Atten Percept Psychophys (2011) 73:572–580
DOI 10.3758/s13414-010-0028-z

Diagnostic spatial frequencies and human efficiency for discriminating actions

Steven M. Thurman · Emily D. Grossman

Published online: 16 November 2010

Abstract Humans extract visual information from the world through spatial frequency (SF) channels that are sensitive to different scales of light-dark fluctuations across visual space. Using two methods, we measured human SF tuning for discriminating videos of human actions (walking, running, skipping and jumping). The first, more traditional, approach measured signal-to-noise ratio (*s/n*) thresholds for videos filtered by one of six Gaussian band-pass filters ranging from 4 to 128 cycles/image. The second approach used SF "bubbles", Willenbockel et al. (*Journal of Experimental Psychology. Human Perception and Performance*, *36*(1), 122–135, 2010), which randomly filters the entire SF domain on each trial and uses reverse correlation to estimate SF tuning. Results from both methods were consistent and revealed a diagnostic SF band centered between 12-16 cycles/image (about 1-1.25 cycles/body width). Efficiency on this task was estimated by comparing *s/n* thresholds for humans to an ideal observer, and was estimated to be quite low (>.04%) for both experiments.

Keywords Action recognition · Spatial frequency · Ideal observer · Biological motion · Bubbles

Introduction

The human visual system is organized to process visual information through spatial frequency (SF) channels that are each sensitive to a particular range of frequencies of repeating light-dark patterns across the visual field (see De Valois & De Valois, 1980). Since the discovery of the contrast sensitivity function for sinusoidal gratings (Campbell & Robson, 1968), one tradition in vision science has been to determine what SF information is critical for recognizing objects. SF tuning has been measured for various stimuli such as faces (Costen, Parker, & Craw, 1996; Fiorentini, Maffei, & Sandini, 1983; Gold, Bennett, & Sekuler, 1999), letters (Chung, Legge, & Tjan, 2002; Parish & Sperling, 1991), and objects (Norman & Ehrlich, 1987). These studies illustrate that diagnostic stimulus information is available in specific SF bands for different objects, and that observers readily extract this information for visual categorization and recognition (Sowden & Schyns, 2006).

Measuring SF tuning for objects gives vision researchers information about the scale of the diagnostic features for recognizing a given object or for discriminating different exemplars within an object class. For example, low spatial frequencies carry primarily information about global object shape and large-scale relations among features, while higher spatial frequencies carry information about object shape as well as fine-grained features of the object. Importantly, local and global stimulus features can be extracted to different degrees from a very large range of spatial frequencies because the coarse-to-fine processing mode is orthogonal to the local-to-global mode (e.g. Oliva & Schyns, 1997). Nonetheless, determining SF tuning for a visual stimulus is one method for estimating the relative importance of large and small-scale stimulus features for recognition. Since there is currently debate about the role of such features for perception of biological motion, the following question arises naturally: What are the diagnostic spatial frequency bands for recognizing dynamic human actions?

Although this question has not been addressed directly, a study by Kuhlmann and Lappe (2006) investigated perception

S. M. Thurman (✉) · E. D. Grossman
Department of Cognitive Sciences,
University of California, Irvine,
3151 Social Science Plaza,
Irvine, CA 92697-5100, USA
e-mail: sthurman@uci.edu

of human actions from natural scenes that were systematically blurred with a Gaussian filter, essentially removing high SF content (equivalent to a low-pass filter). The results from this study demonstrated that human actions could be reliably discriminated on the basis of very crude low SF information, even if only a few frames of the action sequence were shown. However if only a single frame was shown, performance dropped precipitously as blur level increased compared to the full action sequence. Results from Kuhlmann and Lappe (2006) highlight a few important points about action perception. First, action recognition is robust even when filtered to contain only low spatial frequencies, and there appears to be a critical point at which performance deteriorates when the image is too blurry. Second, observer performance improves when motion information is present in the action video by displaying three consecutive frames, though Kuhlmann and Lappe (2006) argue that motion aids primarily in segmenting the actor from the background, and not in the recognition process. However, it is still unclear exactly what role high SF information plays in action recognition and how this compares to intermediate and lower spatial frequencies.

The goal of the current study was to determine which spatial frequencies carry the most diagnostic information for discriminating actions. Researchers have used a variety of methods for measuring SF tuning of object recognition. For instance, one common method is to apply a series of increasing band-pass SF filters to a stimulus, ranging from low to high spatial frequencies, and then measuring signal to noise ratio (*s/n*) thresholds for each band-pass level (e.g. Gold et al., 1999; Parish & Sperling, 1991). Other methods for identifying critical SF bands for object recognition include low pass filtering (Rubin & Siegel, 1984), combining low and high-pass filtering (Fiorentini et al., 1983; Solomon & Pelli, 1994), and critical-band masking (Majaj, Pelli, Kurshan, & Palomares, 2002). Gold et al. (1999) present a useful table summarizing many previous studies including methods and results.

