13

Multimodal Perception and Simulation

Peter Werkhoven and Jan van Erp

This chapter discusses mechanisms of multimodal perception in the context of multimodal simulators and virtual worlds. We review some notable findings from psychophysical experiments with a focus on what we call *touch-inclusive multimodal perception*—that is, the sensory integration of the tactile system with other sensory systems such as vision and hearing.

The Relevance of Understanding Multimodal Perception

Humans have evolved to interact with their natural environment through multiple and highly sophisticated sensory systems, each consisting of dedicated receptors, neural pathways, and specialized parts of the brain in which processing takes place. Human sensory systems for vision, hearing (audition), touch (somatic sensation), taste (gustation), and smell (olfaction) enable us to sense physical properties from different sensory modalities such as light, sound, pressure, the flavor of substances, and volatile chemicals. Another sensory system, the vestibular system, lets us sense the gravitational force and our body accelerations.

Not trivially, our brains merge the information derived from the various sensory systems into a coherent and unambiguous multisensory percept of the world. They constantly process voluminous and parallel streams of sensory information and try to relate those sensory signals that originate from the same event, regardless of their modality. To understand the mechanisms underlying multimodal perception, we need to understand not only how information from each individual sensory modality is processed but also how information from one sensory system is integrated with and modulated by other sensory systems.

A thorough understanding of the mechanisms of multimodal perception has become of particular interest to the community of developers of rapidly evolving multimodal simulators and virtual worlds.

DOI: 10.1037/14135-014
Human Information Processing: Vision, Memory, and Attention, C. Chubb, B. A. Dosher, Z.-L. Lu, and R. M. Shiffrin (Editors)

Multimodal Simulation

Today we see interactive multimodal simulators that combine three-dimensional (3D) sound and vision, tactile feedback, and high-tech motion platforms. Among the most advanced simulators in the world is the multipurpose and multimodal simulator Desdemona, developed by the research institute TNO (Toegepast Natuurwetenschappelijk Onderzoek) Human Factors Institute in the Netherlands in collaboration with the Austrian company AMST Systemtechnik (see Figure 13.1). It has been designed to realistically simulate complex movements ranging from F16 dog fights to off-road vehicle driving and even roller coasters. It has a cabin that can contain an F16 cockpit, mounted on a fully gimbaled system that is able to rotate around any conceivable axis. The system as a whole allows 2 m of vertical movement, combined with 8 m along a horizontal sledge. The sledge itself is able to spin as well. Centrifugation enables Desdemona to generate constant G-forces up to a maximum of 3 G. Desdemona is also used for experiments on multimodal interfaces that support the pilot in keeping spatial orientation and situation awareness.

Research focuses on how to provide people with consistent multisensory cues for spatial orientation (proprioceptive, vestibular, visual, auditory) with the ambition of arriving at accurate sensory integration models (using Bayesian frameworks) that can account for the user variability observed in psychophysical experiments. The complexity of such models may be illustrated by the findings of Mesland, Bles, Werkhoven, and Wertheim (1998), who investigated the

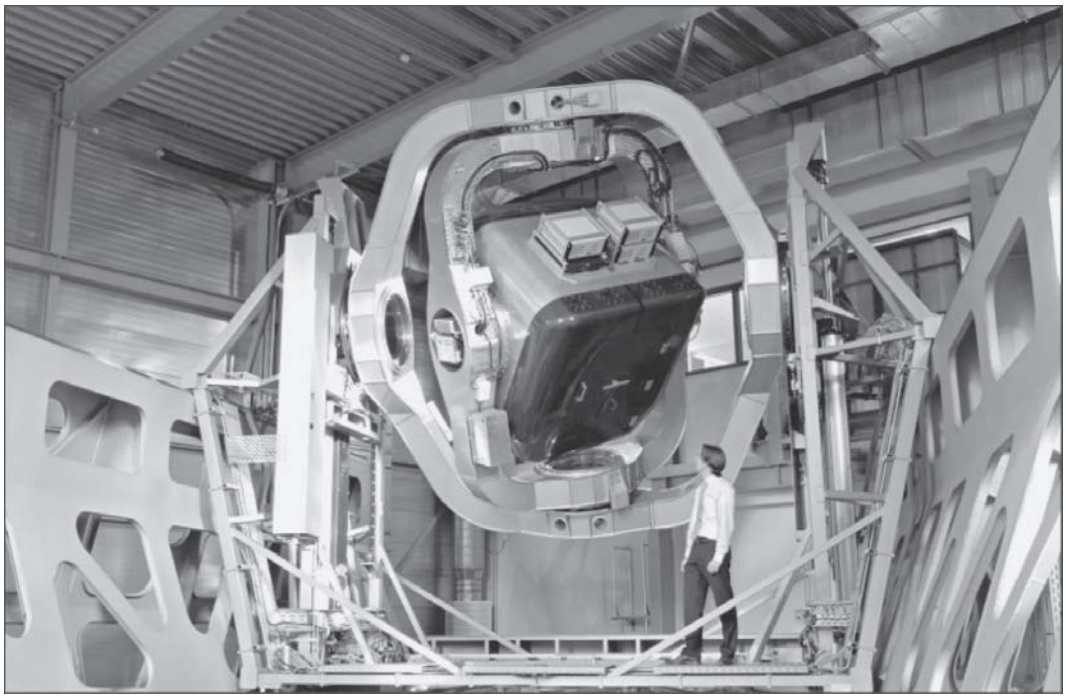

Figure 13.1. The multipurpose and multisensory motion simulator Desdemona, developed by TNO and AMST.

percept of passive horizontally oscillating self-motion simulated using a combination of a linear horizontal accelerator ("sled") and head-mounted displays. Remarkably, the self-motion percept for such simulation gained in quality not when the visual and proprioceptive stimuli were correctly in phase but when the visual stimulus had a small lead.

