An advantage for detecting dynamic targets in natural scenes

Quoc C. Vuong Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Andries F. Hof Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Heinrich H. Bülthoff Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Ian M. Thornton Department of Psychology, University of Wales Swansea, Swansea, United Kingdom

In the present study, we tested the extent to which observers use dynamic information to detect targets in natural scenes. For this purpose, we used composite stimuli in which target sequences were superimposed onto distractor sequences. We varied target visibility in the composite sequence, and the presence or absence of motion. Across four experiments, we found a dynamic advantage for target detection: Observers performed more accurately with dynamic than static target scenes. This advantage depended on the availability of target motion, irrespective of whether the target was upright or inverted in the image plane (Experiments 1–4). The magnitude of this advantage also depended on the availability of segmentation cues (Experiments 1 and 2) and on the distractors used (Experiments 2 and 4). Overall, the dynamic advantage reported extends previous work using isolated dynamic objects to more complex scenes.

Keywords: natural scenes, dynamic information, target detection

Introduction

The visual system of an active organism has to deal efficiently with a continuously changing environment. This capacity is important because it is needed to navigate through that environment and to recognize and interact with objects and other organisms. That we can often spot a friend in a crowd, for instance, attest to our ability to detect objects of interest in the scene despite a cluttered and dynamic background.

Both behavioral and physiological data suggest that human observers can rapidly process and interpret natural scenes despite their visual complexity (e.g., Johnson & Olshausen, 2003; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). For example, when observers are briefly flashed natural images (~20 ms) and are instructed to respond only when an animal is present in the image, responses made 280–290 ms after stimulus onset are mostly to target trials and not to distractor trials. Even more striking, brain signals evoked approximately 150 ms after stimulus onset reliably distinguish target trials from distractor trials. Observers can also rapidly detect targets in peripheral vision (Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001), and they can rapidly categorize scenes (e.g., as a forest, beach, or city street) that are briefly flashed and masked (e.g., Renninger & Malik, 2004; Schyns & Oliva, 1994). To date, researchers have predominantly used static images to study high-level processing of natural scenes (e.g., detecting specific categories of objects or categorizing scenes). What has not been investigated is how dynamic information in the environment might contribute to the processing of these scenes.

In the present study, we tested whether observers could use motion in the environment to help them detect target objects (e.g., a friend in a crowd). On the one hand, motion may simply add "noise" to the visual input (e.g., object features may deform or become occluded), and this could impair the observers' detection. On the other hand, given that the visual system evolved in a dynamic environment and is known to be sensitive to changing visual information, the availability of motion could facilitate detection. There are at least two ways that motion might facilitate performance. First, observers could use motion in the environment to segment objects from background clutter (e.g., Brady & Kersten, 2003; Cunningham, Shipley, & Kellman, 1998; for segmentation mechanisms in the fly visual system, see Bülthoff, 1981). Second, observers could detect specific motion patterns in the scene even if the visual input is noisy or degraded.

Several lines of research suggest that observers are well equipped to select and use characteristic patterns of motion in their environment. Some of the most compelling demonstrations come from work on biological motion

doi: 10.1167/6.1.8 Received July 15, 2004; published January 31, 2006 ISSN 1534-7362

(Johansson, 1973). Here it has been shown that observers can extract a range of information, such as the gender, identity, and even the emotional state of an actor, when movements are portrayed via a few points of light placed on the major joints. Similarly, the motion of other objects has been shown to play a direct role in object recognition. For example, several recent studies have found an advantage for dynamic image sequences over static images for stimuli such as faces (e.g., Pilz, Thornton, & Bülthoff, 2005; Thornton & Kourtzi, 2002) and novel objects (e.g., Vuong & Tarr, 2004). Typically, observers respond more quickly when matching a test stimulus to a preceding dynamic stimulus than to a preceding static image.

One common factor in the above studies is that target objects were always presented in isolation against a uniform background. In the current work, we asked whether a dynamic advantage could also be observed using complex, naturalistic scenes. Furthermore, we wanted to know whether such an advantage would rely on simple segmentation cues or on the detection of specific target patterns.

The target patterns we used were always walking human figures, which could either be static or in motion. These patterns were superimposed onto distractor scenes containing machines or animals, again either static or in motion. The visibility of the human scenes in the composite stimulus could be varied from trial to trial, and the observers' task in all experiments was to detect whether humans were present or absent on that trial. Observers might detect targets by first segmenting the composite stimulus (or solving the transparency problem that results from this manipulation) into its constituent scenes and then judging whether a human target is in one of them. In this case, elements in the constituent scenes may be grouped by their shared motion, thereby facilitating the segmentation of the composite stimulus. This segmentation process might or might not occur independently of the semantics of the motion (e.g., that it is a human or animal gait). In a series of experiments, we varied the presence or absence of motion (Experiments 1–4), the strength and availability of segmentation cues (Experiments 1, 2 and 4), and the familiarity of target motion (Experiment 3) to explore the relationship between segmentation and target-specific motion.