Willenbockel and colleagues (2010) recently introduced a novel method for measuring SF tuning inspired by the "bubbles" technique using reverse correlation (Gosselin & Schyns, 2001). The SF bubbles method involves filtering the SF domain of a stimulus on each trial with a random sampling vector, which can be envisioned as applying multiple, random band-pass filters of varying amplitude and bandwidth. Performing a multiple linear regression of observer accuracy with the sampling vectors across trials results in a classification vector revealing the SF bands that tend to lead to accurate discriminations. One benefit of the SF bubbles method is that it does not allow observers to adapt to a particular SF range during the experiment, and that it is basically a combination of all possible SF filtering experiments since it is equivalent to either low-pass, high-pass, or multiple band-pass filtering on each trial (Willenbockel et al., 2010).

In the current experiment we measured SF tuning for discriminating human actions using two different methods. We used a more traditional band-pass SF filtering method similar to Parish and Sperling (1991), and then we used the SF bubbles method described above (Willenbockel et al., 2010). The purpose of the current experiment was threefold. First, we wanted to determine if particular spatial frequency bands are more diagnostic than others for action discrimination. Second, we sought to compare the patterns of SF tuning derived from both techniques. Last, we performed ideal observer analysis in order to estimate how efficient observers are at extracting information at various spatial scales while discriminating human actions. The results of this experiment shed light on the spatial scale of diagnostic features for biological motion perception and allow us to quantitatively compare human efficiency for discriminating actions to previous estimates of efficiency for discriminating other types of objects.

Experiment 1

Participants

Seven participants were recruited at the University of California, Irvine, and were offered course credit for participation in the experiment. Author S.T. was one of the participants in this experiment. All participants had normal or corrected to normal vision.

Stimuli and apparatus

Stimuli included videos of nine human actors performing four different actions: walking, running, skipping and jumping. Videos were selected from an online database of freely available human actions (see Gorelick, Shechtman, Irani, & Basri, 2007). The actions chosen for this experiment represented four different types of ambulation; hence the actions differed primarily in terms of limb articulation, speed and body posture. We chose not to include non-ambulating actions, such as jumping jacks or hand waving, as it would be trivial to discriminate between ambulating figures and non-translating figures. The videos were recorded at a resolution of 180 by 144 pixels at 50 frames per second, and each of the videos had the same wall as a common background. In order to avoid edge artifacts that result from aliasing when Fourier analysis is performed with image dimensions that are not a power of 2, the videos were cropped and resized using bi-cubic interpolation to be 256 by 256

pixels. Each video was converted to grayscale and edited to consist of 25 frames. All image processing was performed using MATLAB.

The action stimuli were filtered with one of six Gaussian band-pass filters, each separated by one octave with a standard deviation of 0.5 octaves. The transfer functions of the filters are displayed in Fig. 1a. The centers of the filters were 4, 8, 16, 32, 64, and 128 (high-pass) cycles/image, which corresponded to center frequencies of 0.28, 0.57, 1.13, 2.27, 4.54, 9.08 cycles/degree visual angle. We created Gaussian noise fields by drawing independent samples from a Gaussian distribution (mean = 0, SD = 1) for each pixel in a 256 x 256 array. The noise fields were then filtered with one of the six band-pass filters, creating six sets of filtered Gaussian noise. Each set contained 100 unique filtered noise fields, and dynamic noise was created by randomly choosing 25 frames from the set of 100 on each trial.

The stimuli were presented on a 16 x 12 inch ViewSonic CRT monitor with a resolution of 1024 × 768 pixels and refresh rate of 100 Hz. Participants were seated in a dark room 38 cm from the screen with a chinrest to help maintain a constant viewing distance. At this distance the stimuli subtended 14.08 × 14.08 deg and were presented at a rate of 50 Hz. The experiment was programmed in Matlab using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997), powered by a 2 Ghz Intel core Apple Mac Mini.

Procedure

Observers first performed a training block to ensure that the actions could be reliably discriminated and to familiarize the observers with SF filtered stimuli. The training block consisted of 140 trials in which seven types of action videos (6 filtered and 1 unfiltered original) were displayed 20 times each, in random order. The actor and action in the video was chosen randomly on each trial. Since this was training, no noise was added to the stimuli. After checking to ensure that observers had less than a 5% error rate on the unfiltered original action videos, the experiment commenced.

