the observers in the whole experiment. Our experience (To, Lovell, Troscianko, & Tolhurst, 2010) is that, generally, averages of normalized ratings have a lower coefficient of variation than do averages of raw ratings.

Observers

The experiment tested seven observers, recruited from the student or postdoctoral researcher populations at the University of Cambridge, UK. They all gave informed consent, and had normal vision after prescription correction, as verified using Landolt C acuity chart and the Ishihara color test (10th Edition).

Construction of stimuli

Original movies

The monochrome test stimuli were constructed from six original color video sequences, each lasting 10 s. Three of the stimulus movies were of "environmental movement": e.g., grass blowing, or water rippling or splashing. One was of a duck swimming on a pond. The other two were of people's hands: one person was peeling a potato, and the other was demonstrating sign language for the deaf. A seventh video sequence was used to make a standard video pair.

The video sequences were originally taken at 200 fps with a Pulnix TMC-6740 CCD camera (JAI-Pulnix A/S, Copenhagen, Denmark; Lovell, Gilchrist, Tolhurst, & Troscianko, 2009). Each frame was saved as a separate uncompressed image: 10 bits in each of three color planes, on a Bayer matrix of 640 × 480 pixels. Each frame image was converted to a floating-point 640 × 480 pixel RGB image file with values corrected for any nonlinearities in the luminance versus output relation. Each image was then reduced to 320 × 240 pixels by taking alternate rows and columns (the native pixelation of R and B in the Bayer), and was converted to monochrome by averaging together the three color planes. Finally the images were cropped to 240 × 240 pixels. Some very bright highlights were removed by clipping the brightest image pixels to a lower value.

Stimulus movie clips

From each of the 10-s, 200-fps movies, we took two nonoverlapping sequences of 300 frames: these will be referred to as *clips A* and *B*. For each of the six testing movies, the 300-frame clips A and B were each subject to six temporal manipulations (shown schematically in Figure 1):

A. Alternate frames were simply discarded to leave 150 frames to be displayed at 120 fps (Figure 1A). These had *natural movement*. These clips were the references for pairwise rating comparison with the modified clips below.
B. A *glitch*, or pause, was inserted midway by repeating one frame image near the middle of the movie for a number of frames before jumping back to the original linear sequence of frames (Figure 1B). The duration of the glitch was varied from 83–660 ms between the six movies, and between clips A and B. After discarding alternate frames, the modified clips still lasted 150 frames, and still began and ended on the same frame images.
C. The movie was temporally coarse quantized to make it *judder* throughout by displaying every *n*th frame *n* times (Figure 1C). The degree of temporal coarse quantization (effectively reducing the video clip from 120 fps to 6–25 fps) varied between stimuli

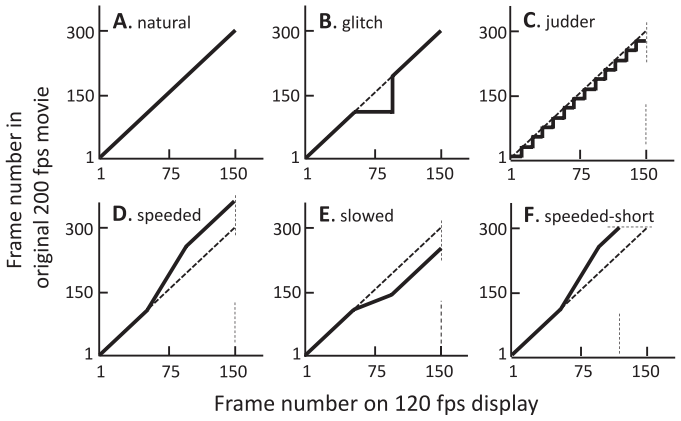

Figure 1. Schematics of the sequences of frames used to make movie clips. Panel A shows natural movement: alternate frames are taken from the original clip and are played back at about half speed. Panels B–F show various ways (described in the text) for distorting the natural movement. Note that method F results in movie clips with fewer frames than the other five methods.

but all clips still lasted 150 frames and still began and ended on the same frame images.

D. The speed of the middle part of the movie was increased by skipping frames (Figure 1D). The degree of skipping (a factor of 1.2–3.7) and the duration (from 200–500 ms) of the speeded part varied between clips and movies. After discarding alternate frames, the *speeded* movies still comprised 150 frames; they began at the same point as the natural movement clips, but they ended further through the original 10 s video sequence.
E. The speed of the middle part of the movie could be reduced, by replicating frames in the middle of the clip (Figure 1E). The degree of replication (speed reduced by factor of 0.25–0.8) and the duration (from 300–660 ms) of the slowed part varied between clips and movies. After discarding alternate frames, the *slowed* movies still comprised 150 frames overall; they began at the same point as the natural movement clips, but they ended at an earlier point in the original sequence.
F. The speed of the middle part of the movie could be increased as in Figure 1D, except that, once normal speed was resumed, frames were added only until the movie reached the same point as in the natural movement controls; thus, these *speeded-short* movies lasted less than 150 frames (Figure 1F).

The magnitudes and durations of the manipulations varied between movies and between clips A and B. Of the five manipulations, the first four (B through E) left the modified clips with the 150 frames. However, in the case of the last (F), there was a change in the total number of frames; these stimuli were excluded from most of our modeling analysis (see Supplementary Materials). For a given original video sequence the frames in all the derived video clips were scaled by a single maximum so that the brightest single pixel in the whole set was 255. In summary, for each original video sequence, we had two clips and five modified variants of each clip.