It is worth briefly commenting on our choice of human figures as a target category. There is now growing evidence from both behavioral (e.g., Reed, McGoldrick, Shackelford, & Fidopiastis, 2004) and neuroimaging studies (e.g., Grossman et al., 2000) that suggests that the perception of the human body may differ from the processing of other types of object (for a recent collection of related work, see Knoblich, Thornton, Grosjean, & Shiffrar, 2006). However, it is not this issue that is the main focus of the current study. Specifically, we are not directly concerned with comparing the detection of humans with other categories of objects: Our target objects are always human walkers. The choice of this target category was motivated primarily by practical issues, such as availability, uniformity, and familiarity. Human bodies all have a similar shape and they make highly familiar movements (e.g., walking) that we are very good at detecting (Johansson, 1973), even in noisy conditions (e.g., Cutting, Moore, & Morrison, 1988; Thornton, Pinto, & Shiffrar, 1998). These factors are likely to make human gait patterns highly salient in the composite stimulus and therefore a good target category to expose possible differences between dynamic and static stimuli. We return to the possible "special" status of human motion in Experiment 3.

In Experiment 1, we introduce the main task used to compare static and dynamic target detection using human/machine composites. Experiment 2 addresses the role of segmentation by adding a motion component to all stimuli, even when the target item in the composite is static. In Experiment 3, we inverted all the stimuli to gauge whether the observed dynamic advantage was limited to highly familiar or even preferentially processed stimuli. This manipulation can affect the high-level perception of motion (e.g., Pavlova & Sokolov, 2000; Sumi, 1984) and the neural structures subserving the processing of biological motion (e.g., Grossman & Blake, 2001) but it does not affect the low-level statistics of the image sequences, such as their Fourier amplitude or phase spectra (for a discussion, see Dong & Atick, 1995). Finally, in Experiment 4 we used animals as our distractors instead of machines to assess whether the target/distractor similarity influences performance. In all experiments, we found that dynamic target scenes resulted in better detection performance than static target scenes.

Experiment 1

The purpose of Experiment 1 was to test whether there is a dynamic advantage for detecting targets in natural scenes, as had been previously found for isolated faces and objects (Pilz et al., 2005; Thornton & Kourtzi, 2002; Vuong & Tarr, 2004). Thus, in this experiment we compared performance for detecting humans in natural image sequences and in natural images. On each trial, two composite sequences were presented briefly to both sides of fixation. The two sequences were either both moving or both static. On half the trials, one of the composite sequences contained a human figure, the visibility of which was systematically varied. The distractor images involved mechanical devices of various kinds. The observers' task was simply to report the presence or absence of the human target. If a dynamic advantage does occur for natural scenes, then we would predict greater accuracy in detecting dynamic human targets across all levels of visibility.

Method

Participants

Ten participants (three females/seven males) from the Tübingen community participated for pay in Experiment 1.

Stimuli

The stimuli consisted of composite image sequences created by averaging pixel intensity values on a frame-by-frame basis from two video clips, thereby superimposing the two sequences. The clips were natural scenes of either individual humans walking through a park or various machine movements. The scenes containing humans served as the target for this and all experiments. The original human and machine video clips were recorded with a SONY digital video recorder at 25 frames/s and subsampled to 12 frames/s. From each video clip, 36 frames were selected and resized to 150 × 112 pixels (4.8 × 3.7 deg). Two manipulations were applied to attenuate potential color and luminance differences between images. First, individual frames were converted to grayscale. Second, these images were equalized to have an approximately uniform intensity profile with a mean value of ~128. In total, there were 41 human sequences and 8 machine sequences. The machine sequences included moving cars, a spinning carousel, falling stones, an electric saw, pedals from a bicycle, and three separate moving machines in a factory. Note that there was a diversity of shapes and movements for the machines in comparison to the uniformity of these factors for human targets.

Figure 1 illustrates how composite sequences were constructed from the original 36-frame sequences. First, eight consecutive frames were randomly selected on a trial-by-trial basis from two sequences. Then for corresponding frames in the two 8-frame subsequences, a weighted average of the pixel value at spatially corresponding locations was taken. Specifically, images from one sequence were weighted by α, which can vary from 0 to 1, and images from the other sequence were weighted by $(1 - \alpha)$. If $\alpha = 0$ or $\alpha = 1$, then only one sequence is visible in the final composite stimulus. As α varies from 0 to 1, one sequence in the composite stimulus becomes more and more visible whereas the other sequence becomes less and less visible.

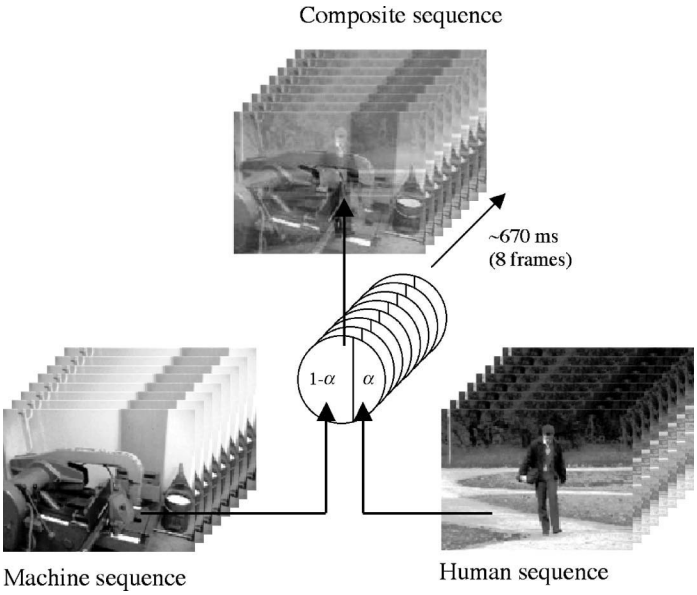


Figure 1. An illustration of how composite sequences were created from a weighted average of video clips of human figures and machines. The weight, α, controlled the visibility of human scenes in the composite stimulus.