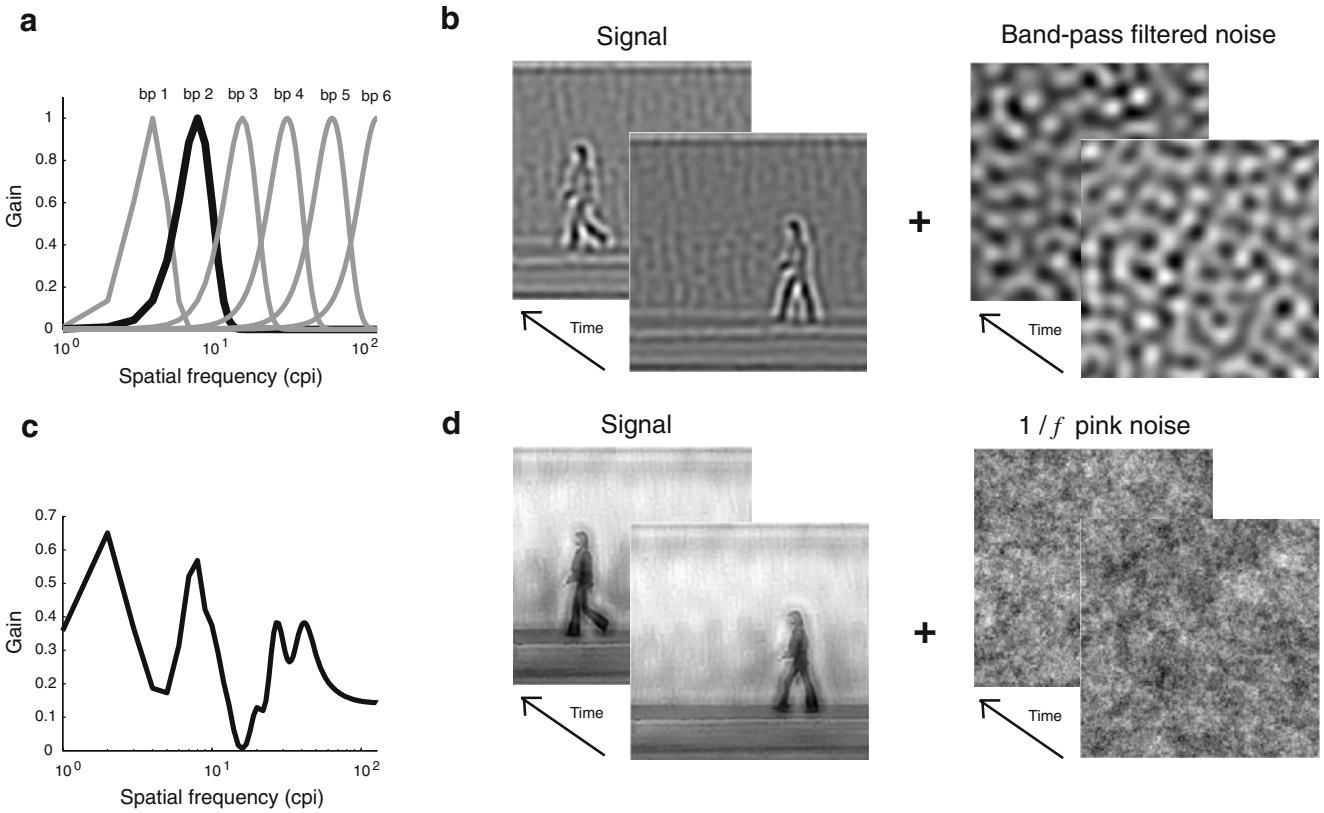

Fig. 1 SF filters, example stimuli and example noise fields from Experiments 1 and 2. **a**. Transfer functions for each of the Gaussian band-pass filters used in Experiment 1, labeled as bp1 through bp 6 from low to high spatial frequencies. **b**. Selected frames from an example action video (left), filtered with the darkened filter in plot A (bp 2), plus an example of dynamic Gaussian noise filtered in the same SF band (right). **c**. Transfer function of an example SF sampling vector using the SF bubbles method. **d**. Selected frames from a stimulus filtered with the transfer function in plot C (left), plus an example of dynamic Gaussian pink noise (right)

Signal to noise ratio thresholds were estimated using the method of constant stimuli with a four-choice discrimination task. Observers performed two blocks of 360 trials. In each block, each of the six band-pass filtered action stimuli was displayed a total of 60 times. Those 60 trials of action stimuli consisted of five different *s/n* ratio levels shown 12 times each. The five *s/n* levels were 0.005, 0.01, 0.05, 0.1, 0.5, and were chosen based on pilot data to sample the psychometric function from about chance performance to above threshold performance across SF bands. Signal to noise ratio was measured by computing signal power, *s*, from the contrast variance of the signal (filtered action video) and noise power, *n*, from the contrast variance of the noise (filtered Gaussian fields), and dividing *s* by *n*. On each trial the contrast variance of the noise field was multiplied by a scaling factor in order to achieve the target *s/n* level. This method for computing and adjusting *s/n* is described in detail by Parish and Sperling (1991).

On each trial a stimulus consisted of a randomly chosen filtered action video plus an identically filtered noise field with a target *s/n* randomly chosen from the list of five *s/n* levels (see Fig. 1b). The stimulus was presented for 500 ms, followed by an answer screen displaying the mapping between keyboard responses and the four possible actions. The response screen remained until the observer responded by using the numbers 1-4 on a keyboard. No feedback was given. The next trial commenced after a pause of 3 seconds. The entire experiment lasted about one hour.

Data analysis

For each observer, accuracy was computed for each of the five *s/n* levels in the six SF bands. In total, accuracy was computed from 24 data samples per condition (30 total conditions). Data for each of the seven observers was combined and a best-fitting (least-squares) Weibull psychometric function was fitted to the mean data in order to estimate *s/n* thresholds at 80% performance for each SF band.

Ideal observer

As argued by Gold et al. (1999), optimal performance for this task can be computed by maximizing the cross-correlation between the filtered, noise-masked stimulus and each of the templates (unfiltered action videos). This is analogous to the spatial correlator ideal discriminator described by Parish and Sperling (1991). We used this method to estimate ideal observer thresholds for the current experiment. We performed Monte Carlo simulations testing the accuracy of the ideal observer discriminating action videos that were filtered in each of the same six SF bands and masked with identically filtered Gaussian noise at a variety of *s/n* levels. The response of the ideal observer on each simulated trial was chosen by computing the correlation between the test stimulus and each of the 36 action templates (4 actions by 9 actors), and then choosing the template with the maximum correlation. We tested performance of the ideal observer on 80 trials for each condition (6 SF bands by 5 *s/n* levels) and estimated *s/n* thresholds with a best-fitting Weibull function.

