Other ways of seeing: From behavior to neural mechanisms in the online “visual” control of action with sensory substitution

Michael J. Proulx⁎, James Gwinnut, Sara Dell’Erba, Shelly Levy-Tzedeck, Alexandra A. de Sousa and David J. Brown

Abstract. Vision is the dominant sense for perception-for-action in humans and other higher primates. Advances in sight restoration now utilize the other intact senses to provide information that is normally sensed visually through sensory substitution to replace missing visual information. Sensory substitution devices translate visual information from a sensor, such as a camera or ultrasound device, into a format that the auditory or tactile systems can detect and process, so the visually impaired can see through hearing or touch. Online control of action is essential for many daily tasks such as pointing, grasping and navigating, and adapting to a sensory substitution device successfully requires extensive learning. Here we review the research on sensory substitution for vision restoration in the context of providing the means of online control for action in the blind or blindfolded. It appears that the use of sensory substitution devices utilizes the neural visual system; this suggests the hypothesis that sensory substitution draws on the same underlying mechanisms as unimpaired visual control of action. Here we review the current state of the art for sensory substitution approaches to object recognition, localization, and navigation, and the potential these approaches have for revealing a metamodal behavioral and neural basis for the online control of action.

Keywords: Sensory substitution, blindness, prosthetics, object recognition, perception for action, active sensing

Our senses allow us to interact with the world. Vision in particular is adapted to providing information about distant, silent objects by decoding the reflected or emitted light to perceive what objects are and where they are (Marr, 1982). The distinction between recognition and localization maps led on to the discovery of two processing streams in the visual brain for “where” information in the dorsal stream and “what” information in the ventral stream (Mishkin & Ungerleider, 1982). It has been suggested that these categories should be modified: the “where” category includes perception for action in addition to mere spatial perception, and would more appropriately be termed “how” rather than just “where” (Milner & Goodale, 2008). It is worth noting that, although there is this appar-
ent dissociation of processing into two streams in the brain, there certainly must be many interactions between the two: one must know what something is to know where it is and how to act upon it. This is apparent in any experience where one simultaneously processes what and where something is while acting upon such information, but also in an extensive line of recent work that has shown that action can impact perception, and vice-versa (Biegastraaten, de Grave, Brenner, & Smeets, 2007; Cressman & Hentiques, 2011). The adaptive nature of action, such as hand motion, in response to perceptual information is termed the perceptual online control of action (Gritsenko, Yakovenko, & Kalaska, 2009). Perception for action research provides a dynamic perspective given that one must continually map and re-map the surroundings in a reference frame to accommodate for self-movement and the movement of other objects (Pasqualotto & Proulx, 2012; Pasqualotto, Spiller, Jansari, & Proulx, 2013b).

For example, past research has shown that when hand motion is directed towards a visual target, an initial set of instructions is sent to the muscles controlling both hand and eye motion, and that this information is updated from retinal and extra-retinal signals as the motion occurs to fine tune the trajectory of the hand. This allows one to account for any changes in location of the target (Goodale, Pelisson, & Prablanc, 1986; Sarlegna et al., 2004) and to correct small performance errors (Gritsenko et al., 2009). This fine tuning of motion is termed online control of action and is suggested to be automatic rather than voluntary (Gritsenko et al., 2009). Transcranial magnetic stimulation (TMS) can disrupt individuals’ online corrective behavior when placed over the primary parietal cortex, suggesting it plays an important role in the online updating of motor actions (Gomi, 2008).

The online control of action can occur in the other senses, for example, humans can use auditory feedback and even echolocation to direct movements toward a target (Boyer et al., 2013; Thaler, 2013; Thaler, Arnott, & Goodale, 2011; Thaler & Goodale, 2011; Thaler, Milne, Arnott, Kish, & Goodale, 2014). Recent work to develop prosthetics for sight restoration aims to provide “visual” online control through methods of sensory substitution for vision using the other senses. Sensory substitution devices aim to provide various features and dimensions of visual information for the visually impaired by translating visual information into tactile or auditory form (Bach-y-Rita & Kercel, 2003; Meijer, 1992; Proulx, 2010).