Design

The observers' task in this and subsequent experiments was always to detect human targets in composite sequences at different levels of α, which controlled the visibility of human target scenes. We used a 2 (target present, target absent) × 2 (dynamic target, static target) × 7 (levels of α from .20 to .44 in 0.04 steps) within-subjects design. There were 20 repetitions in each of these conditions, for a total of 560 trials.

On target-present trials, the composite sequence was created from a human sequence and a machine sequence. Human sequences were randomly selected from the 41 video clips available. Likewise, machine sequences were randomly selected from the eight clips available.

On dynamic target trials, the composite sequence consisted of a human sequence averaged with a machine sequence. On static target trials, the composite stimuli consisted of a human image averaged with a machine image. The static images were randomly selected from the 36 possible frames of both human and machine sequences on each trial. Note that there is no image motion on static target-present and target-absent trials.

Procedure

Each trial began with a black fixation cross at the center of a gray background, followed 1000 ms later by a 750-Hz warning tone. On dynamic trials, two composite sequences were simultaneously presented 1.5 deg to the left and right of the fixation cross after the tone (cf. Thorpe et al., 2001). The display was presented for approximately 670 ms; that is, each of the eight images comprising the sequences was presented for approximately 80 ms. On static trials, two composite images were simultaneously presented to the left and right of fixation after the tone for ~670 ms.

On target-present trials, the composite stimuli containing the human scene appeared equally often to the left or right of fixation. The other stimulus was a composite of two different machines presented in the same condition (i.e., dynamic or static) as the composite stimulus containing the target. In addition, one of the machines had the same visibility as the human target. On target-absent trials, a randomly selected machine sequence replaced the human sequence. The purposes of presenting two composite sequences simultaneously were, first, to make the detection task more difficult and, second, to prevent observers from fixating at a particular location of the monitor (as all humans were filmed in the center of the scene).

Observers were instructed to maintain fixation as best as possible on each trial (eye movements were not monitored). They were further instructed to respond "present" only if they were highly confident that a human was present in one of the two composite stimuli. They made their responses by pressing either a "present" or "absent" key with their left and right hand, which was counterbalanced across observers. Observers took a self-timed break after every 140 trials. There were also 16 practice trials at the beginning of the experiment to familiarize observers with the task and the response keys. For these trials, $\alpha = .32$ and .44 were used.

The experiment was conducted in a dimly lit room. Observers sat approximately 60 cm from a 21-in. SONY Trinitron monitor, which had an 1152 × 870 pixel resolution and 75 Hz refresh rate. Nothing was used to constrain the observers' head movement. A G4 Mac running PsychToolbox (Brainard, 1997; Pelli, 1997) was used to control stimulus presentation and data collection.

Results and discussion

There was no systematic pattern in the false alarm rates on target-absent trials (responding "present"). Furthermore, our initial analyses of response times only revealed a significant effect of target visibility. As expected, observers responded more quickly as the targets became more visible, with no difference in response times between dynamic and static targets. These findings probably reflect the fact that observers were asked to respond "present" only if they were confident they saw a target. Thus, in this and subsequent experiments, we only analyzed hit rates.

Figure 2 shows the hit rates for dynamic target and static target trials as a function of the visibility of the human target for Experiment 1. The solid lines in this and subsequent figures are the linear regressions to the average data. The hit rates were then submitted to a 2 × 7 repeated measures analysis of variance (ANOVA) with target type (dynamic target, static target) and target visibility (α) as within-subjects factors. Not surprisingly, there was a main effect of target visibility, $F(6,54) = 78.58$, $p < .001$. More importantly, however, the main effect of target type was significant, $F(1,9) = 27.65$, $p < .01$. Observers were more accurate with dynamic targets ($M = 56.4\%$, $SE = 3.6\%$) than with static targets ($M = 37.5\%$, $SE = 4.7\%$).

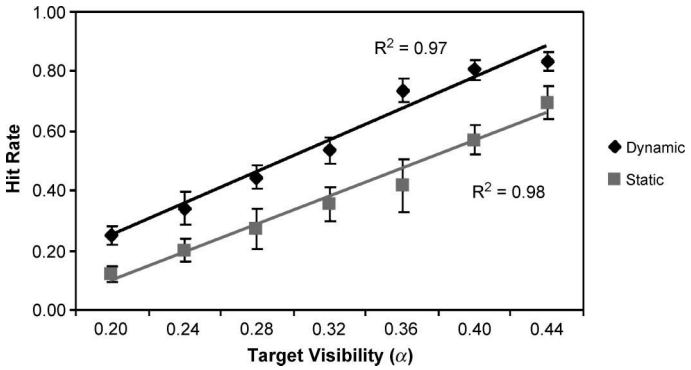


Figure 2. The mean hit rates for detecting targets in Experiment 1. The mean hit rates are fitted by linear regression (solid lines). Error bars in this and subsequent figures represent ±1 *SEM*.

The main finding in Experiment 1 is a dynamic advantage for detecting targets in natural scenes, which extends previous work using isolated faces and objects (Pilz et al., 2005; Thornton & Kourtzi, 2002; Vuong & Tarr, 2004). A second finding in this experiment is that motion from the target or the distractor scenes does not add "noise" to the visual input. On the contrary, the availability of motion improved the observers' performance.