Results

Figure 2 shows human observer and ideal observer performance as a function of *s/n* level. The ideal observer was able to discriminate actions at all band-pass filter levels, so information was clearly available in all SF bands to perform the task. Human observers reached the 80% threshold in all SF bands except the first band. In fact observer performance discriminating actions in the first SF band during the practice block, which contained no noise, was only 77% on average (*SD* = 12%), so it was not surprising that observers failed to reach threshold with the *s/n* levels used.

The mean observer data and ideal observer data was fit with psychometric functions and 80% performance thresholds were estimated and plotted in Fig. 3a. Human observer tolerance to noise peaked in SF band 3, which

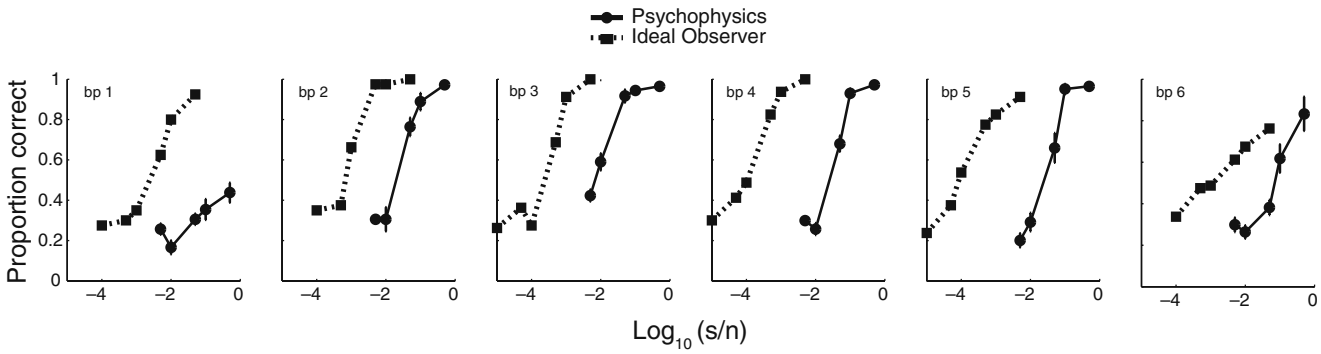

Fig. 2 Mean human (n = 7) and ideal observer performance as a function of *s/n* level. Each plot represents a single SF band, labeled bp1 through bp 6 from left to right. The x-axis represents the log transform of the *s/n* level. Error bars represent the standard error of the mean

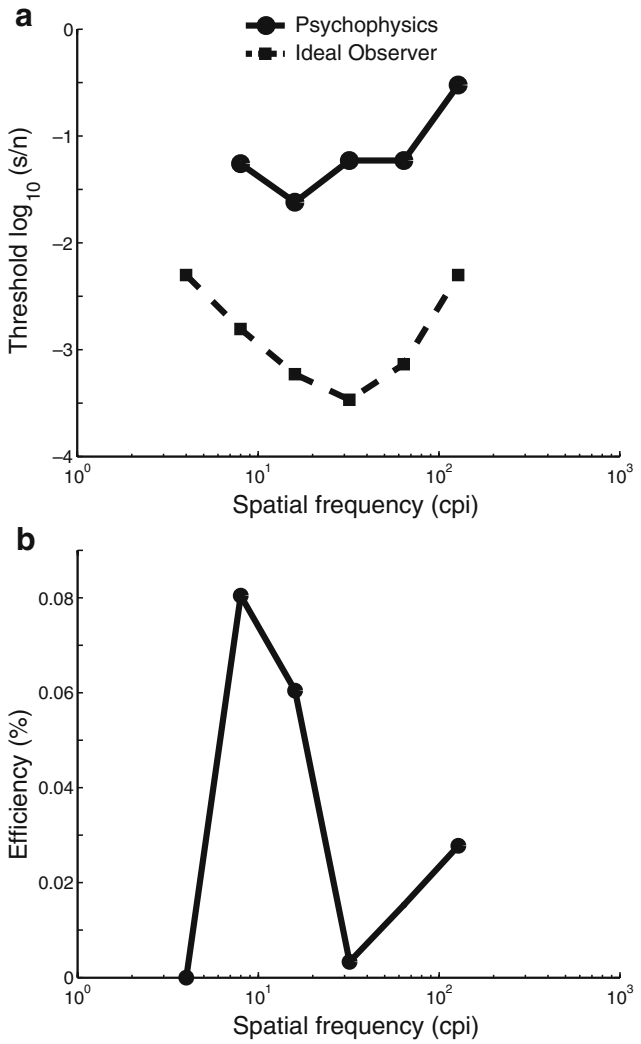

Fig. 3 **a**. Signal-to-noise ratio thresholds (80% accuracy) for human and ideal observers. Thresholds were estimated using a best-fitting (least-squares) Weibull function. Data points represent the center of each SF band. **b** Human efficiency for each SF band

corresponded to 16 cycles/image. Since the average body torso in the videos was about 20 pixels wide and the average body height was about 120 pixels tall, this was equivalent to 1.25 cycles/body width and 7.5 cycles/body height.