In the experiment, each clip was paired against its five variants, and the natural clip A was also paired against natural clip B. This resulted in 11 pairs of clips per video sequence, for a total of 66 pairs in the complete experiment. Some examples of movie clip pairs are given in the Supplementary Materials.

Spatial frequency filtering

In addition to the temporal manipulations, each of the 66 pairs of video clips was replicated with (a) low-pass spatial filtering, and (b) high-pass spatial filtering. The frames of each clip (natural and modified) were filtered with very steep filters in the spatial frequency domain (Figure 2A, B). The low-pass filter cut off at 1.08 c/°, while the high-pass filter cut-in at 2.17 c/° but also included 0 c/°, retaining the average luminance of that frame. After filtering, the frames were scaled by the same factor used for the unfiltered clips. Overall, there were thus 3×66 movie-clip pairs used in the experiment, or 198.

Some examples of single frames from spatially filtered and unfiltered movie clips are shown in Figure 2C through E.

Spatio-temporal blending

A fuzzy border (30 pixels, or 1.5° wide) was applied around each edge of each square frame image, to blend it spatially into the gray background. The 30 pixels covered half a Gaussian with spread of 10 pixels. The onset and offset of the movie clip were ramped up and down in contrast to blend the clip temporally into the background interstimulus gray. The contrast ramps were raised cosines lasting 25 frames (208 ms).

Standard stimulus pair

In addition to the 198 test pairs, a standard pair was generated from a seventh video sequence of waves on water. The two clips in this pair differed by a noticeable but not extreme glitch or pause (291 ms) midway in one of the pair. The magnitude of this difference was defined as 20 in the ratings experiment. Observers had to rely on this 20 difference to generate their ratings (see above).

This standard movie pair served as a common anchor and its purpose was to facilitate observers using a same reference point. Its role would be especially important at the beginning of the experiments when

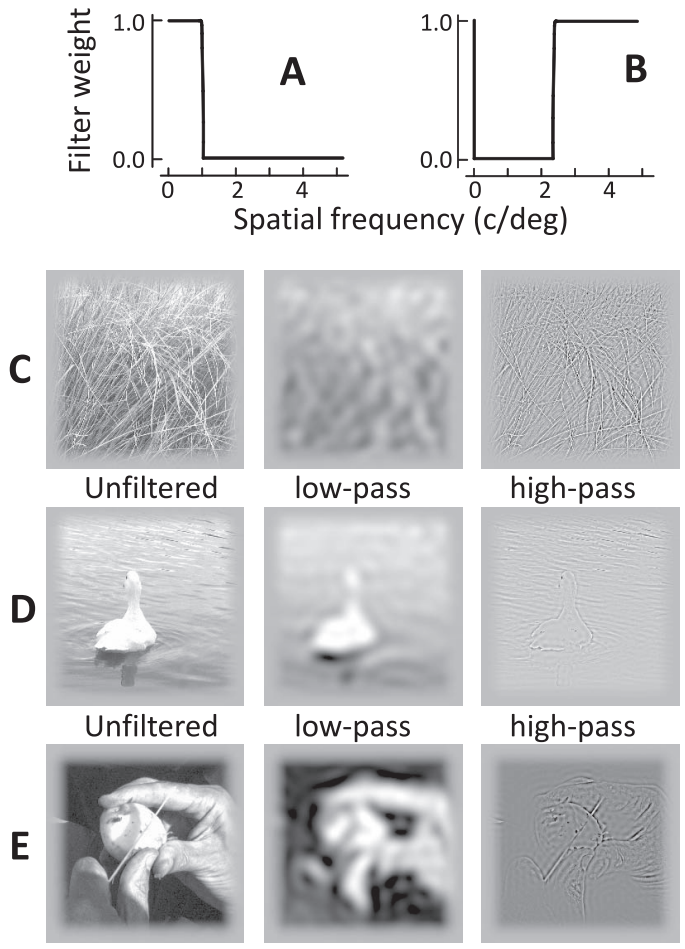

Figure 2. (A) The low-pass spatial filter applied to the raw movie clips to make low-pass movies. (B) The high-pass spatial filter; note that it does retain 0 c/°, the average brightness level. (C–E) Single frames taken from three of the movie families, showing, from left to right, the raw movie, the low-pass version, and the high-pass version.

observers generated an internal rating scale. This would prevent a situation in which, for example, an observer might rate some pairs using some maximum allowed value only to come across subsequent pairs that they perceive to be even more different. This standard pair was made from a different movie clip from the test stimuli, and was repeatedly presented to the observer throughout the experiments in the demonstration, training, and testing phases. Observers were instructed that their ratings of the subjective difference between any other movie pair must be based on this standard pair using a ratio scale, even when the test pairs differed along different stimulus dimensions from the standard pair. Our previous discrimination experiments with static natural images relied on a similar use of ratings based on a standard pair (e.g., To, Lovell, Troscianko, & Tolhurst, 2008; To et al., 2010).

While our choice of the particular standard pair and the particular standard difference may have affected the magnitudes of the observers' difference ratings to all other stimuli, it is unlikely that a different choice might have modified the rank order across the experiment.

Visual Difference Predictor (VDP) modeling of perceived differences

Comparing a single pair of frames

We have been developing a model of V1 simple and complex cells in order to help explain perceived differences between pairs of static naturalistic images (To et al., 2011; To et al., 2010; Tolhurst et al., 2010). This model is inspired by the cortex transform of Watson (1987) as developed by Watson and Solomon (1997). The model computes how millions of stylized V1 neurons would respond to each of the images under comparison; the responses of the neurons are then compared pairwise to give millions of tiny difference cues, which are combined into a single number by Minkowski summation. This final value is a prediction of an observer's magnitude rating of the difference between those two images (Tolhurst et al., 2010). For present purposes, we are interested only in the monochrome (luminance contrast) version of the model.