Experiment 2

While the results of Experiment 1 clearly show a dynamic advantage for detecting targets in natural scenes, it is unclear exactly what role motion is playing in this task. Specifically, is the advantage simply due to the motion aiding segmentation or are observers also sensitive to the specific motion patterns associated with the target scenes? To address this question, in Experiment 2 both dynamic and static human target scenes were always superimposed onto dynamic machine scenes. If observers simply used target and distractor motion to segment the composite scenes, then they should be equally accurate on dynamic target and static target trials, as segmentation is possible in both cases. By comparison, if observers were also sensitive to motion from target scenes, then they should perform better on dynamic target trials.

Movies 1 and 2 illustrate an example of the distinction between dynamic target and static target trials in Experiment 2. For these movies, the same human target was averaged with the same machine sequence with $\alpha = .32$.

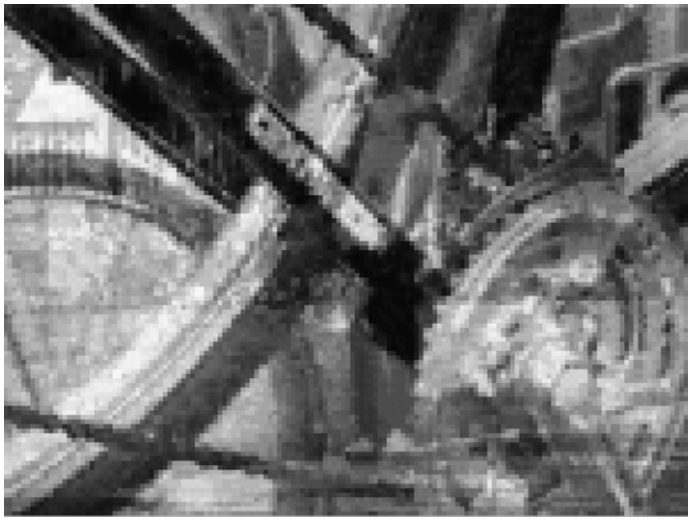
Method

Participants

Ten new participants (five females/five males) from the Tübingen community participated for pay in Experiment 2.

Stimuli

The same human and machine sequences from Experiment 1 were used in Experiment 2. The critical change in this experiment was how composite stimuli on static target trials were created. On each of these trials, a human target was first randomly selected from one of the 41 possible sequences, and then a single image was randomly selected from this sequence. Lastly, the selected human

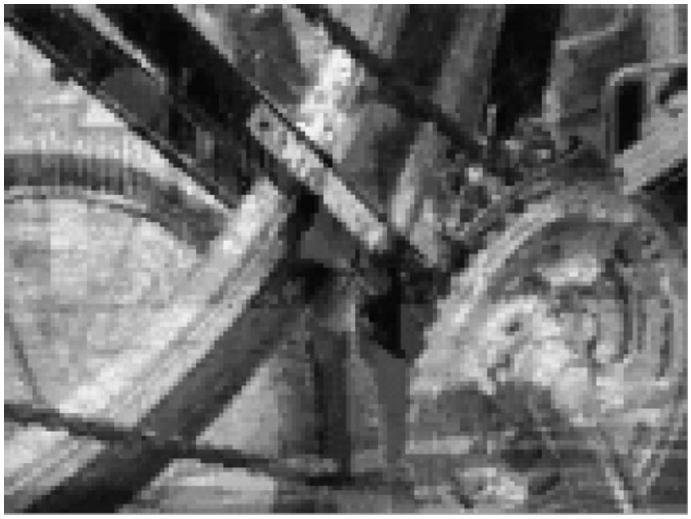


Movie 1. A composite sequence of a dynamic human target averaged with a machine (bicycle pedal) with $\alpha = .32$.

image was averaged with each of the different images from a randomly selected machine sequence. On static-absent trials, the same procedure was used but a randomly selected machine image replaced the human image. Thus, there was image motion on every trial (i.e., from the machine distractor sequence), but there was target motion only on dynamic target trials (i.e., from the human target sequence).

Design and procedure

The design and procedure in Experiment 2 were identical to those of Experiment 1. Again, there were two composite sequences presented simultaneously on each trial. On target-present trial, one of the composite sequences contained a human target. The other composite sequence consisted of two different machine sequences, one of which was presented in the same condition as the human target (i.e., dynamic or static).



Movie 2. A composite sequence of a static version of the same human target averaged with the same machine as in Movie 1, again with $\alpha = .32$.

Results and discussion

Figure 3 shows the hit rates for dynamic and static targets as a function of the visibility of the human target for Experiment 2. The hit rates were submitted to a 2×7 repeated measures ANOVA with target type (dynamic target, static target) and target visibility (α) as within-subjects factors. The pattern of results in Experiment 2 is similar to the pattern found in the first experiment. As in Experiment 1, there were main effects of target type, $F(1,9) = 28.38$, $p < .01$, and target visibility, $F(6,54) = 82.41$, $p < .001$. Observers were overall more accurate with dynamic targets ($M = 49.8\%$, $SE = 3.1\%$) than with static targets ($M = 40.6\%$, $SE = 4.2\%$).