Figure 3B shows human efficiency for each SF band measured. Efficiency was estimated using the same formula as Parish and Sperling (1991) as the ratio of the *s/n* thresholds of human and ideal observers at the 80% threshold criterion. Efficiency in SF band 1 was set to zero since observers failed to reach threshold in this condition. Efficiency had a sharp peak at 8 cycles/image, but overall was low across all SF bands, especially in comparison to the efficiency measured for other stimuli such as letters (Parish & Sperling, 1991; but see Gold et al., 1999).

Experiment 2

Participants

Ten participants were recruited at the University of California, Irvine, and were given the option of obtaining course credit for participation in the experiment. All participants had normal or corrected to normal vision. None of the participants in Experiment 2 participated in the first experiment.

Stimuli and apparatus

The same set of action videos from Experiment 1 were used in Experiment 2. However unlike Experiment 1, which used a series of SF band-pass filters to measure SF tuning, Experiment 2 used the SF "bubbles" method. Willenbockel et al. (2010) present a helpful schematic of the general SF bubbles method, and what follows is an overview of the method and the specific parameters used in the current experiment.

On each trial the SF domain was sampled with a smooth random vector that was constructed in a 3-part process. First, a vector of zeros was created with a length of w x k, where w is the width of the image (256 pixels) and k is a smoothing constant. As k increases, the smoothing of the vector increases. We used an intermediate smoothing level, k = 20, which has been shown to produce reasonable SF tuning curves (Willenbockel et al., 2010). Then a constant number of "bubbles", represented by b, were randomly distributed in this vector by taking b samples from a uniform distribution of integers ranging from 1 to w x k, thus creating a binary vector with b ones and (wk – b) zeros. As the number b increases, the variance of the sampling vector tends to decrease due to smaller peaks and troughs. We used b = 45, the same as Willenbockel et al. (2010), but also tested smaller values of b (10 and 25) on one pilot subject (author SMT) and found that it did not substantially change the resulting SF tuning estimates. In the second step the random binary vector was convolved with a Gaussian kernel, or SF "bubble", with an arbitrary standard deviation of 1.5. This created a "smooth" sampling vector of length w x k. Lastly, a segment of length w/2 was randomly chosen from the smooth vector and transformed with a logarithmic function in order to match the SF sensitivity of the human visual system. The transformed vector served as the random SF filter for a given trial. The transfer function of an example SF bubbles vector is displayed in Fig. 1c.

Instead of using band-pass filtered Gaussian noise masks as in Experiment 1, we chose to use 1/f filtered Gaussian pink noise for Experiment 2 (see Fig. 1d). The primary reason was because all frequencies are randomly sampled

independently on each trial in a SF bubbles experiment, we did not want to bias observers to use a particular SF band by using a band-pass filtered mask. We also chose not to use unfiltered, broadband Gaussian white noise such as that used by Willenbockel et al (2010), because such a mask perceptually masks high frequencies better than low frequencies (Cass, Alais, Spehar, & Bex, 2009). Because $1/f$ pink noise perceptually masks all spatial frequencies about equally (Cass et al., 2009), it is less likely to bias observers to rely on particular SF bands. We created dynamic pink noise masks by filtering the same 100 Gaussian noise masks from Experiment 1 with a transfer function of $1/f$, where f is spatial frequency (cycles/image).

The apparatus and viewing parameters were the same as in Experiment 1.

Procedure

Observers performed a training block, as in Experiment 1. Error rates were checked to be below 5% before commencing the main experiment.

Observers performed four-choice discrimination of actions in three separate blocks of 200 trials. On each trial a random vector was created with the SF bubbles method, and used as the SF filter for a randomly chosen action video. Dynamic Gaussian pink noise was then added to the filtered stimulus (Fig. 1d). The contrast variance of the pink noise was adjusted on a trial-by-trial basis to maintain performance on average at 80% threshold. This was accomplished with a 3-1 double interleaved staircase procedure in which three consecutive correct responses led to a decrease in s/n ratio, and one incorrect response led to an increase in s/n. The first block of the experiment started with $s/n = 0.01$ and a step size of .002. Each subsequent block started with the threshold estimate of s/n from the previous block for that observer. Each stimulus was presented for 500 ms, followed by an answer screen displaying the mapping between keyboard responses and the four possible actions. The response screen remained until the observer responded by using the numbers 1-4 on a keyboard. No feedback was given. The next trial commenced after a pause of 3 seconds. The entire experiment lasted about one hour.

Data analysis

For each observer, the s/n threshold was estimated by averaging the s/n level across all trials in the last block of the experiment. The staircase procedure ensured that this s/n level resulted in about 80% accuracy in the action discrimination task.

SF tuning curves were computed by reverse correlating observer responses with the random SF sampling vectors across trials. Thus to determine which frequencies tended to lead to correct responses for each observer, we performed a multiple linear regression on the SF sampling vectors and observer accuracy across the 600 total trials. As described by Willenbockel et al (2010), this is computationally equivalent to taking a weighted sum of all 600 sampling vectors, with correct responses weighted as $1 - P(\text{correct})$ and incorrect responses weighted as $-P(\text{correct})$. Note that the staircase procedure kept the probability of a correct response, $P(\text{correct})$, at around 0.8 for each observer. This analysis resulted in a *classification vector* of regression coefficients that was transformed into Z-scores in order to perform statistical tests. We computed group classification vectors by summing classification vectors across all observers, and then dividing by the square root of n, the number of observers (Willenbockel et al., 2010). The group classification vector represents how diagnostic each SF was for discriminating the videos of human actions.