Briefly, each of the two images under comparison is convolved with 60 Gabor-type simple cell receptive field templates: six optimal orientations, five optimal spatial frequencies, and odd- and even-symmetry (each at thousands of different locations within the scene). These linear interim responses in luminance are converted to *contrast* responses by dividing by the local mean luminance (Peli, 1990; Tadmor & Tolhurst, 1994, 2000). Complex cell responses are obtained by taking the rms of the responses of paired odd- and even-symmetric simple cells; To et al. (2010) found that a complex-cell model was a better predictor than a simple-cell model for ratings for static images. The complex-cell responses are weighted to match the way that a typical observer's contrast sensitivity depends upon spatial frequency. The response versus contrast function becomes sigmoidal when the interim complex cell responses are divided by the sum of two normalizing terms (Foley, 1994; Tolhurst et al., 2010). First, there is nonspecific suppression (Heeger, 1992) where the response of each neuron is divided by the sum of the responses of all the neurons whose receptive fields are centered on the same point. Second, there is orientation- and spatial-frequency-specific suppression by neurons whose receptive fields surround the field of the neuron in question (Blakemore & Tobin, 1972; Meese, 2004). The details, assumptions and parameters are given in To et al. (2010) and Tolhurst et al. (2010).

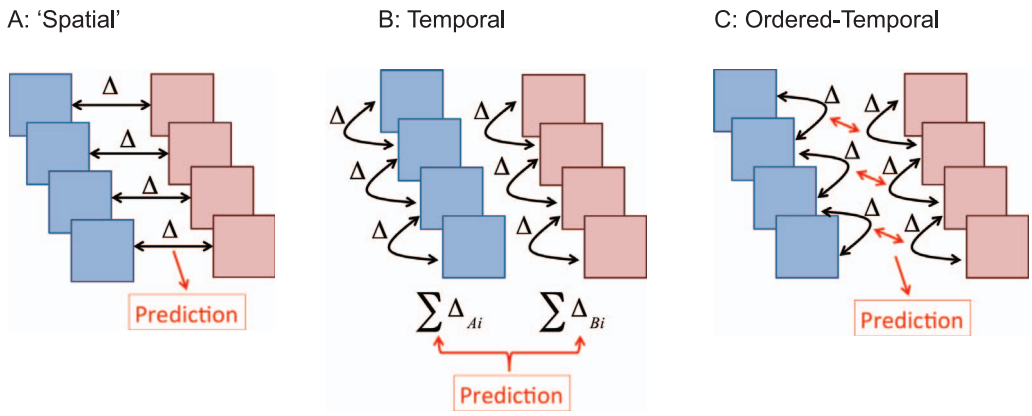

Figure 3. Schematics showing the three methods we used to compare movie clips, based on the idea that we should compare pairs of movie frames directly, either between the movies (A) or within movies (B, C). See details in the text.

Comparing all the frames in a pair of movie clips

Such VDP modeling is confined to extracting a single predicted difference value from comparing two images. Here, we elaborate the VDP to comparing two movie clips, each of which comprises 150 possibly different frames or brief static images. We have investigated three different methods of breaking movies into static image pairs for application of the VDP model. These methods tap different spatio-temporal aspects of any differences between the paired movies, and are shown schematically in Figure 3.

The Spatial Model: Stepwise between-movie comparison (Figure 3A). Each of the 150 frames in one movie is compared with the corresponding frame of the other movie, giving 150 VDP outputs which are then combined by Minkowski summation into a single number (Watson, 1998; Watson et al., 2001). This is one prediction of how different the two movie clips would seem to be. The VDP could show differences for two reasons. First, if the movies are essentially of the same scene but are moving temporally at a different rate, the compared frames will get out of synchrony. Second, at an extreme, the VDP would show differences even if the two movie clips were actually static but had different *spatial* content. This method of comparing frames will be sensitive to differences in spatial structure as well as to differences in dynamics, and is a measure of *overall difference in spatio-temporal structure*.

The Temporal Model: Stepwise within-movie comparison. Here the two movies are first analyzed separately. Each frame in a movie clip is compared with the successive frame in the same movie; with 150 frames, this gives 149 VDP comparison values which are combined to a single value by Minkowski summation (Figure 3B). This is a measure of the overall dynamic content of the movie clip, and is calculated for each of the paired movies separately. The difference between the values for the two clips is a measure of *the difference in overall dynamic content*. It would be uninfluenced by any spatial differences in scene content.

The Ordered-Temporal Model: Hybrid two-step movie comparison. The previous measure gives overall dynamic structure but may not, for instance illustrate whether particular dynamic events occur at different times within the two movie clips (Figure 3C). Here, each frame in a movie clip is compared directly with the succeeding frame (as before) but this dynamic frame difference is now compared with the dynamic difference between the two equivalent frames in the other movie clip. So within each movie, the nth frame is compared to the $(n+1)$th frame, and then these two differences (the nth difference from each movie) are compared to generate a between-movie difference. The 149 between-movie differences are combined again by Minkowski summation. This gives a measure of the perceived difference in *ordered dynamic content*, irrespective of any spatial scene content differences. This model has the advantage of being sensitive to dynamic changes occurring both within each movie and between the two.