Parallel to the development of professional simulators, we see the emergence of 3D multimodal virtual game environments for social interaction, concept development, decision making, and learning. Such game environments allow people to interact with virtual worlds similar to the way they act in the real world and allow them to experience the consequences of their actions. Interaction and consistent multimodal representations (Ernst & Bulthoff, 2004) are crucial for this process of learning by doing. Consistency issues are even more complicated in the case of augmented worlds in which virtual and real worlds are combined. For example, using see-through displays in combination with auditory and tactile displays, we can perceive virtual objects embedded in our real environment. Virtual objects must behave correctly in this real world with respect to visual perspective, occlusion, shading, sound, and touch. However, technical limitations (spatiotemporal resolutions and dynamic ranges) and principle limitations (e.g., constraints of color spaces) still yield many inconsistencies.

Altogether, it is of crucial importance to have sufficient knowledge about human tolerances for inconsistencies between modalities and about modality interference effects. Furthermore, knowledge about multisensory illusions and metameric classes of sensory stimuli may lead to alternative and more feasible ways of creating a similar percept.

Sensory Substitution and Synesthetic Media

So far, we have reflected on how to convey visual properties of simulated environments to the visual sense, auditory properties to the auditory sense, and tactile properties to the touch senses in multimodal simulation. There is, however, a growing interest in exploiting our senses in less conventional ways—that is, to transduce information from one modality such that it can be sensed by other sensory systems.

Sensory transduction may serve to substitute a failing sensory system. Perhaps the most successful application of sensory substitution is Braille. Information usually acquired with our visual sensory system (reading) is transduced such that it can be acquired through the tactile sensory system (fingertips). Bach-y-Rita and Kercel (2003) suggested that reading itself can be considered the first sensory substitution system because it transduces auditory information (spoken words) such that it can be read by the visual system. With sufficient signal processing power and miniaturization of feedback devices, it has also become possible to create real-time sensory transducers. One example is the "seeing with sound" system, "vOICe," with which blind people can sense a visual scene (e.g., a street to cross). This sensory substitution system transduces images from a head-mounted camera into sound patterns that carry directional and distance information. An earlier example is the Tactile Vision Substitution System, developed by Bach-y-Rita, Collins, Saunders, White, and Scadden (1969), which transduces

visual patterns from a head-mounted camera to vibrotactile patterns on the torso, enabling blind people to "see with their skin." Sensory substitution can also be applied within a sensory system, such as the finger-to-forehead of a person who has lost peripheral sensation (Bach-y-Rita, 1995).

Second, sensory transduction can be applied not as a substitution but as an augmentation of our senses or to enhance a single communication channel, such as speech or writing, into information that sends stimuli to several human senses. Waterworth (1997) termed such applications *synesthetic media.* For example, Smoliar, Waterworth, and Kellock (1995) developed a system to transduce the auditory properties such as dynamics, tempo, articulation, and synchronization of piano play to suitable visual representations to facilitate the communication between the student and the piano teacher. In another domain, TNO developed a tactile suit that transduces directional and gravitational information to vibrotactile patterns on the torso. This suit (see Figure 13.2)

Figure 13.2. The vibrotactile vest (TNO).

has been successfully tested for supporting pilots in landing their helicopter in Afghanistan during "brownouts," when clouds of dust and sand make landings based on visual information nearly impossible (van Erp, 2007). Similarly, the tactile vest has proven to support the spatial orientation of astronauts effectively under microgravity conditions in the International Space Station (van Erp & van Veen, 2006). However, it can also be used to complement the visual system, for example, in gaming applications.

Benefits of Multimodal Human–Computer Interaction

Waterworth (1997) stated that human–computer interaction (HCI) design should be seen as the art and science of sensory ergonomics for developing appropriate artifacts for sensory enhancement and communication. He assumed computers to serve as sensory transducers rather than cognitive artifacts. Today's technology makes it possible to design advanced sensory transducing multimodal human–computer interfaces, which may have various potential benefits, if designed carefully.

First, multimodal interfaces yield more robust performance. They present information in consistent complementary or redundant forms (or both). For example, the visual shape and the sound of a bird are complementary information, allowing disambiguation and enhancing the human detection and recognition performance of objects. Visual and tactile information about the size of the bird can be redundant, increasing the robustness of perceptual performance in case some sensory systems fail. Multimodal HCI can greatly improve the performance stability and robustness of recognition-based systems through disambiguation (Oviatt & Cohen, 2000).

Second, multimodal interfaces can reduce mental load. Current HCIs are strongly unimodal (usually visual), mainly consuming resources of a single sensory system and possibly causing mental overload. It has been shown, for example, that the tactile sensory systems can take over tasks of the visual system by adequately transducing the visual information to tactile patterns (van Erp, 2007).