The main finding in Experiment 2 was a dynamic advantage for detecting human targets in natural scenes, which replicates the results of Experiment 1. This advantage was found despite the fact that segmentation cues were available on both dynamic target and static target trials. On dynamic target trials, the human and machine scenes could be segmented based on the shared motion of elements in these separate scenes—that is, target and distractor motion is used solely to segment the composite stimulus. Likewise, a similar segmentation process could operate on static target trials if we consider that elements in the static human scenes have a shared null motion. That we found a dynamic advantage in this experiment suggests that observers also used target motion to perform the detection task. Given that human movements are the most salient movements in the target scenes as outlined in the Introduction, we believe that observers relied predominantly on these movements in the scene. It is possible that observers also used movements of background elements but we remain neutral on this issue.

A second finding was that the magnitude of the dynamic advantage was smaller in this experiment than in

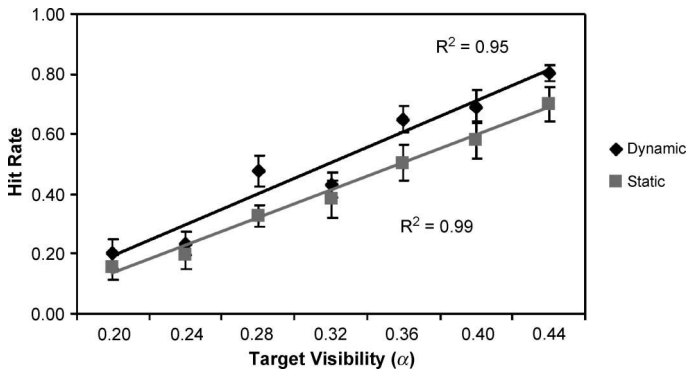


Figure 3. The mean hit rates for detecting targets in Experiment 2.

Experiment 1. This finding suggests that being able to segment the composite stimulus also contributes to target detection. Overall, there was an average of 18.9% difference between dynamic target and static target trials in the first experiment but only a 9.1% difference in this experiment.

Experiment 3

In Experiment 3, we asked whether the dynamic advantage found in Experiments 1 and 2 depended specifically on the use of human targets. As mentioned in the Introduction, there is now considerable evidence to suggest that observers are particularly sensitive to this form of motion (Knoblich et al., 2006). Thus, it is possible that the current dynamic advantage would not generalize to other forms of target stimuli. To explore this issue, we made use of a well-known manipulation that has been shown to preferentially disrupt the processing of human motion patterns. This manipulation simply involves picture plane inversion of the stimuli (e.g., Grossman & Blake, 2001; Pavlova & Sokolov, 2000; Sumi, 1984). We chose to invert the stimuli rather than changing the target category because inversion preserves all other spatial and temporal characteristics of the stimuli relative to the upright versions. Note that we inverted both the human and machine scenes in the composite. We did this to reduce the possibility that observers detected a conflict in the alignment of the top–bottom axis in the composite stimuli.

If the dynamic advantage in the previous experiments relied on the preferential processing of canonical human motion, inverting the composite stimulus should impair the observers' ability to detect this motion pattern. We would therefore not expect to find a difference between dynamic target and static target trials. However, if observers detected a task-defined target motion (i.e., inverted human motion), then we would still expect to find a difference between dynamic target and static target trials.

Method

Participants

A new group of 10 observers (6 females/4 males) from the Tübingen community participated in Experiment 3 for pay.

Stimuli

The stimuli used in Experiment 3 were identical to those used in Experiment 2, with the exception that the composite stimuli on each trial was inverted (flipped 180 deg out of the image plane rather than rotated 180 deg in the image plane).

Design and procedure

The design and procedure for Experiment 3 was again identical to that of Experiment 2. All observers were informed that the stimuli would be inverted on every trial.

Results and discussion

Figure 4 shows the hit rates for dynamic and static targets as a function of the visibility of the human target for Experiment 3. As in previous experiments, the hit rates were submitted to a 2 × 7 repeated measures ANOVA with target type (dynamic target, static target) and target visibility (α) as within-subjects factors. We again found main effects of target type, $F(1,9) = 19.89$, $p < .01$, and target visibility, $F(6,54) = 41.58$, $p < .001$. The accuracy on present trials was higher for dynamic targets ($M = 56.0\%$, $SE = 2.7\%$) than for static targets ($M = 45.1\%$, $SE = 2.7\%$).

We also compared the hit rates from Experiments 2 and 3 in a 2 × 2 × 7 mixed design ANOVA with target orientation (upright, inverted) as a between-subjects factor and target type (dynamic target, static target) and target visibility (α) as within-subjects factors. We found significant effects of target type, $F(1,18) = 44.77$, $p < .05$, and target visibility, $F(6,108) = 184.12$, $p < .001$. We also found a significant interaction between target orientation and target visibility, $F(6,108) = 4.32$, $p < .001$, but there was no consistent pattern in the data to allow us to interpret this interaction (compare Figures 3 and 4). Lastly, there was no main effect of target orientation, $F(1,18) = 0.88$, *ns,* suggesting that observers were equally accurate with upright and inverted human scenes. However, there was a slight indication that observers were more accurate with inverted targets ($M = 52.1\%$, $SE = 2.2\%$) than with upright targets ($M = 46.4\%$, $SE = 2.2\%$).