Modeling realistic impulse functions

Direct VDP comparisons of paired frames is not a realistic model since the human visual system is almost certainly unable to code each 8 ms frame as a separate event. As a result, it is necessary to model human dynamic sensitivity. Thus, before performing the paired VDP comparisons, we convolved the movie clips with two different impulse functions (Figure 4A) representing "sustained" (P-cell, red curve) or "transient" (M-cell, blue curve) channels. The impulse functions have shape and duration that matches the temporal-frequency sensitivity functions (Figure 4B) inferred for the two kinds of channel by Kulikowski and Tolhurst (1973). The three methods for comparing movie frames

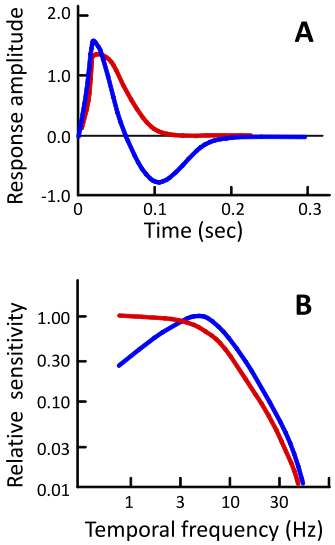

Figure 4. (A) The impulse functions used to convolve the movie clips: Sustained (red) and Transient (blue). These impulse functions have the temporal-frequency filter curves shown in (B) and are like those proposed by Kulikowski and Tolhurst (1973) for sustained and transient channels.

were performed separately on the movies convolved with the sustained and transient impulse functions, giving six potential predictors of human performance. Note that, in the absence of realistic temporal filtering, the analysis would effectively be with a very brief *sustained* impulse function lasting no more than 8 ms (compared to the 100 ms duration of the realistic sustained impulse; Figure 4A).

Experimental results

Reliability of the observers' ratings

Seven naïve observers were presented with 198 pairs of monochrome movies and were asked to rate how different they perceived the two movie clips in each pair to be. We have previously shown in To et al. (2010) that observers' ratings for static natural images are reliable. To verify the robustness of the difference ratings for movies pairs in the present experiment, we considered the correlation coefficient between the ratings given by any one observer with those given by each other observer; this value ranged from 0.36 to 0.62 (mean 0.50).

In our previous visual discrimination studies (To et al., 2010), we have scaled and averaged the difference ratings of all seven observers to each stimulus pair (see Methods), and have used the averages to compare with various model predictions. Although this procedure might remove any potential between-observer differences in strategy, it was used to average out within-observer variability. As in our previous experiments and analyses, averaging together the results of observers produced a reliable datasets for modeling.

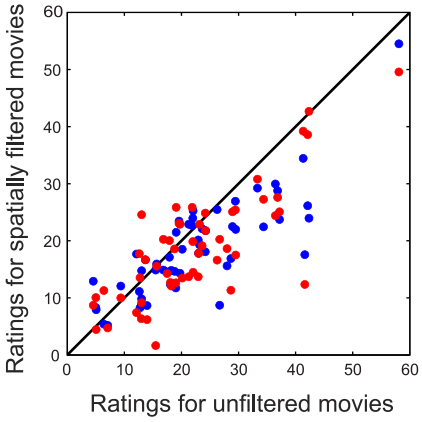

Figure 5. The averaged ratings given by observers for the spatially filtered movies are plotted against the ratings for the raw, unfiltered versions of the movies. Low-pass filtered (blue) and high-pass filtered (red).

When observers' averaged ratings for unfiltered, high-pass and low-pass movies were compared (see Figure 5), we found that the ratings for filtered movies were generally lower than those for unfiltered movies. This would seem understandable, as unfiltered movies contain more details.

Observers' ratings versus model predictions

The average of the normalized ratings for each movie pair was compared with the predictions obtained from a variety of different models of V1 complex cell processing. The correlation coefficients between the ratings and predictions are presented in Table 1. The different versions of the model all rely ultimately on a Visual Difference Predictor (To et al., 2010; Tolhurst et al., 2010), which calculates the perceived difference between a pair of frames. Given that the movie clips are composed of 150 frames each, there are a number of ways of choosing paired frames to compare a pair of movie clips (see Methods). The Spatial Model takes each frame in one movie and compares it with the equivalent frame in the second movie, and is sensitive to differences in spatial content and dynamic events in the movie clips. The Temporal Model compares successive frames within a movie clip to assess the overall dynamic content of a clip. The Ordered-Temporal Model is similar, but it compares the timings of dynamic events between movie clips.

Models	Transient and/or sustained	All movies ($n = 162$)	Unfiltered ($n = 54$)	Low-pass ($n = 54$)	High-pass ($n = 54$)
Spatial	–	0.288	0.338	0.274	0.336
Temporal	–	0.357	0.409	0.102	0.546
Ordered-Temporal	–	0.438	0.505	0.266	0.584
Multilinear (three parameters)	–	0.507	0.608	0.392	0.704
Spatial	S	0.396	0.425	0.306	0.582
Temporal	S	0.299	0.292	0.241	0.414
Ordered-Temporal	S	0.415	0.503	0.370	0.488
Multilinear (three parameters)	S	0.509	0.631	0.408	0.749
Spatial	T	0.453	0.391	0.444	0.565
Temporal	T	0.454	0.518	0.481	0.425
Ordered-Temporal	T	0.342	0.311	0.301	0.469
Multilinear (three parameters)	T	0.598	0.627	0.636	0.675
Multilinear (six parameters)	T/S	0.759	0.788	0.702	0.806

Table.1. The correlation coefficients between the observers' ratings (average of 7 per stimulus) and various V1-based models. The correlations for the spatially unfiltered, for the low-pass, and the high-pass filtered subsets are shown separately, as well as the fits for the entire set of 162 movie pairs. Correlations are shown for raw movies and for temporally-filtered (sustained and transient) models. The final six-parameter multilinear fit is a prediction of ratings from the three models and the two kinds of impulse function.