Third, multimodal HCI has the potential to greatly expand the accessibility of virtual worlds to a larger diversity of users by adequately selecting the most appropriate combinations of modalities with respect to age, skill, style, impairments, and language (Oviatt & Cohen, 2000). For example, vision allows for a prolonged (foveal) attention to complex visual scenes and can attract attention peripherally, whereas hearing is transient but is omni-directional and has a longer short-term memory storage (Wickens, 1992). In contrast to vision and hearing, touch is capable of simultaneously sensing and acting when exploring the environment.

Fourth, multimodal presentation can promote new forms of HCI that were not previously available. For example, interfaces may adapt information presentation to the most appropriate sensory system given the preferences, needs, tasks, and context of the user, including aspects such as privacy, environmental noise and lighting, weather conditions, and protecting cloth. Adaptation of multimodal information presentation can be system controlled (the system

finds out itself based on user profiles and user behavior monitoring) versus user controlled (explicit user-articulated preferences).

However, one also has to be aware of the trade-offs in multimodal HCI (Sarter, 2006). One of them is the trade-off between the benefits of adaptive multimodal HCI on the one hand and the increased cost of interface management and monitoring user demands on the other. Another is that the expected increase of robustness and decrease of mental load through the use of multimodal interfaces may lead to a higher risk that our brain can no longer converge modalities into a coherent percept and that sensory systems start to interfere, if such interfaces are not designed carefully.

Obviously, the successful design of multimodal HCI and virtual worlds relies heavily on a thorough knowledge of the underlying mechanisms of multimodal perception.

What Is Known About Touch-Inclusive Multimodal Perception

Research into the characteristics of unimodal perception has a long history; is spread across multiple disciplines, including the neurosciences, psychology, philosophy, physics, and computer sciences; and has been written down and reviewed in a huge body of literature. The first author's early contributions to the literature of human perception were also focused on unimodal visual motion perception, for example, motion detection mechanisms underlying the local extraction of linear motion patterns (Werkhoven & Koenderink, 1990) and also on rotary motion perception (Werkhoven & Koenderink, 1991), second-order motion perception (Werkhoven, Chubb, & Sperling, 1993; Werkhoven, Sperling, & Chubb, 1994), and structure from motion perception (De Vries & Werkhoven, 1995; Werkhoven & van Veen, 1995). The fact that these topics are just tiny building blocks of a single sensory system makes one realize that the world of multimodal perception will continue to challenge us with research questions for many decades to come.

Interestingly, although human perception generally has a multimodal nature, researchers only recently started to study the underlying mechanisms of multisensory perception and integration (Rock & Victor, 1964; Welch & Warren, 1980). Stein and Meredith (1993) and later Calvert, Spence, and Stein (2004) wrote excellent overviews based on the strongly fragmented literature on multisensory processing and have highlighted the most notable advances in this field. It must be noted, however, that the literature on unimodal as well as multimodal perception is strongly centered on the visual and auditory sensory systems. The synthesis of these sensory systems with the tactile sensory system has received minimal attention from the research community, given the great relevance and potential of touch in interactive virtual worlds.

In this chapter, we discuss what we call *touch-inclusive multimodal perception*: To what extent do tactile and other sensory modalities differ in their

spatiotemporal characteristics? To what extent can touch-inclusive multimodal presentations enhance detection and recognition? To what extent do incongruent tactile and other sensory modalities interfere with each other?

To What Extent Do Tactile and Other Sensory Modalities Differ in Their Spatiotemporal Characteristics?

Given the increasing complexity of visual interfaces, system designers are increasingly looking toward the auditory and tactile sensory systems to provide alternative or supplementary means of information transfer (Spence & Driver, 1997; van Erp, 2001). Effective multimodal interfaces require that stimulation from several sensory channels be coordinated and made congruent informationally as well as temporally (Kolers & Brewster, 1985). Given the context of touch-inclusive multimodal interfaces and that time is an important and common dimension of all sensory modalities, we here focus on the temporal characteristics of the tactile and the visual system and aim at identifying cross-modal sensitivities and biases. For example, are time intervals when estimated by the tactile system the same as when estimated by the visual system?

Cross-Modal Biases

The relationships among the auditory, visual, and tactile channels regarding temporal duration has not been studied extensively. Only the perceived durations of visual and auditory time intervals have been compared. For temporal intervals on the order of 1 s, visual intervals had to be set longer than auditory intervals to be judged as equal in duration (Goldstone, Boardman, & Lhamon, 1959).

Modality differences have also been reported for other time-related measures and tasks, for example, in duration discrimination (Lhamon & Goldstone, 1974), temporal order judgment (Kanabus, Szelag, Rojek, & Poppel, 2002), stimulus sequence identification (Garner & Gottwald, 1968), perception of temporal rhythms (Gault & Goodfellow, 1938), and a temporal-tracking and continuation-tapping task (Kolers & Brewster, 1985). We know of only two studies that have addressed auditory–tactile interval duration comparisons. Both Ehrensing and Lhamon (1966) and Hawkes, Deardorff, and Ray (1977) found perceived tactile durations to equal auditory ones.