Here we review the research on sensory substitution for vision restoration in the context of providing the means of online control for action in the blind or blindfolded. First we will describe sensory substitution in more detail, and then we will examine how such devices are used for object recognition, localization, and action, including pointing, grasping, and navigation behaviors. We summarize with a review of the neural basis for perception through such devices and note the excellent potential to reveal new, underlying neural mechanisms for the online control of action (Levy-Tzck et al., 2012b).

I. Substituting vision with auditory or tactile perception

Vision has special properties that are challenging to convey to the other senses in that the bandwidth of vision, and the capacity for parallel processing, exceeds that of the other senses (Pasqualotto & Proulx, 2012). Parallel processing is crucial for multisensory integration and thus multisensory perception and learning. This is disrupted in congenital visual impairment because vision provides the best spatial acuity and processing capacity to integrate information from the other senses (Proulx, Brown, Pasqualotto, & Meijer, 2014). The information capacity of the human eye has been estimated to be around 4.3 × 10^6 bits per second (Jacobson, 1951). This is four orders of magnitude greater bandwidth than estimates of the human fingertip of 100 bits per second (Kokjer, 1987); and other areas of skin estimated even lower, from 5 to 56 bits per second (Schmidt, 1981). The information capacity of the human ear is between these two estimates, and the highest after vision, with a capacity at 104 bits per second (Jacobson, 1950). For this reason, although the original tactile-visual sensory substitution systems used the skin of the back or stomach as an analogue of the retina (Bach-y-Rita, Collins, Saunders, White, & Scadden, 1969), the modern version uses more sensitive areas such as the forehead or tongue (Bach-y-Rita & Kercel, 2003). Even so, compared to central vision it is low-resolution.

The first general-purpose auditory sensory substitution system was invented by Meijer (1992) to provide a higher resolution substitution mechanism. Although the skin/retina analogue is obvious in some ways due to

...
the topographic representation of spatial locations on each, the auditory system, compared to the skin, provides a higher informational capacity to convey visual information to the brain. Meijer's invention, "The vOICe" (the middle letters signify, "Oh, I see!") converts images captured by a camera into sound through headphones. Each image is translated into sound with a left-to-right scan every one or two seconds with timing and stereo cues representing the horizontal axis, frequency representing the image's vertical axis and loudness representing pixel brightness. The input can take many forms, however, is always two-dimensional in nature, with one camera providing sufficient visual information for this. The software can run on Android phones, and can use the phone's camera, or with a web cam input for a PC. Spy sunglasses, with a webcam on the bridge of the nose, allow the visual input to be near the impaired or blindfolded eyes as a form of monocular vision. There are other auditory approaches to substitution. The Prosthesis for Substitution of Vision by Audition, for example, uses frequency to code for both axes by increasing from bottom to top and from left to right of the converted image, with increased representation of the center of the image as occurs with the fovea (Capelle, Trullemans, Arno, & Veraart, 1998).

Both of these devices, with extensive training, result in experiences that seem qualitatively similar to vision in some ways. Participants report distal attribution with tactile and auditory substitution for vision, rather than just merely attributing perceptual experience to the stimulated modalities (Auvray et al., 2005). Fascinatingly, a long-term user of The vOICe has acquired blindness later in life, and reports having visual imagery evoked by the device that is reminiscent of vision, including depth perception and smooth movement (Ward & Meijer, 2010). These findings of distal attribution and visual phenomenology converge with theoretical work concerning whether the experience of sensory substitution is more like the sensory modalities which have been substituted or those which are doing the substituting (Hurley & Noe, 2003), with some concluding that the experience is an extension of the sensory capacities of the individual to a novel sensory modality domain (Auvray & Myn, 2009). There is a unifying role for sensorimotor action for learning to use sensory substitution, and for describing the experience of it (O'Regan & Noe, 2001). Although visual information might be transformed into sound or touch, interacting with the world in an active, visual-like way will confer something other than just the substituting sensory modality, and possibly even something resembling the substituted modality.