Without temporal filtering

We first used the VDP to compare pairs of frames as they were actually presented on the display; this presumes unrealistically that an observer could treat each 8 ms frame as a separate event and that the human contrast sensitivity function (CSF) is flat out to about 60 Hz. Overall, the best predictions by just one model were generated by the Ordered-Temporal Model ($r = 0.438$). The difference predictions were slightly improved when the three models were combined into a multiple linear regression model and the correlation coefficient increased to $r = 0.507$. When considering the accuracy of the models' predictions in relation to the different types of movies presented (i.e., spatially unfiltered, low-pass, and high-pass), the predictions were consistently superior for the high-pass movies, and poorest predictions were for difference ratings for the low-pass movies. More specifically, in the case of the *Multilinear Model*, the correlation values between predictions and measured ratings for unfiltered, low-pass and high-pass were 0.608, 0.392, and 0.704, respectively.

Realistic temporal filtering

The predictions discussed so far were generated by models that compare the movies pairs frame-by-frame, without taking into account the temporal properties of V1 neurons. Now we examine whether the inclusion of sustained or/and transient temporal filters (Figure 4) improves the models' performance at predicting perceived differences between two movie clips.

The Multilinear Model based only on the sustained impulse function caused little improvement overall ($r = 0.509$ now vs. 0.507 originally) or separately to the three different kinds of spatially filtered movie clips, and the predictions given by the different models separately for the different spatially filtered movie versions were also little improved. In particular, the fit for low-pass movies remained poor ($r = 0.408$ vs. $r = 0.392$). This is hardly surprising given that low spatial frequencies are believed to be detected primarily by transient channels rather than sustained channels (Kulikowski & Tolhurst, 1973). However, it is worth noting that, for the high-pass spatially filtered movies, the realistic sustained impulse improves the predictions of the Spatial Model (r increases from 0.336 to 0.582) but lessens the predictions of the Ordered-Temporal Model (r decreases from 0.584 to 0.488). This would seem to be consistent with the idea that higher spatial frequencies are predominantly processed by sustained channels (Kulikowski & Tolhurst, 1973).

As expected, when the modeling was based on the transient impulse function alone with its substantially different form (Figure 4), the nature of the correlations changed. In particular, it generated much stronger fits for low-pass spatially filtered ($r = 0.636$ vs. $r = 0.408$ in previous case). However, the fits for high-pass were noticeably weaker ($r = 0.675$ vs. $r = 0.749$). When taking all the movies into account, the improvement of the predictions for the low-pass movies meant that the overall correlation for the multilinear fit improved from 0.509 to 0.598. It therefore becomes clear that the sustained and transient filters operate quite specifically to improve the predictions for only one type of movie: high-pass movies in the case of sustained impulse functions and low-pass movies in the case of transient impulse functions.

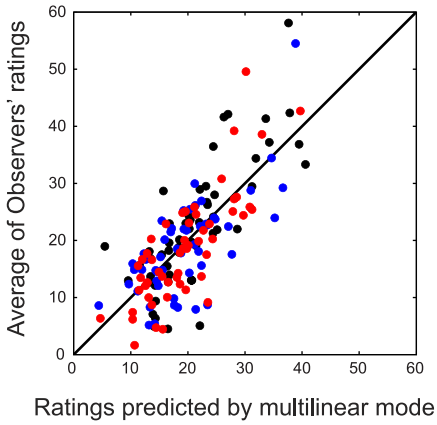

Figure 6. The averaged ratings given by observers for the 162 movie pairs consisting of exactly 150 frames are plotted against our most comprehensive spatiotemporal VDP model: raw movies (black), low-pass filtered (blue) and high-pass filtered (red). The model is a multilinear regression with six model terms and an intercept (a seventh parameter). The six-model terms are for the three methods of comparing frames, each with sustained and transient impulse functions. The correlation coefficient with seven parameters is 0.759. See the Supplementary Materials for more details of this regression.

Finally, the predictions of the V1 based model were substantially improved by incorporating *both* the transient and sustained filters; that is, a multiple linear fit with six parameters (three ways for comparing frames, for each of the sustained and transient temporal filters). Figure 6 plots the observers' averaged ratings against the six-parameter model predictions. The correlation coefficients between the predictions generated by this six-parameter multilinear model and observers' measured ratings was 0.759 for the full set of 162 movie pairs, a very substantial improvement over our starting value of 0.507 ($P = 0.0001$, $n = 162$). The correlations for the six parameter fits also improved separately for the spatially unfiltered (black symbols), the low-pass movies (blue), and the high-pass movies (red). The greatest improvement from our starting point without temporal filtering was in the multilinear fit for the low-pass spatially filtered movies (r increases from 0.392 to 0.702).

Discussion

We have investigated whether it is possible to model how human observers rate the differences between two dynamic naturalistic scenes. Observers were presented with pairs of movie clips that differed on various dimensions, such as glitches and speed, and were asked to rate how different the clips appeared to them. We then compared the predictions from a variety of V1-based computational models with the measured ratings. All the models were based on multiple comparisons of pairs of static frames, but the models differed in how pairs of frames were chosen for comparison. The pairwise VDP comparison is the same as we have reported for static photographs of natural scenes (To et al., 2010), but augmented by prior convolution in time of the movie clips with separate realistic sustained and transient impulse functions (Kulikowski & Tolhurst, 1973).