On the basis of the biased auditory visual relation and the unbiased auditory tactile relation, one might expect a bias in tactile visual comparisons as well: Visual intervals will probably have to be longer than tactile intervals to be judged as equal in duration. However, because manual interactions with the environment are often controlled using visual feedback, it may be expected that the perception of time intervals for the eye and for the fingertips has evolved to be consistent. Evidence for this comes from the development of visual–haptic interactions in children (Birch & Lefford, 1963, 1967).

van Erp and Werkhoven (2004) investigated the expected cross-modal bias for tactually and visually defined empty time intervals. In a forced-choice discrimination task, participants judged whether the second of two intervals was shorter or longer than the first interval. Two pulses defined the intervals. The pulse was either a vibrotactile burst presented to the fingertip or a foveally presented flash of a white square. The comparisons were made for unimodal (visual–visual or tactile–tactile) and cross-modal intervals (visual–tactile and tactile–visual) in the range of 100 to 800 ms. The standardized bias was defined as the time interval of one modality (say, visual) that was subjectively equal to the standard interval of another (say, tactile), normalized with respect to that standard interval and, thus, expressed in percentages. The results showed significant cross-modal biases between the tactile and sensory systems. Tactile empty intervals had to be set 8.5% shorter on average to be perceived as long as visual intervals. The cross-modal bias was largest for small intervals (15% for the 100-ms intervals).

Cross-Modal Sensitivity

Unimodal threshold studies have shown that the temporal resolution of the skin lies between those of hearing and vision (Kirman, 1973). This relation goes for numerous time-related measures and tasks, including discrimination of duration (Goodfellow, 1934), synchronization of finger taps (Kolers & Brewster, 1985), and adjusting empty intervals to equal pulse duration (Craig, 1973).

To be able to compare visual and tactile information in a cross-modal setting, there must be a common representation of the information from both senses. Several mechanisms for cross-modal visual–haptic comparisons have been suggested, based on two fundamentally different models (Summers & Lederman, 1990). The first is based on modality-specific representations that are used for unimodal comparisons (Lederman, Klatzky, Chataway, & Summers, 1990). These modality-specific representations must be translated into a common representation for cross-modal comparisons. This implies that cross-modal comparisons require an extra translation compared with unimodal comparisons. On the basis of the assumption that this extra translation increases the variability in the judgments, this model predicts a lower sensitivity for cross-modal comparisons than for unimodal comparisons. The second model (Ernst & Banks, 2002) states that information from the different modalities is directly processed and translated into a common (amodal) representation. This representation is used for both unimodal and cross-modal comparisons. In the latter model, unimodal and cross-modal comparisons are based on the same representation and are, therefore, hypothesized to have the same sensitivity.

van Erp and Werkhoven (2004) investigated the human sensitivity to discriminate empty intervals as a function of interval length and compared cross-modal sensitivity with unimodal sensitivity. Variances were derived from the same unimodal and cross-modal interval comparison experiments mentioned earlier. The Weber fractions (the threshold divided by the standard interval) were 20% and were constant over the standard intervals. This indicates that the Weber law holds for the range of interval lengths tested (100–800 ms).

Furthermore, the Weber fractions are consistent over unimodal and cross-modal comparisons, which suggests that there are no additional costs involved in the cross-modal comparison.

To What Extent Can Touch-Inclusive Multimodal Presentations Enhance Stimulus Detection and Recognition?

Generally, humans perceive real-world scenes through multiple sensory systems, and, most often, the information from the different sensory modalities involved is congruent, in either a complementary or a redundant way. Important questions are how scene information is processed within each modality, how information is integrated across modalities, and if and how this increases detection and recognition performance of objects or events.

Benefits of Multimodal Perception

Studies on multisensory integration have demonstrated that human perception can significantly increase in quality when the same environmental property is perceived in more than one sensory modality. For example, multimodal redundant stimuli have been shown to improve reaction time (Hershenson, 1962), stimulus identification (Doyle & Snowden, 2001), contrast detection (Lippert, Logothetis, & Kayser, 2007), perceptual organization (Vroomen & de Gelder, 2000), temporal boundaries (Vroomen & de Gelder, 2004), spatial localization (Alais & Burr, 2004), height estimation (Ernst & Banks, 2002), the reliability of depth cues (Landy, Maloney, Johnston, & Young, 1995), and size and stiffness estimates (Wu, Basdogan, & Srinivasan, 1999). Extensive reviews on neural, perceptual, and behavioral aspects of sensory integration can be found by Stein and Meredith (1993) and Calvert et al. (2004). So for congruent stimuli (derived from the same source), multisensory interaction indeed seems to improve the quality of perception. In fact, multisensory integration allows the brain to arrive at a statistically optimized integrated perceptual estimate under conditions in which the stimuli from the individual modalities involved are congruent, although each may be noisy, incomplete, and perhaps slightly different.

Interactions Between Sensory Systems

For incongruent stimuli, multisensory integration would obviously be ineffective. However, the brain cannot always determine correctly if individual signals are congruent or not. In some cases, our brain values a holistic percept so highly that incongruent stimuli lead to illusionary percept. For example, Shams, Kamitani, and Shimojo (2000, 2002) discovered that we perceive an illusory second flash when a single flash of light is accompanied by multiple auditory beeps, leading to a decrease in numerosity judgment performance. Andersen, Tiippana, and Sams (2005) extended this work by showing that the number of perceived flashes can be both increased (called *fission*) and decreased (called *fusion*) by presenting a larger or smaller number of irrelevant beeps in combination with the flashes.