Ideal observer

A spatial correlator ideal observer similar to Experiment 1 was implemented for Experiment 2. We performed Monte Carlo simulations testing the accuracy of the ideal observer discriminating action videos that were filtered with random SF sampling vectors and masked in $1/f$ dynamic pink noise. The s/n was adjusted online using the same staircase procedure as used with human observers. Thus, the ideal observer was tested under the same conditions as the human observers. The response of the ideal observer on each simulated trial was chosen by computing the cross-correlation between the test stimulus and each of the thirty-six action video templates (4 actions by 9 actors), and then choosing the template with the maximum correlation. We tested the performance of the ideal observer in 10 separate blocks of Monte Carlo simulations, each with 600 trials, similar to the human observers. This was done so that we could compare the group variance in classification vectors between human and ideal observers, and to equate the statistical power of the z-scored classification vectors between human and ideal observers.

Results

Figure 4 shows classification vector results for human and ideal observers. The overall pattern of diagnostic spatial frequencies for individual subjects (thin gray lines) was quite consistent. The average correlation between each human observer and the group-averaged classification vector was $r = 0.88$, $SD = 0.06$. Similarly for individual ideal observers, based on the same number of simulated trials each, the average correlation was $r = 0.71$, $SD = 0.19$. The dotted line in Fig. 4 represents the threshold for statistical

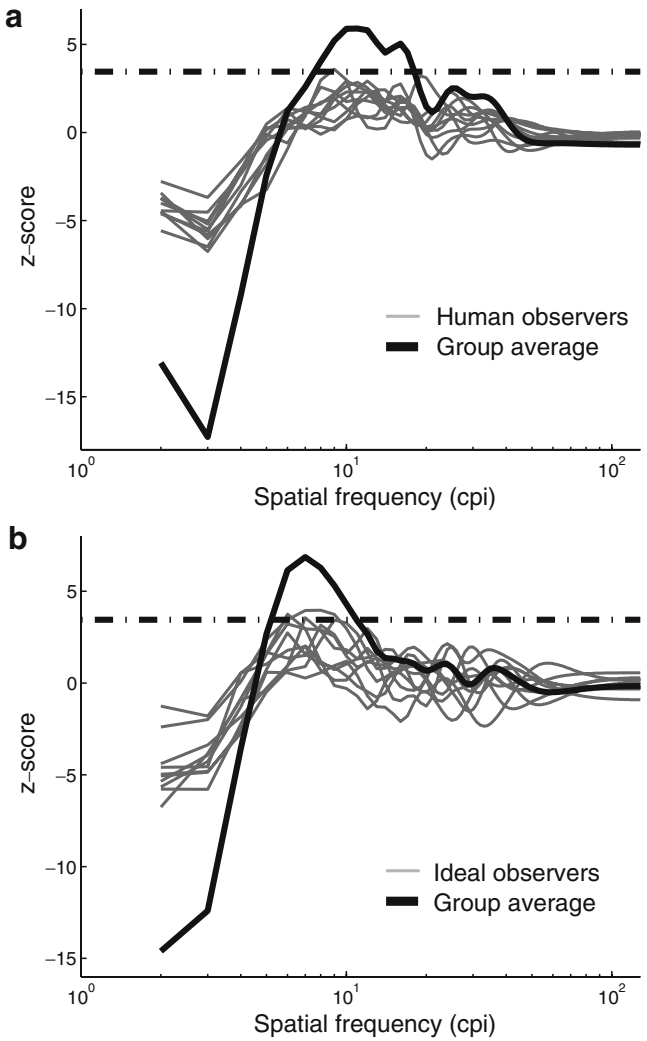

Fig. 4 Classification vector results using SF bubbles. **a**. *Z*-scored classification vectors for individual human observers and the group-averaged classification vector. **b**. Classification vectors for independent simulations of the ideal observer and the group-average. The dotted line represents the critical *z*-score level for statistical significance (Zcrit = 3.45, p = 0.05)

significance at the $p = 0.05$ level, corrected for multiple comparisons (*Z*crit = 3.45; see Chauvin, Worsley, Schyns, Arguin, & Gosselin, 2005 and Willenbockel et al., 2010)

For human observers, a statistically significant band of spatial frequencies centered at 12 cycles/image, with a bandwidth of 1.5 octaves (ranging from 6 to 18 cycles/image), was the most diagnostic for the action discrimination task. The center was estimated as the median SF out of all SF's in the cluster with *z*-scores above the statistical threshold. Figure 5 plots an example action video filtered with this significant diagnostic SF band. The statistically significant diagnostic spatial frequencies for the ideal observers were centered at 7 cycles/image, with a bandwidth of 1.5 octaves (ranging from 4 to 10 cycles/image). It is clear from the plots in Fig. 4 that spatial frequencies less than 5 cycles/image were the least diagnostic for both human and ideal observers. In fact, the extremely negative *z*-scores for the lowest spatial frequencies indicate that low SF information is very unreliable as compared to higher spatial frequencies, and that this pattern was highly consistent across individual observers. This is consistent with the result from Experiment 1 in which human observers only reached 43% accuracy even with the highest *s/n* level condition for band-pass filtered stimuli at 4 cycles/image.