The use of magnitude estimation ratings in visual psychophysical experiments might be considered too subjective and, hence, not very reliable. We have previously shown that ratings are reliable (To et al., 2010): repeated ratings within subjects are highly correlated/reproducible (even when the stimulus set was 900 image pairs), and the correlation for ratings between observers is generally good. We have chosen this method in this and previous studies for three reasons. First, when observers generate ratings of how different pairs of stimuli appear to them, they have the freedom to rate the pairs as they please, regardless of which features were changed. The ratings therefore allow us to measure discrimination across and independently of feature dimensions, separately and in combination (To et al., 2008; To et al., 2011). Second, unlike threshold measurements, ratings can be collected quickly. We are interested in studying vision with naturalistic images and naturalistic tasks, and it is a common criticism of such study that natural images are so varied that we must always study very many (198 movie pairs in this study). Using staircase threshold techniques for so many stimuli would be impractically slow. Third, ratings are not confined to threshold stimuli, and we consider it necessary to be able to measure visual performance with suprathreshold stimuli as well. Overall, averaging the ratings of several observers together has provided us with data sets that are certainly good enough to allow us to distinguish between different V1-based models of visual coding (To et al., 2010; To et al., 2011). Although our study is not aimed toward the invention of image quality metrics, we have essentially borrowed our ratings technique from that applied field.

We have investigated a number of different measures for comparing the amount and timing of dynamic events in within a pair of movies. Our Spatial Model computed the differences between corresponding frames in each clip in the movie pair and then combined the values using Minkowski summation. This is the method used by Watson (1998) and Watson et al. (2001) in their model for evaluating the degree of perceived distortion in compressed video streams. It seems particularly appropriate in such a context where the compressed video would have the same fundamental spatial and synchronous temporal structure as

the uncompressed reference. In a sense, this measure imagines that an observer views and memorizes the first movie clip, and can then replay the memory while the second clip is played. In our experiments, it will be sensitive to any differences in the spatial content of the movies but also to temporal differences caused by loss of synchrony between otherwise identical movies. The greater the temporal difference or the longer the difference lasts, the more the single frames will become decorrelated and the greater the measured difference. Others have suggested that the video distortion metrics would be enhanced by using additional measures, where observers evaluate the dynamic content within each movie clip separately (e.g., Seshadrinathan & Bovik, 2010; Vu & Chandler, 2014; Wang & Li, 2007). We investigated a Temporal Model, which combined the differences between successive frames within each movie, and then compared the summary values between the movies. While this measure might summarize the degree of dynamic activity within each movie clip, it will be insensitive to differences in spatial content, and it is likely to be insensitive to whether two movie clips have the same overall dynamic content but a different order of events. Thus, our Ordered-Temporal Model is a hybrid model that included elements from both the Spatial and Temporal Models; this should be particularly enhanced with a transient impulse function, which itself highlights dynamic change.

It is easy to imagine particular kinds of difference between movie clips that one or other of our metrics will fail to detect, but we feel that most differences will be shown up by combined use of several metrics. Some may still escape since we have not explicitly modeled the coding of lateral motion. Might our modeling miss the rather obvious difference between a leftward and rightward moving sine-wave grating, or a left/right flip in the spatial content of the scene? Natural movement in natural scenes is less likely to be so regularly stereotyped as a single moving grating. More interestingly, a literal head-to-head comparison of movie frames will detect almost any spatio-temporal difference, but it may be that some kinds of difference are not perceived by the observer, even though a visual cortex model suggests that they should be visible, perhaps because such differences are not of any survival benefit (see To et al., 2010). Our different spatial or temporal metrics aim to measure the differences between movie clips; they are not aimed to specifically map onto the different kinds of decision that the observers actually make. Observers often reported that they noticed that the "rhythm" of one clip was distorted or that the steady or repetitive movement within one clip suddenly speeded up. This implies a decision based primarily upon interpreting just one of the clips rather than on any direct comparison. Our model can only sense the rhythm of one clip or a sudden change within it by comparing it with the control clip that has a steady rhythm and steady continuous movement.

As well as naturalistic movie clips, we also presented clips that were subject either to low-pass spatial filtering or to high-pass filtering. In general, the sustained filter was particularly useful at strengthening predictions for the movies containing only high spatial frequencies. On the other hand, the transient filter improved predictions for the low-pass spatially filtered movies. These results seem to be consistent with the "classical" view of sustained and transient channels (a.k.a. the parvocellular vs. magnocellular processing pathways) with their different spatial frequency ranges (Hess & Snowden, 1992; Horiguchi et al., 2009; Keesey, 1972; Kulikowski & Tolhurst, 1973; Tolhurst, 1973, 1975). In order to get good predictions of observers' ratings, it was necessary to have a multilinear regression that included *both* sustained and transient contributions.

Although a multiple linear regression fit (Figure 6) with all six candidates' measures (three ways for comparing frames, for each of the sustained and transient temporal filters) yielded strong predictions of observers' ratings ($r = 0.759$, $n = 162$), it is necessary to point out that there are significant correlations between the six measures. Using a stepwise linear regression as an exploratory tool to determine the relative weight of each, we found that, of the six factors, only two stand out as being particularly important (see Supplementary Materials): the spatial/sustained factor, which is sensitive to spatial information as well as temporal, and the ordered-temporal/transient factor handling the specific sequences of dynamic events. A multilinear regression with only these two measures gave predictions that were almost as strong as with all six factors ($r = 0.733$, Supplementary Materials). Again, this finding seems to be consistent with the classical view of sustained and transient channels, that they convey different perceptual information about spatial structure and dynamic content.