Besides these audiovisual illusions, comparable effects have been reported for almost all other combinations of modalities (Bresciani, Dammeier, & Ernst, 2006; Courtney, Motes, & Hubbard, 2007; Ernst & Bulthoff, 2004; Hötting & Röder, 2004; Violentyev, Shimojo, & Shams, 2005).

Findings for incongruent stimuli have shed some light on the sensory integration process. The illusory flash effect has been explained by Bayesian models (Andersen et al., 2005; Bresciani et al., 2006; Ernst & Bulthoff, 2004; Shams, Ma, & Beierholm, 2005). These models propose that the more reliable estimate (with the smallest standard deviation) has a larger influence on the final percept. Auditory estimates are generally more reliable than visual estimates in temporal tasks, giving them dominance over visual perception in the illusory flash experiment (Andersen et al., 2005). Similarly, the tactile modality is more reliable than the visual modality and can induce visual flash illusions (Violentyev et al., 2005). In turn, the tactile modality can be modulated by auditory stimuli (Bresciani et al., 2005). Interestingly, this suggests an order of dominance (influence) for numerosity estimates equal to the order of performance found by Lechelt (1975) for unimodal temporal numerosity judgment: Hearing is best followed by touch and vision. That is, the more reliable modality in the illusory flash paradigm corresponds with the more accurate modality in his temporal numerosity judgment task.

Multimodal Numerosity Estimation

Given the growing body of models and experimental data on numerosity judgment tasks with incongruent stimuli, it may come as a surprise that multisensory numerosity judgments of congruent stimuli have hardly been studied. Only recently, Gallace, Tan, and Spence (2007) tested a spatial numerosity judgment task in which multimodal pulses were presented simultaneously at multiple locations as opposed to sequentially at a single location, as in temporal numerosity judgment. Participants had to count and sum the pulses presented in the tactile and visual modality. They found that bimodal numerosity judgments were significantly less accurate than unimodal judgments and suggested that numerosity judgments rely on a unitary amodal system. It would be interesting to know if this extends to temporal numerosity judgment.

Philippi, van Erp, and Werkhoven (2008) investigated a temporal numerosity judgment task and tested whether multimodal presentations can reduce the numerosity underestimation biases observed in unimodal conditions. Participants were presented with two to 10 pulses at different interstimulus intervals (ISIs) under unimodal conditions (visual, auditory, and tactile senses) as well as multimodal combinations. The results showed that for short ISIs (between 20 and 80 ms), multimodal presentation significantly reduced the underestimation of numerosity compared with unimodal presentation and, thus, enhanced performance. Interestingly, however, we found no differences in the variance of numerosity estimation between unimodal and multimodal presentations, suggesting that the integration process did lead to performance enhancement, but not through the variance reduction predicted by current (Bayesian) integration models.

The Cost of Multimodal Integration

We have seen convincing examples of perceptual task improvement under multimodal presentation conditions, generally for low-level perceptual tasks. For such tasks, the benefits of multimodal presentation seem to outweigh the cost of integration in terms of processing multiple resources. For higher level perceptual tasks such as object recognition, multisensory integration may come at a higher cost, for example, when information from different modalities is derived from the same scene but with different scene orientations for different modalities. Newell, Woods, Mernagh, and Bülthoff (2005) investigated the visual, haptic, and cross-modal recognition of scenes of familiar objects. Participants first learned a scene of objects in one sensory modality and were then asked to detect positional switches between objects in the same or a different modality. Newell and colleagues found a cost in scene recognition performance when there was a change in sensory modality and scene orientation between learning and test and suggested that differences between visual and haptic representations of space may affect the recognition of scenes of objects across these modalities.

To What Extent Do Incongruent Tactile and Other Sensory Modalities Interfere With Each Other, and What Is the Role of Attention in Sensory Integration?

Sensory Integration Models

Various studies on human perception of incongruent sensory inputs have determined how one sensory system was biased by task-irrelevant stimulation of other sensory systems. Some studies have shown an asymmetry of bias effects—that is, the more "appropriate" system for a particular task seemed to dominate less appropriate sensory systems. Guest and Spence (2003), for example, investigated the visuotactile assessment of roughened textile samples, in the presence of a congruent or an incongruent textile distracter, and concluded that vision influenced touch more than touch-influenced vision. The results further suggested that modality appropriateness was a function of the discriminative ability of the modality as well as ecological validity. Obviously, the order of dominance observed for temporal tasks does not seem to extend directly to spatial tasks such as texture assessment.

Early qualitative perceptual integration models assume that the more "appropriate" sensory system (i.e., most sensitive for the specific stimulus) dominates the less appropriate system, in the most extreme form, the "winner takes all" model (Welch & Warren, 1980). Other studies (Ernst & Bulthoff, 2004) have found mutual influences that can best be explained by the assumption that sensory signals are integrated with weights proportional with relative signal-to-noise ratio (the reliability of the sensory channel). On the basis of this reliability assumption, various quantitative probabilistic "ideal observer" models have been developed to explicitly model multimodal perception.

The maximum likelihood integration (MLI) approach by Andersen et al. (2005) assumes complete integration of the sensory channels and, consequently, that the sum of their weights equals 1. They found that early MLI (integration before stimulus categorization) explained the perceptual integration of rapid beeps and flashes better than late MLI (integration after categorization). Shams et al. (2005) developed a Bayesian integration scheme that could account for situations of partial integration in sound-induced flash illusions as well as complete integration.