Overall, we found a dynamic advantage for detecting inverted human scenes, suggesting that the dynamic advantage does not necessarily rely on the preferential processing of upright human gait patterns. Rather, we think that observers quickly learn to detect unfamiliar inverted

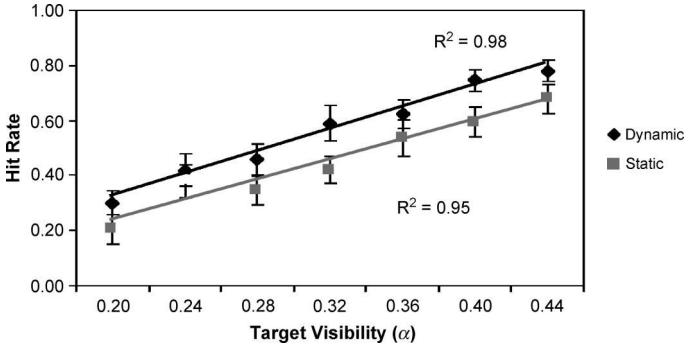


Figure 4. The mean hit rates for detecting targets in Experiment 3.

human motion, as this was the task-defined target. Although inversion is known to impair the processing of human gait, observers can quickly adapt to this manipulation (e.g., Grossman & Blake, 2001; Pavlova & Sokolov, 2000; Sumi, 1984).

The lack of an inversion effect in the current experiment does raise the possibility that the dynamic advantage could be driven by low-level differences between dynamic target and static target trials. For example, differences in their amplitude spectra can differentially affect their stimulus contrast (e.g., Johnson & Olshausen, 2003). We believe, however, that the histogram equalization manipulation attenuates such differences in the composite stimuli. Furthermore, our hypothesis is that target detection is based on both segmentation and detection of dynamic patterns. In conjunction with our previous results, the data suggest that segmentation cues alone do not sufficiently account for the dynamic advantage.

Movie 4. A composite sequence of the static human target of Movie 2 averaged with the zebra $\alpha = 0.32$.

Experiment 4

To further test the extent to which observers' performance depended on being able to detect target motion, in Experiment 4 we used animals instead of machines as our distractors. Animals were chosen because, first, animal movements are more similar to human movements than machines (e.g., both animals and humans have articulations of identifiable limbs) and, second, animals have similar curvilinear contours as humans as opposed to the rectilinear contours of machines. The similarity of animal motion to human motion has two potential consequences: First, the similarity may make it more difficult to segment the composite stimulus, and second, the similarity may make it more difficult to detect target motion. If observers were simply segmenting the composite stimulus based on shared motion in the constituent scenes, then they might show the same dynamic advantage found in Experiments 1 and 3. By comparison, if observers were also sensitive to target motion, then the magnitude of this advantage may be reduced because the target/distractor similarity may reduce their ability to segment and to detect target motion.

Again we present movies to demonstrate what observers perceived. Movies 3 and 4 illustrate a dynamic and static version of the same human target as in Movies 1 and 2 superimposed onto a zebra sequence with $\alpha = .32$.

Movie 3. A composite sequence of the dynamic human target of Movie 1 averaged with an animal (zebra) with $\alpha = .32$.

Method

Participants

A last group of 10 observers (8 females/2 males) from the Tübingen community participated for pay in Experiment 4.

Stimuli

The animal clips for Experiment 4 were taken from professionally recorded wildlife footages (Wild Paradise series). These had a frame rate of 25 frames/s but we subsampled them to 12 frames/s. Images from these clips were post-processed in the same manner as the original human and machine video clips (conversion to grayscale, luminance histogram equalization). The animal sequences included fighting vultures, a walking lion, a walking zebra, a walking coyote, a walking cheetah, a walking bear, swimming fish, and a running antelope.

Design and procedure

The design and procedure for Experiment 4 was identical to that of Experiment 2.

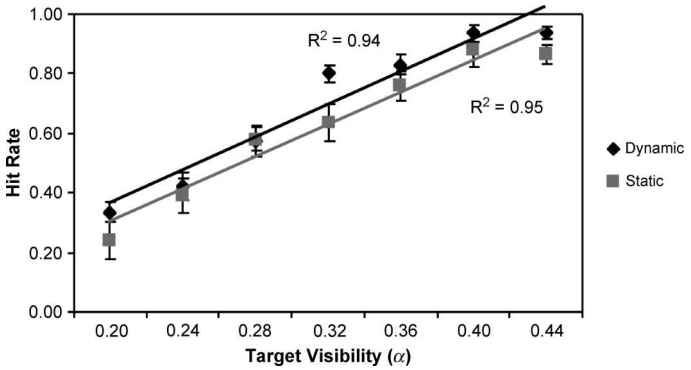


Figure 5. The mean hit rates for detecting targets in Experiment 4.

Results and discussion

Figure 5 shows the hit rates for dynamic and static targets as a function of the visibility of the human target for Experiment 4. The hit rates for Experiment 4 were submitted to a 2 × 7 repeated measures ANOVA with target type (dynamic target, static target) and target visibility (α) as within-subjects factors. As in the previous experiments, we found main effects of target type, $F(1,9) = 17.00$, $p < .01$, and target visibility, $F(6,54) = 106.13$, $p < .001$. The overall accuracy on present trials was higher for dynamic targets ($M = 69.1\%$, $SE = 1.6\%$) than for static targets ($M = 62.1\%$, $SE = 2.2\%$).