Since the overall *s/n* level was adjusted using the staircase procedure to maintain threshold performance, we can estimate overall efficiency collapsed across all spatial frequencies. The average threshold *s/n* level for human observers was 0.0072 and the threshold *s/n* for ideal observers was 0.0001. Thus, overall human efficiency was 0.02%, which is comparable to the efficiency estimates from Experiment 1.

Discussion

In this study, we used two different methods to determine which spatial frequencies were most diagnostic for discriminating videos of human actions. SF tuning has been measured for a number of different objects and stimulus types, most commonly faces and letters (Gold et al., 1999), but this is the first experiment to our knowledge to measure SF tuning for dynamic natural scenes of human actions.

Results from Experiment 1, using band-pass filtered stimuli in identically filtered noise, revealed that human observers were most tolerant to noise in the SF band centered at 16 cycles/image. In terms of object-based spatial frequencies, this corresponded to 1.25 cycles/body width and 7.5 cycles/body height. While computing object-based spatial frequencies is relatively straightforward with stationary objects, such as static images of human faces, one challenge in estimating object-based spatial frequencies in the current study is due to variability in body size and body postures over time in the action videos. That is, the width and height of the human form varies across postures in the action sequence and slightly across different actors in the stimulus set, so we computed object-based spatial frequencies from an estimate of the average body width and height across all postures and actors. We acknowledge this issue and chose to plot the data in absolute terms of cycles/image, while making reference to rough estimates of the corresponding object-based spatial frequencies in the text.

In Experiment 2, the SF bubbles method was used to estimate SF tuning from classification vectors. The results of the group classification vector for human observers revealed a diagnostic 1.5-octave band of spatial frequencies

Fig. 5 Selected frames from an example video of walking filtered with the diagnostic SF band for human observers from Experiment 2. The transfer function for the diagnostic filter was created by replacing all spatial frequencies less than threshold from the group classification vector in Fig. 4a with zeros and then scaling the values above threshold to range from 0 to 1

centered at 12 cycles/image, or 0.94 cycles/body width, 5.6 cycles/body height. The peak SF from Experiment 1 (16 cycles/image) was contained within this diagnostic SF band, and it is evident by comparing Figs. 3 and 4 that the overall results from both experiments were very consistent.

By comparing the data from both experiments, it is clear that one significant advantage of the SF bubbles method over the band-pass filtering method is the resolution of the SF tuning estimates. With the SF bubbles method it is possible to determine the relative importance of the entire range of spatial frequencies, instead of just a few SF bands, and with fewer trials. Another benefit of the SF bubbles method is that all spatial frequencies are represented on each trial, just in different proportions, so observers are not able to adapt to particular SF bands during the experiment. Furthermore, the researcher does not have to choose the SF centers, bandwidths, and shape of the filters for a given experiment. Since these parameters are quite variable across different experiments (see Table 1 in Gold et al., 1999), it can be somewhat difficult to generalize across the previous experiments.

Another goal of the current experiment was to estimate human efficiency discriminating action videos to compare to efficiency estimates for other object types. In order to measure efficiency we had to obtain a metric of the available information in the action videos for the current task by implementing ideal observer analysis. We chose to employ a spatial correlator ideal observer (e.g. Gold et al., 1999; Parish & Sperling, 1991) that discriminated actions by maximizing the cross-correlation between a filtered test action sequence and the unfiltered template sequences. In Experiment 1 we found that human efficiency varied slightly across SF bands, but overall was quite poor. Efficiency peaked at 0.08% for SF band 2 (8 cycles/image) and varied between 0 and 0.06% efficiency for the other SF bands. The mean efficiency across all six SF bands was 0.034%. In Experiment 2 using the SF bubbles method, overall efficiency was estimated to be 0.02%. Efficiency estimates from each method were very similar and suggest that human observers are not very efficient at utilizing the information available to discriminate videos of human action.

Previous studies have identified a large range of human efficiencies for different discrimination tasks and experimental conditions. For instance, peak efficiency for letter identification has been estimated to be 13% (Solomon & Pelli, 1994), 42% (Parish & Sperling, 1991), and less than 1.5% (Gold et al., 1999). Gold et al. (1999) reported similarly low efficiencies for observers in a comparable face identification task. Our efficiency estimates in the current experiment are much lower than previous estimates for other object identification tasks, suggesting that observers had trouble utilizing all of the available information in the action stimuli.

One possible factor causing low efficiency in the current task was the short stimulus duration (500 ms). A reason for using this short duration was a limitation of the recorded action videos. We edited the videos to contain only frames in which a human actor was visible, and not off screen, so this limited us to only a 25 frame action segment across all videos in the stimulus set that we used (Gorelick et al., 2007). One possible extension of the current study would be to measure efficiency with a new set of action videos and with longer stimulus durations to get an estimate of the upper limit of efficiency under different conditions. An interesting follow-up study could also measure SF tuning at different viewing distances to determine if the observed critical SF band is object-based or retina-based. For instance, Parish and Sperling (1991) found that the critical SF band for letter identification was invariant with respect to viewing distance, suggesting that the critical SF information was in object-based coordinates (but see Chung et al., 2002; Loftus & Harley, 2005; Majaj et al., 2002; Nasanen, 1999; Willenbockel et al., 2010, Exp. 3a).