Bresciani et al. (2006) studied experimental conditions in which visual and tactile signals were only partially integrated. They interpreted partial integration as a coupling between sensory channels and quantified the integration process using a Bayesian integration scheme with a coupling prior (e.g., prior knowledge about the probability that two sensory channel inputs originate from the same source). The free model parameter "Coupling Strength" distinguishes the model of Bresciani et al. from those of Andersen et al. (2005) and Shams et al. (2005).

The Complicating Role of Attention

Previous evidence on multimodal integration is based on experimental paradigms in which the participants' task was to ignore the "irrelevant" channel. This was done to show that multimodal integration occurs even if you want to ignore it. However, in such conditions, the strength of the integration and the ability to ignore a channel are confounded. Therefore, we explicitly distinguish two effects: (bottom-up) sensory system integration effects occurring in situations where both modalities are attended, and (top-down) sensory system suppression effects by selectively attending to the task relevant modality and ignoring the irrelevant.

Isolating Sensory System Integration

To disentangle bottom-up integration and top-down attention effects, Werkhoven, van Erp, and Philippi (2009) carried out multimodal perception experiments in which participants were exposed to incongruent sequences of visual flashes and tactile taps in two conditions. In one condition (the cue condition), they were instructed before stimulus presentation to report the number of events in a particular modality and to ignore the other (i.e., the traditional paradigms, in which sensory system integration and suppression effects are combined). In a second condition (the no-cue condition), they were instructed on which modality to report only after stimulus presentation and therefore could not ignore a channel (i.e., isolating the effect of sensory system integration). By comparing the results, Werkhoven et al. (2009) could quantify to what extent sensory integration and selective attention influence whole or partial perceptual integration.

The effects measured were fission effects and fusion effects: The task-irrelevant modality can increase (fission) or decrease (fusion) the number of perceived events in the task-relevant modality. Results showed that no-cue conditions yielded overall stronger fission and fusion effects than cue conditions, indicating that previous studies were based on the combined effects of

sensory integration and selective attention. Furthermore, in no-cue conditions, the influence of vision on touch is stronger than the influence of touch on vision. However, in cue conditions, irrelevant flashes are more easily ignored than irrelevant taps. Together, these results suggest that the bottom-up influence of vision on touch is stronger but that vision is also more easily suppressed by top-down selective attention.

Consequences and Chances for Multimodal Simulation

So what has been learned so far? Because the experiments mentioned here can at most be considered to be tiny pieces of a giant puzzle of perceptual organization and because generalization of their results cannot be scientifically justified, we will just briefly speculate on some possible consequences for multimodal simulation.

Our sensory systems are all optimized for specific aspects of the outside world, often leading to different spatial and temporal characteristics. The crossmodal biases for time interval estimation between the tactile and the visual system, for example, are substantial. There is no right or wrong about an internal representation of a sensory system; there are only differences. To make time intervals congruent in simulated environments, we may have to adjust them a little bit relative to each other.

Furthermore, we have seen that multimodal estimation can improve many perceptual tasks, such as numerosity estimation. Improving numerosity estimation was not trivial back in the 1940s. At that time, Taubman (1950) reported that observers had difficulty adequately discriminating the short tones as part of characters in the international Morse code. This problem arose not only for the perception of auditory pulses, it also existed if the code consisted of flashes of light (i.e., blinker code). We may not need to optimize Morse code any longer, but the use of tactile displays (van Erp et al., 2006) to transduce information through simple spatial and temporal patterns on the skin may certainly benefit from this knowledge. The tactile vest has a wide range of current and potential applications, varying from guidance for people who are blind or have severe visual impairment, to spatial orientation in vehicles, to feedback during revalidation, to sports feedback, to teleoperations, to entirely new forms of multimodal touch-inclusive gaming.

Last but not least, we saw that incongruent stimuli can lead to illusory percepts (e.g., illusory flashes or taps) due to automatic sensory integration process. More specifically, the bottom-up interaction between sensory channels can be asymmetric, and perceptual attention can have a strong top-down influence. Such illusory percepts, when sufficiently understood, can be of interest to the community of multimodal simulator designers. The simulation of some perceptual aspects can be constrained by technical limitations or can come at high cost, similar to force feedback effects. In such cases, it may be interesting to find alternative stimulus combinations that create the same (illusory) percept. It may even allow for illusory percepts that cannot occur in the real world, such as the "rubber-hand illusion" (Ehrsson, Spence, & Passingham, 2004).

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal cross-modal integration. *Current Biology, 14,* 257–262.

Andersen, T. S., Tiippana, K., & Sams, M. (2005). Maximum likelihood integration of rapid flashes and beeps. *Neuroscience Letters, 380,* 155–160. doi:10.1016/j.neulet.2005.01.030

Bach-y-Rita, P. (1995). *Nonsynaptic diffusion neurotransmission and late brain reorganization.* New York, NY: Demos-Vermande.

Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969, March 8). Vision substitution by tactile image projection. *Nature, 221,* 963–964. doi:10.1038/221963a0

Bach-y-Rita, P., & Kercel, S. W. (2003). Sensory substitution and the human-machine interface. *Trends in Cognitive Sciences, 7,* 541–546. doi:10.1016/j.tics.2003.10.013

Birch, H. G., & Lefford, A. (1963). Intersensory development in children. *Monographs of the Society for Research in Child Development, 28*(5). doi:10.2307/1165681

Birch, H. G., & Lefford, A. (1967). Visual differentiation, intersensory integration, and voluntary motor control. *Monographs of the Society for Research in Child Development, 32*(2). doi:10.2307/1165792

Bresciani, J. P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision, 6,* 554–564. doi:10.1167/6.5.2

Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: Auditory signals can modulate tactile taps perception. *Experimental Brain Research, 162,* 172–180. doi:10.1007/s00221-004-2128-2

Calvert, C., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes.* Cambridge, MA: MIT Press.

Courtney, J. R., Motes, M. A., & Hubbard, T. L. (2007). Multi- and unisensory visual flash illusions. *Perception, 36,* 516–524. doi:10.1068/p5464

Craig, J. (1973). A constant error in the perception of brief temporal intervals. *Perception & Psychophysics, 13,* 99–104. doi:10.3758/BF03207241

De Vries, S. C., & Werkhoven, P. (1995). Cross-modal slant and curvature matching of stereo- and motion-specified surfaces. *Perception & Psychophysics, 57,* 1175–1186. doi:10.3758/BF03208373

Doyle, M. C., & Snowden, R. J. (2001). Identification of visual stimuli is improved by accompanying auditory stimuli: The role of eye movements and sound location. *Perception, 30,* 795–810. doi:10.1068/p3126

Ehrensing, R. H., & Lhamon, W. T. (1966). Comparison of tactile and auditory time judgments. *Perceptual and Motor Skills, 23,* 929–930. doi:10.2466/pms.1966.23.3.929

Ehrsson, H. H., Spence, C., & Passingham, R. E. (2004, August 6). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science, 305,* 875–877. doi:10.1126/science.1097011

Ernst, M. O., & Banks, M. S. (2002, January 24). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433. doi:10.1038/415429a

Ernst, M. O., & Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8,* 162–169. doi:10.1016/j.tics.2004.02.002

Gallace, A., Tan, H. Z., & Spence, C. (2007). Multisensory numerosity judgments for visual and tactile stimuli. *Perception & Psychophysics, 69,* 487–501. doi:10.3758/BF03193906

Garner, W. R., & Gottwald, R. L. (1968). The perception and learning of temporal patterns. *The Quarterly Journal of Experimental Psychology, 20,* 97–109. doi:10.1080/14640746808400137

Gault, R. H., & Goodfellow, L. D. (1938). An empirical comparison of audition, vision, and touch in the discrimination of temporal patterns and ability to reproduce them. *Journal of General Psychology, 18,* 41–47. doi:10.1080/00221309.1938.9709888

Goldstone, S., Boardman, W. K., & Lhamon, W. T. (1959). Intersensory comparisons of temporal judgments. *Journal of Experimental Psychology, 57,* 243–248. doi:10.1037/h0040745

Goodfellow, L. D. (1934). An empirical comparison of audition, vision, and touch in the discrimination of short intervals of time. *The American Journal of Psychology, 46,* 243–258. doi:10.2307/1416558

Guest, S., & Spence, C. (2003). Tactile dominance in speeded discrimination of textures. *Experimental Brain Research, 150,* 201–207.

Hawkes, G. R., Deardorff, P. A., & Ray, W. S. (1977). Response delay effects with cross-modality duration judgments. *Journal of Auditory Research, 17,* 55–57.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology, 63,* 289–293. doi:10.1037/h0039516

Hötting, K., & Röder, B. (2004). Hearing cheats touch, but less in congenitally blind than in sighted individuals. *Psychological Science, 15,* 60–64. doi:10.1111/j.0963-7214.2004.01501010.x

Kanabus, M., Szelag, E., Rojek, E., & Poppel, E. (2002). Temporal order judgement for auditory and visual stimuli. *Acta Neurobiologiae Experimentalis, 62,* 263–270.

Kirman, J. H. (1973). Tactile communication of speech: A review and analysis. *Psychological Bulletin, 80,* 54–74. doi:10.1037/h0034630

Kolers, P. A., & Brewster, J. M. (1985). Rhythms and responses. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 150–167. doi:10.1037/0096-1523.11.2.150

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35,* 389–412. doi:10.1016/0042-6989(94)00176-M

Lechelt, E. C. (1975). Temporal numerosity discrimination: Intermodal comparisons revisited. *British Journal of Psychology, 66,* 101–108. doi:10.1111/j.2044-8295.1975.tb01444.x

Lederman, S. J., Klatzky, R. L., Chataway, C., & Summers, C. G. (1990). Visual mediation and the haptic recognition of two-dimensional pictures of common objects. *Perception & Psychophysics, 47,* 54–64. doi:10.3758/BF03208164

Lhamon, W. T., & Goldstone, S. (1974). Studies of auditory-visual differences in human time judgment: 2. More transmitted information with sounds than lights. *Perceptual and Motor Skills, 39,* 295–307. doi:10.2466/pms.1974.39.1.295

Lippert, M., Logothetis, N. K., & Kayser, C. (2007). Improvement of visual contrast detection by a simultaneous sound. *Brain Research, 1173,* 102–109. doi:10.1016/j.brainres.2007.07.050

Mesland, B. S., Bles, W., Werkhoven, P., & Wertheim, A. H. (1998). How flexible is the self-motion system? Introducing phase and amplitude differences between visual and proprioceptive passive linear horizontal self-motion stimuli. In B. S. Mesland (Ed.), *About horizontal self-motion perception* (pp. 99–142). Utrecht, The Netherlands: University of Utrecht.