As in Experiment 3, we compared the results across Experiments 2 and 4, which only differed in the distractor used. For this analysis, the hit rates from these two experiments were submitted to 2 × 2 × 7 mixed design ANOVA with distractor type (machine distractor, animal distractor) as a between-subjects factor and target type (dynamic target, static target) and target visibility (α) as within-subjects factors. We found significant effects of distractor type, $F(1,18) = 29.56$, $p < .001$, target type, $F(1,18) = 44.77$, $p < .05$, and target visibility, $F(6,108) = 184.12$, $p < .001$. Importantly, we found a significant interaction between distractor type and target visibility, $F(6,108) = 4.32$, $p < .001$. Thus, although observers still showed an advantage for detecting dynamic targets, the type of distractor modulated this advantage. Lastly, we point out that the present results cannot be strictly based on detecting curvilinear (humans and animals) versus rectilinear contours (machines).

General discussion

In the current series of experiments, observers were shown composite stimuli and had to detect targets whose visibility was systematically varied. Our key manipulation was whether scenes containing targets were presented dynamically or statically. The critical finding across all experiments was a dynamic advantage for detecting targets in natural scenes. That is, observers were more accurate at detecting targets presented as image sequences as opposed to the same targets presented as static images. The results suggest that this advantage is based on the detection of dynamic cues available in the target scenes rather than on strictly static cues (Experiments 1–3). We also found that the magnitude of this advantage was modulated by the availability of segmentation cues (Experiments 1 and 2) and the type of distractors used (Experiments 2 and 4).

A similar dynamic advantage was previously found for faces and novel objects (Pilz et al., 2005; Thornton & Kourtzi, 2002; Vuong & Tarr, 2004). However, a potential limitation with these earlier studies was that the dynamic objects used were presented in isolation against a uniform background. By comparison, natural scenes are cluttered with irrelevant objects and irrelevant movements.

Thus, the present results help extend the role of dynamic information to the processing of natural scenes. Under natural viewing conditions, we must be able to segment the scene—a process which, in itself, is a challenging problem with or without motion for any organism (e.g., Brady & Kersten, 2003; Bravo & Farid, 2004; Bülthoff, 1981). Arguably, the method of averaging pixel values of two images produces an effect of transparency and might therefore render the composite stimulus somewhat "unnatural". To address this point, we note two observations. First, the composite stimuli have a similar amplitude spectrum as natural image sequences (e.g., Dong & Atick, 1995). Second, for humans there are natural occurrences of transparency as when an indoor scene is partially reflected in a window (e.g., Kersten, 1991). That said, we are using other paradigms to further examine the contributions of dynamic cues to target detection without introducing this transparency artifact. For example, we are currently using a visual search task to compare how efficiently different types of movements (e.g., humans versus animals) are processed. We believe that this task would provide a more direct test for whether the visual system is sensitive to task-specified target motion.

Although dynamic cues in target scenes seem critical to the dynamic advantage reported here, the present results indicate that being able to segment the composite scenes into their constituent scenes also contribute to target detection. This segmentation process can be aided by the availability of shared motion of elements in the constituent scenes and does not necessarily depend on the semantics of the motion. In fact, early studies on the fly visual system demonstrated that insects can separate moving targets from stationary backgrounds or stationary targets from moving backgrounds (e.g., Bülthoff, 1981; Reichardt & Poggio, 1979). These studies suggest that segmentation can be achieved by comparing the outputs of simple motion detectors (Reichardt, Poggio, & Hausen, 1983). These low-level segmentation mechanisms can

provide an account for some of the current findings. In particular, low-level segmentation cues could explain why observers in Experiment 2 performed better in the static target condition than observers in Experiment 1.

Based on the results across the four experiments, however, we argue that more complex mechanisms must be involved for detecting targets in natural scenes. First, observers performed better with dynamic targets than with static targets, even if there was image motion in both dynamic and static conditions to equate the availability of segmentation cues (Experiments 1–4). Second, observers appeared to be sensitive to the type of distractors used (Experiments 2 and 4). Thus, our stimuli and task seem to require a high-level interpretation of objects in the scene (humans, machines, and animals) and possibly some knowledge of their motion patterns.

The results of Experiment 3 further suggest that the ability to interpret dynamic patterns is not necessarily restricted to highly familiar types of motion. Observers in this experiment were quickly able to learn and use unfamiliar inverted target motions to accomplish the task. Based on the combined results from Experiments 1–4, we think that the current dynamic advantage should generalize to other target categories, a prediction that we are currently in the process of testing.

One issue that has not been directly addressed in the current study are differences in the information content between dynamic and static trials. That is, on dynamic trials the spatial relations of features in the composite image would be different from frame to frame, thereby providing multiple independent samples of the stimulus, which can then be used to detect targets. By comparison, on static trials the same image is presented for the entire stimulus duration, thereby providing only a single sample. Our current task does not really allow us to equate information content in the absence of motion. However, in previous studies of dynamic effects in recognition, the simultaneous presentation of multiple images has not been shown to increase performance in the same manner as the presentation of an image sequence (e.g., Lander & Bruce, 2000; Pilz et al., 2005; Wallis & Bülthoff, 2001). Indirectly, we can note that in Experiment 2, the addition of distractor motion during static target trials will have changed the spatial relations between image features of the target and the machine. However, this frame-to-frame change in these static relations did not improve performance.