Our VDP is intended to be a model of low-level visual processing, as realistic a model of V1 processing as we can deduce from the vast neurophysiological single-neuron literature. The M-cell and P-cell streams (Derrington & Lennie, 1984; Gouras, 1968)—the Transient and Sustained candidates—enter V1 separately and, at first sight, travel through V1 separately to reach different extrastriate visual areas (Merigan & Maunsell, 1993; Nassi & Callaway, 2009). However, it is undoubtedly true that there is some convergence of M-cell and P-cell input onto the same V1 neurons (Sawatari & Callaway, 1996; Vidyasagar, Kulikowski, Lipnicki, & Dreher, 2002) and that extrastriate areas receive convergent input, perhaps via different routes

(Nassi & Callaway, 2006; Ninomiya, Sawamura, Inoue, & Takada, 2011). Even if the transient and sustained pathways do not remain entirely separate up to the highest levels of perception, it still remains that perception of high spatial frequency information will be biased towards sustained processing, while low spatial frequency information will be biased towards transient processing with its implication in motion sensing.

Our measures of perceived difference rely primarily on head-to-head comparisons of corresponding frames between movies or of the dynamic events at corresponding points. This is core to models of perceived distortion in compressed videos (e.g., Watson, 1998). It is suitable for the latter case because the original and compressed videos are likely to be exactly the same length. However, this becomes a limitation in a more general comparison of movie clips, when the clips might be of unequal length. The bulk of the stimuli in our experiment were, indeed, paired movie clips of equal length, but 36 out of the total 198 did differ in length. We have not included these 36 in the analyses in the Results section because we could not simply perform the head-to-head comparisons between the paired clips. In the Supplementary Materials, we discuss this further: by truncating the longer clip of the pair to the same length as the shorter one, we were able to calculate the predictive measures. The resulting six-parameter multilinear regression for all stimuli had a less good fit than for the subset of equal length videos, which suggests that a better method of comparing unequal length movie clips could be sought.

The inclusion of sustained and transient filters allowed us to model low-level temporal processing. However one surprise is that we achieved very good fits without explicitly modeling neuronal responses to lateral motion. It may be relevant that, in a survey of saliency models, Borji, Sihite and Itti (2013) report that models incorporating motion did not perform better than the best static models over video datasets. One possibility is that lateral motion produces some signature in the combination of the various measures that we have calculated. Future work could include lateral movement sensing elements in our model, either separately or in addition to the temporal impulse functions and a number of such models of cortical motion sensing both in V1 and in MT/V5 are available (Adelson & Bergen, 1985; Johnston, McOwan, & Buxton, 1992; Perrone, 2004; Simoncelli & Heeger, 1998). However, the current model does surprisingly well without such a mechanism.

Keywords: computational modeling, visual discrimination, natural scenes, spatial processing, temporal processing

Acknowledgments

We are pleased to acknowledge that the late Tom Troscianko played a key role in the planning of this project, and that P. George Lovell and Chris Benton contributed to the collection of the raw movie clips. The research was funded by grants (EP/E037097/1 and EP/E037372/1) from the EPSRC/Dstl under the Joint Grants Scheme. The first author was employed on E037097/1. The publication of this paper was funded by the Department of Psychology, Lancaster University, UK.

Commercial relationships: None.
Corresponding author: Michelle P. S. To.
Email: m.to@lancaster.ac.uk.
Address: Department of Psychology, Lancaster University, Lancaster, UK.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 2*(2), 284–299.

Blakemore, C., & Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research, 15*(4), 439–440.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*(1), 55–69. doi:10.1109/TIP.2012.2210727.

Derrington, A. M., & Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *Journal of Physiology, 357,* 219–240.

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 11*(6), 1710–1719.

Gouras, P. (1968). Identification of cone mechanisms in monkey ganglion cells. *Journal of Physiology, 199*(3), 533–547.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640. doi:10.1126/science.1089506.

Heeger, D. J. (1992). Normalization of cell responses in

cat striate cortex. *Visual Neuroscience, 9*(2), 181–197.

Hess, R. F., & Snowden, R. J. (1992). Temporal properties of human visual filters: Number, shapes and spatial covariation. *Vision Research, 32*(1), 47–59.

Hirose, Y., Kennedy, A., & Tatler, B. W. (2010). Perception and memory across viewpoint changes in moving images. *Journal of Vision, 10*(4), *2*, 1–19, http://www.journalofvision.org/content/10/4/2, doi:10.1167/10.4.2. [PubMed] [Article]

Horiguchi, H., Nakadomari, S., Misaki, M., & Wandell, B. A. (2009). Two temporal channels in human V1 identified using fMRI. *Neuroimage, 47*(1), 273–280. doi:10.1016/j.neuroimage.2009.03.078.

Itti, L., Dhayale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. *Vision Research, 44*(15), 1733–1755.

Johnston, A., McOwan, P. W., & Buxton, H. (1992). A computational model of the analysis of some first-order and second-order motion patterns by simple and complex cells. *Proceedings of the Royal Society B: Biological Sciences, 250*(1329), 297–306. doi:10.1098/rspb.1992.0162.

Keesey, U. T. (1972). Flicker and pattern detection: A comparison of thresholds. *Journal of the Optical Society of America, 62*(3), 446–448.

Kulikowski, J. J., & Tolhurst, D. J. (1973). Psychophysical evidence for sustained and transient detectors in human vision. *Journal of Physiology, 232*(1), 149–162.