Newell, F. N., Woods, A. T., Mernagh, M., & Bülthoff, H. H. (2005). Visual, haptic and cross-modal recognition of scenes. *Experimental Brain Research, 161,* 233–242. doi:10.1007/s00221-004-2067-y

Oviatt, S. L., & Cohen, P. R. (2000). Multimodal systems that process what comes naturally. *Communications of the ACM, 43,* 45–53. doi:10.1145/330534.330538

Philippi, T. G., van Erp, J. B. F., & Werkhoven, P. J. (2008). Multisensory temporal numerosity judgment. *Brain Research, 1242,* 116–125. doi:10.1016/j.brainres.2008.05.056

Rock, I., & Victor, J. (1964, February 7). Vision and touch: An experimentally created conflict between the two senses. *Science, 143,* 594–596. doi:10.1126/science.143.3606.594

Sarter, N. B. (2006). Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial Ergonomics, 36,* 439–445. doi:10.1016/j.ergon.2006.01.007

Shams, L., Kamitani, Y., & Shimojo, S. (2000, December 14). What you see is what you hear. *Nature, 408,* 788. doi:10.1038/35048669

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Brain Research. Cognitive Brain Research, 14,* 147–152. doi:10.1016/S0926-6410(02)00069-1

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport, 16*(17), 1923–1927. doi:10.1097/01.wnr.0000187634.68504.bb

Smoliar, S. W., Waterworth, J. A., & Kellock, P. R. (1995). PianoFORTE: A system for piano education beyond notation literacy. *Proceedings of the Third ACM International Conference on Multimedia,* 457–465.

Spence, C., & Driver, J. (1997). Cross-modal links in attention between audition, vision, and touch: Implications for interface design. *International Journal of Cognitive Ergonomics, 1,* 351–373.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* Cambridge, MA: MIT Press.

Summers, D. C., & Lederman, S. J. (1990). Perceptual asymmetries in the somatosensory system: A dichhaptic experiment and critical review of the literature from 1929 to 1986. *Cortex, 26,* 201–226.

Taubman, R. E. (1950). Studies in judged number: I. The judgment of auditory number. *The Journal of General Psychology, 43,* 195–219. doi:10.1080/00221309.1950.9710620

van Erp, J. B. F. (2001). Tactile navigation display. In S. Brewster & R. Murray-Smith (Eds.), *Haptic human-computer interaction* (pp. 165–173). Berlin, Germany: Springer-Verlag. doi:10.1007/3-540-44589-7_18

van Erp, J. B. F. (2007). *Tactile displays for navigation and orientation: Perception and behaviour.* Leiden, The Netherlands: Mostert & Van Onderen.

van Erp, J. B. F., & van Veen, H. A. H. C. (2006). Touch down: The effect of artificial touch cues on orientation in microgravity. *Neuroscience Letters, 404,* 78–82. doi:10.1016/j.neulet.2006.05.060

van Erp, J. B. F., & Werkhoven, P. J. (2004). Perception of vibro-tactile asynchronies. *Perception, 33,* 103–111.

Violentyev, A., Shimojo, S., & Shams, L. (2005). Touch-induced visual illusion. *Neuroreport, 16,* 1107–1110. doi:10.1097/00001756-200507130-00015

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1583–1590. doi:10.1037/0096-1523.26.5.1583

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 513–518. doi:10.1037/0096-1523.30.3.513

Waterworth, J. A. (1997). Creativity and sensation: The case for synaesthetic media. *Leonardo, 30,* 327–330. doi:10.2307/1576481

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667. doi:10.1037/0033-2909.88.3.638

Werkhoven, P., & Koenderink, J. J. (1990). Extraction of motion parallax structure in the visual system I. *Biological Cybernetics, 63,* 185–191. doi:10.1007/BF00195857

Werkhoven, P., & Koenderink, J. J. (1991). Visual processing of rotary motion. *Perception & Psychophysics, 49,* 73–82. doi:10.3758/BF03211618

Werkhoven, P., Chubb, C., & Sperling, G. (1993). The dimensionality of texture-defined motion: A single channel theory. *Vision Research, 33,* 463–485. doi:10.1016/0042-6989(93)90253-S

Werkhoven, P., Sperling, G., & Chubb, C. (1994). Motion perception between dissimilar gratings: Spatiotemporal properties. *Vision Research, 34,* 2741–2759. doi:10.1016/0042-6989(94)90230-5

Werkhoven, P., van Erp, J. B. F., & Philippi, T. (2009). Counting visual and tactile events: The effect of attention on multisensory integration. *Attention, Perception, & Psychophysics, 71,* 1854–1861. doi:10.3758/APP.71.8.1854

Werkhoven, P., & van Veen, H. A. H. C. (1995). Extraction of relief from visual motion. *Perception & Psychophysics, 57,* 645–656. doi:10.3758/BF03213270

Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd ed.). New York, NY: HarperCollins.

Wu, W.-C., Basdogan, C., & Srinivasan, M. A. (1999). Visual, haptic, and bimodal perception of size and stiffness. In virtual environments. *ASME Dynamic Systems and Control Division, 67,* 19–26.