Finally, an important goal for future studies is to understand the neural mechanisms underlying the present dynamic advantage. Earlier studies have used the fly visual system as a model to understand image segmentation from dynamic cues (e.g., Bülthoff, 1981; Reichardt & Poggio, 1979; Reichardt et al., 1983). More recently, human brain imaging studies have found that the superior temporal sulcus (STS) is involved in high-level interpretations of visual motion. For instance, this cortical area is responsive to facial movements (e.g., Puce, Allison, Bentin, Gore, & McCarthy, 1998) and human articulation from point-light displays (e.g., Grossman et al., 2000). Furthermore, earlier animal studies suggest that the monkey homologue of this area (STPa) pools information from both ventral and dorsal streams (Oram & Perret, 1994). These streams are believed to predominantly process static and motion information respectively (Ungerleider & Mishkin, 1982). Thus, it would be interesting to study STS in the context of our experimental paradigm.

To conclude, our visual system is faced with the difficult task of interpreting a complex changing visual input to navigate through the environment and to recognize objects. We believe that the dynamic nature of the visual input is an important source of information that allows us to successfully perform these tasks. Consistent with this claim, the present results clearly demonstrate that the addition of target motion in complex scenes improves the observers' ability to detect those targets. It remains a matter for future research to determine the exact nature of the dynamic information that is extracted by observers in this task.

Acknowledgments

This work was supported by the Max Planck Gesellschaft. We would like to thank Nils Aguilar for filming the human and machine scenes and Scott McDonald and Douglas Cunningham for helpful discussions.

Commercial relationships: none.
Corresponding author: Quoc C. Vuong.
Email: quoc.vuong@tuebingen.mpg.de.
Address: Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany.

References

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. *Journal of Vision, 3*(6), 413–422, http://journalofvision.org/3/6/2/, doi:10.1167/3.6.2. [PubMed] [Article]

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. [PubMed]

Bravo, M. J., & Farid, H. (2004). Recognizing and segmenting objects in clutter. *Vision Research, 44*(4), 385–396. [PubMed]

Bülthoff, H. H. (1981). Figure-ground discrimination in the visual system of Drosophila melanogaster. *Biological Cybernetics, 41,* 139–145.

Cunningham, D. W., Shipley, T. F., & Kellman, P. J. (1998). The dynamic specification of surfaces and boundaries. *Perception, 27*(4), 403–415. [PubMed]

Cutting, J. E., Moore, C., & Morrison, R. (1988). Masking the motions of human gait. *Perception and Psychophysics, 44*(4), 339–347. [PubMed]

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural Systems, 6,* 345–358.

Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. *Vision Research, 41*(10–11), 1475–1482. [PubMed]

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in the perception of biological motion. *Journal of Cognitive Neuroscience, 12*(5), 711–720. [PubMed]

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14,* 201–211.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*(7), 499–512, http://journalofvision.org/3/7/4/, doi:10.1167/3.7.4. [PubMed] [Article]

Kersten, D. J. (1991). Transparency and the cooperative computation of scene attributes. In Landy M. & Movshon A. (Eds.), *Computational models of visual processing,* 209–228. Cambridge, MA: MIT Press.

Knoblich, G., Thornton, I. M., Grosjean, M., & Shiffrar, M. (Eds.). (2006). *The human body: Perception from the inside out*. New York, NY: Oxford University Press.

Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology, 12*(4), 259–272.

Oram, M. W., & Perrett, D. I. (1994). Response of anterior superior temporal polysensory (STPa) neurons to "biological motion" stimuli. *Journal of Cognitive Neuroscience, 6*(2), 99–116.

Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception and Psychophysics, 62*(5), 889–899. [PubMed]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. [PubMed]

Pilz, K. S., Thornton, I. M., & Bülthoff, H. H. (2005). A search advantage for faces learned in motion. *Experimental Brain Research.* [PubMed]

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in human viewing eye and mouth movements. *Journal of Neuroscience, 18*(6), 2188–2199. [PubMed] [Article]

Reed, C. L., McGoldrick, J. E., Shackelford, J. R., & Fidopiastis, C. M. (2004). Are human bodies represented differently from other objects? Experience shapes object representations. *Visual Cognition, 11*(4), 523–550.

Reichardt, W. E., & Poggio, T. (1979). Figure-ground discrimination by relative movement in the visual system of the fly: Part I. Experimental results. *Biological Cybernetics, 35,* 81–100.

Reichardt, W. E., Poggio, T, & Hausen, K. (1983). Figure-ground discrimination by relative movement in the visual system of the fly: Part (I). Towards the neural circuitry. *Biological Cybernetics Supplement, 46,* 1–30.

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44*(19), 2301–2311. [PubMed]

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Sumi, S. (1984). Upside-down presentation of the Johansson moving light-spot pattern. *Perception, 13*(3), 283–286. [PubMed]

Thornton, I. M., & Kourtzi, Z. (2002). A matching advantage for dynamic faces. *Perception, 31*(1), 113–132. [PubMed]

Thornton, I. M., Pinto J., & Shiffrar, M. (1998). The visual perception of human locomotion. *Cognitive Neuropsychology, 15,* 535–552.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522. [PubMed]

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*(5), 869–876. [PubMed]

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingel, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorization of natural and artifactual objects. *Perception, 30*(6), 655–668. [PubMed]

Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research, 44*(14), 1717–1730. [PubMed]

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the USA, 98*(8), 4800–4804. [PubMed] [Article]