As mentioned previously, SF tuning measurements help to elucidate the scale of the critical features for object recognition (Sowden & Schyns, 2006). For instance action videos filtered in the highest SF bands show fine-grained features, such as the edges of the body and features of the face and clothing. In contrast lower SF bands, especially those that match the size of the body in the stimulus, contain large-scale features such as body form and articulation over time. Consistent with Kuhlmann and Lappe (2006), observers reliably used low spatial frequen-

cies that likely carried diagnostic information about the representation of the body form. The current experiment extends these results to show that observers also used intermediate spatial frequencies for action recognition. In fact, the effect of filtering with the most diagnostic SF band is an action video in which the limbs and body appear as a somewhat homogenous dark silhouette with a light patch of color surrounding the limbs (see Fig. 5), which likely aided in segmenting the body from background noise.

Further, it is clear by comparing the SF tuning curves of humans to the ideal observer in Experiment 2 that the most diagnostic SF band for human observers (6 to 18 cycles/image) was higher than for the ideal observer (4 to 10 cycles/image). As proposed by Chung et al. (2002), who also found an upward shift in SF tuning for humans as compared to the ideal observer in a letter discrimination task, the SF tuning estimate for humans likely reflects both the diagnostic information in the stimulus as well as physiological properties of the visual system. A parsimonious explanation for the critical spatial frequency range that we measured for human action recognition is that it reflects the best compromise between the optimal SF information for the task, which was measured with the ideal observer analysis, and the contrast sensitivity function inherent to the human visual system.

Acknowledgements This work was supported by a grant from the National Science Foundation (BCS0748314) to E. Grossman.

References

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10* (4), 433–436. doi:10.1163/156856897X00357

Campbell, F. W., & Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *Journal of Physiology, 197* (3), 551–566.

Cass, J., Alais, D., Spehar, B., & Bex, P. J. (2009). Temporal whitening: Transient noise perceptually equalizes the 1/f temporal amplitude spectrum. *Journal of Vision, 9*(10), 1–19. doi:10.1167/9.10.12

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision, 5*(9), 659–667. doi:10.1167/5.9.1

Chung, S. T., Legge, G. E., & Tjan, B. S. (2002). Spatial-frequency characteristics of letter identification in central and peripheral vision. *Vision Research, 42*(18), 2137–2152. doi:10.1016/S0042-6989(02)00092-5

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception and Psychophysics, 58*(4), 602–612.

De Valois, R. L., & De Valois, K. K. (1980). Spatial vision. *Annual Review of Psychology, 31,* 309–341. doi:10.1146/annurev.ps.31.020180.001521

Fiorentini, A., Maffei, L., & Sandini, G. (1983). The role of high spatial frequencies in face perception. *Perception, 12*(2), 195–201. doi:10.1068/p120195

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research, 39*(21), 3537–3560. doi:10.1016/S0042-6989(99)00080-2

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(12), 2247–2253. doi:10.1109/TPAMI.2007.70711

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41* (17), 2261–2271. doi:10.1016/S0042-6989(01)00097-9

Kuhlmann, S., & Lappe, M. (2006). Recognition of biological motion from blurred natural scenes. *Perception, 35*(11), 1495–1506. doi:10.1068/p5500

Loftus, G. R., & Harley, E. M. (2005). Why is it easier to identify someone close than far away? *Psychonomic Bulletin & Review, 12*(1), 43–65.

Majaj, N. J., Pelli, D. G., Kurshan, P., & Palomares, M. (2002). The role of spatial frequency channels in letter identification. *Vision Research, 42*(9), 1165–1184. doi:10.1016/S0042-6989(02)00045-7

Nasanen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research, 39*(23), 3824–3833. doi:10.1016/S0042-6989(99)00096-6

Norman, J., & Ehrlich, S. (1987). Spatial frequency filtering and target identification. *Vision Research, 27*(1), 87–96. doi:10.1016/0042-6989(87)90145-3

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72–107. doi:10.1006/cogp.1997.0667

Parish, D. H., & Sperling, G. (1991). Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Research, 31*(7–8), 1399–1415. doi:10.1016/0042-6989(91)90060-I

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. doi:10.1163/156856897X00366

Rubin, G. S., & Siegel, K. (1984). Recognition of low-pass faces and letters. *Investigative Ophthalmology and Visual Science, 25*(3), 96.

Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature, 369*(6479), 395–397. doi:10.1038/369395a0

Sowden, P. T., & Schyns, P. G. (2006). Channel surfing in the visual brain. *Trends in Cognitive Sciences, 10*(12), 538–545. doi:10.1016/j.tics.2006.10.007

Willenbockel, V., Fiset, D., Chauvin, A., Blais, C., Arguin, M., Tanaka, J. W., et al. (2010). Does face inversion change spatial frequency tuning? *Journal of Experimental Psychology Human Perception and Performance, 36*(1), 122–135. doi:10.1037/a0016465