Lovell, P. G., Gilchrist, I. D., Tolhurst, D. J., & Troscianko, T. (2009). Search for gross illumination discrepancies in images of natural objects. *Journal of Vision, 9*(1):37, 1–14, http://www.journalofvision.org/content/9/1/37, doi:10.1167/9.1.37. [PubMed] [Article]

Meese, T. S. (2004). Area summation and masking. *Journal of Vision, 4*(10):8, 930–943, http://www.journalofvision.org/content/4/10/8, doi:10.1167/4.10.8. [PubMed] [Article]

Merigan, W. H., & Maunsell, J. H. (1993). How parallel are the primate visual pathways? *Annual Review of Neuroscience, 16,* 369–402. doi:10.1146/annurev.ne.16.030193.002101.

Nassi, J. J., & Callaway, E. M. (2006). Multiple circuits relaying primate parallel visual pathways to the middle temporal area. *Journal of Neuroscience, 26*(49), 12789–12798. doi:10.1523/JNEUROSCI.4044-06.2006.

Nassi, J. J., & Callaway, E. M. (2009). Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience, 10*(5), 360–372. doi:10.1038/nrn2619.

Ninomiya, T., Sawamura, H., Inoue, K., & Takada, M. (2011). Differential architecture of multisynaptic geniculo-cortical pathways to V4 and MT. *Cerebral Cortex, 21*(12), 2797–2808. doi:10.1093/cercor/bhr078.

Peli, E. (1990). Contrast in complex images. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 7*(10), 2032–2040.

Perrone, J. A. (2004). A visual motion sensor based on the properties of V1 and MT neurons. *Vision Research, 44*(15), 1733–1755. doi:10.1016/j.visres.2004.03.003.

Sawatari, A., & Callaway, E. M. (1996). Convergence of magno- and parvocellular pathways in layer 4B of macaque primary visual cortex. *Nature, 380*(6573), 442–446. doi:10.1038/380442a0.

Seshadrinathan, K., & Bovik, A. (2010). Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing, 19*(2), 335–350.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38*(5), 743–761.

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision, 13*(8):16, 1–24, http://www.journalofvision.org/content/13/8/16, doi:10.1167/13.8.16. [PubMed] [Article]

Tadmor, Y., & Tolhurst, D. J. (1994). Discrimination of changes in the second-order statistics of natural and synthetic images. *Vision Research, 34*(4), 541–554.

Tadmor, Y., & Tolhurst, D. J. (2000). Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Research, 40*(22), 3145–3157.

To, M., Lovell, P. G., Troscianko, T., & Tolhurst, D. J. (2008). Summation of perceptual cues in natural visual scenes. *Proceedings of the Royal Society B: Biological Sciences, 275*(1649), 2299–2308. doi:10.1098/rspb.2008.0692.

To, M. P. S., Baddeley, R. J., Troscianko, T., & Tolhurst, D. J. (2011). A general rule for sensory cue summation: Evidence from photographic, musical, phonetic and cross-modal stimuli. *Proceedings of the Royal Society B: Biological Sciences, 278*(1710), 1365–1372. doi:10.1098/rspb.2010.1888.

To, M. P. S., Gilchrist, I. D., & Tolhurst, D. J. (2014). Transient vs. Sustained: Modeling temporal impulse functions in visual discrimination of dynamic natural images. *Perception, 43,* 478.

To, M. P. S., Lovell, P. G., Troscianko, T., Gilchrist, I. D., Benton, C. P., & Tolhurst, D. J. (2012). Modeling human discrimination of moving images. *Perception, 41*(10), 1270–1271.

To, M. P. S., Lovell, P. G., Troscianko, T., & Tolhurst, D. J. (2010). Perception of suprathreshold naturalistic changes in colored natural images. *Journal of Vision, 10*(4):12, 1–22, http://www.journalofvision.org/content/10/4/12, doi:10.1167/10.4.12. [PubMed] [Article]

Tolhurst, D. J. (1973). Separate channels for the analysis of the shape and the movement of moving visual stimulus. *Journal of Physiology, 231*(3), 385–402.

Tolhurst, D. J. (1975). Sustained and transient channels in human vision. *Vision Research, 15,* 1151–1155.

Tolhurst, D. J., & Thompson, I. D. (1981). On the variety of spatial frequency selectivities shown by neurons in area 17 of the cat. *Proceedings of the Royal Society B: Biological Sciences, 213*(1191), 183–199.

Tolhurst, D. J., To, M. P. S., Chirimuuta, M., Troscianko, T., Chua, P. Y., & Lovell, P. G. (2010). Magnitude of perceived change in natural images may be linearly proportional to differences in neuronal firing rates. *Seeing Perceiving, 23*(4), 349–372.

Troscianko, T., Meese, T. S., & Hinde, S. (2012). Perception while watching movies: Effects of physical screen size and scene type. *Iperception, 3*(7), 414–425. doi:10.1068/i0475aap.

Vidyasagar, T. R., Kulikowski, J. J., Lipnicki, D. M., & Dreher, B. (2002). Convergence of parvocellular and magnocellular information channels in the primary visual cortex of the macaque. *European Journal of Neuroscience, 16*(5), 945–956.

Vu, P. V., & Chandler, D. M. (2014). Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. Online supplement. *The Journal of Electronic Imaging, 23*(1), 013016.

Wang, Z., & Li, Q. (2007). Video quality assessment using a statistical model of human visual speed perception. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 24*(12), B61–B69.

Watson, A. B. (1987). Efficiency of a model human image code. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 4*(12), 2401–2417.

Watson, A. B. (1998). Toward a perceptual video quality metric. *Human Vision, Visual Processing, and Digital Display VIII, 3299,* 139–147.

Watson, A. B., Hu, J., & McGowan, J. F., III. (2001). Digital video quality metric based on human vision. *Journal of Electronic Imaging, 10,* 20–29.

Watson, A. B., & Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 14*(9), 2379–2391.