2. Pointing and grasping

Vision is key for being able to locate silent, distal objects and also the optimal sense to allow sensorimotor coordination for directing another person via pointing or reaching out and grasping accurately objects within arm’s reach. A few studies on blind or blindfolded subjects have examined object localization, with two conducted to assess identification as well (Auvray, Hametton, & O'Regan, 2007; Brown, Macpherson, & Ward, 2011), and one that focused specifically on perception for action (Proulx, Stoerig, Ludowig, & Knoll, 2008). Over a three week period, participants assigned to use The vOICe practiced with it daily using objects in their own homes. Proulx et al. (2008), in a study on blindfolded subjects, assessed localization with a manual search task that required pointing and responding by pressing LED lit buttons. They used a perimeter device constructed with an LED array. An LED would light up and a tone would be emitted by the controlling computer until the light had been pressed. The participants were able to generalize the self-initiated practice at home to the lab test with significant improvements in speed and accuracy. A second experiment examined the localization and grasping of natural objects placed on a large table. Although again, those trained to use The vOICe had significant improvement in locating the objects, the more impressive finding was that the participants had also learned to reach with grasp-appropriate hand configurations (Dijkerman, Milner, & Carey, 1999). Given the challenge of interpreting the sound, created by a two-dimensional, greyscale image, and determining object borders, distance, and size, this finding was surprising. Such performance implied that participants learned to extract information about size, distance, shape, and orientation, and were able to use that information to adjust grasping movements appropriately to pick up objects of different sizes, shapes, and materials.

Research investigating the relevance of device input location has revealed how action for perception might work in non-visual ways. In the work by Proulx et al. (2008) the camera that served as the visual input capturing device for The vOICe was located
near the eyes in the form of a small camera in spy sunglasses as a form of monocular vision. However the camera need not be located near the eyes, nor kept relatively static in this way. Brown, Macpherson, and Ward (2011) found that different tasks benefitted from different camera locations. Although Proulx et al. (2008) used a head-mounted camera to mimic eye-sight, Avrav et al. (2007) used a handheld webcam for their experiment. Brown et al. (2011) compared experiments that required either object localization or identification with each camera position. They found an interaction of task by camera location: localization performance was superior with the glasses-mounted camera, identification performance was superior with the handheld camera. This finding implies that a sensorimotor account of learning and perception (O'Regan & Noe, 2001) applies to sensory substitution devices for the processing of “where/how”, that is perception-for-action, but not necessarily for perception for object recognition (“what”). Even if using a different sensor modality such as hearing, one can mimic standard perceptual-motor contingencies that are used in normal localization as revealed by having the image input near the eyes. It is important to note that the participants in the Brown et al. (2011) study were blindfolded, but sighted individuals. It would be interesting to see whether this result extends to not only the late, adventitiously blind who have had visual perceptual-motor experience, but also to the congenitally blind who have not had such experience. In this way the role of visual experience in development, versus innate perceptual-motor mechanisms, could be assessed similar to other work on the role of visual experience for cognition (Pasqualotto, Lam, & Proulx, 2012, 2013, Pasqualotto et al., 2013b; Proulx, Brown, Pasqualotto, & Meijer, 2012; Proulx et al., 2014).

These initial studies examining reaching and grasping movements guided by The vOICe sensory substitution device were ecologically valid in many ways, but did not have the ability to measure fine-grained differences in timing and accuracy. A device that has the same basic structure as The vOICe, with the addition of synthesized musical instruments as a feature to represent color, is called the EyeMusic (Levy-Tzedek, Hanassy, Abboud, Maidenbaum, & Amadi, 2012a). These researchers examined the ability of blindfolded sighted individuals to perform fast and accurate reaching movements to targets presented by the EyeMusic, and compared them to visually guided movements. Participants were required to reach rapidly to a target. Surprisingly, participants were able to perform the task nearly as accurately with the device as with seeing after just a short period of training. One limitation, however, is that the participants were given a potentially unlimited amount of time to listen to the object location before the fast reaching movement was required. This delay before initiating the movement illustrates the effortful, conscious nature of using sensory substitution without extensive training, particularly in comparison to the ability of the visual system to process information rapidly and in parallel.

The crossmodal transfer of information from sensory substitution to vision (more as a version of sensory augmentation) was examined in another study by Levy-Tzedek et al. (2012b) that also used The vOICe derivative, the EyeMusic. This study used a version of an online control of action visuomotor rotation task (Krakauer, Pine, Ghilardi, & Ghizer, 2000; Novick & Vaadia, 2011). Participants used a joystick to control a cursor on a computer screen. Their goal was to move the cursor from a starting position to a target presented on the screen. The location of the target was alternately portrayed using vision or SSD. After several trials, the relationship between their hand movement and the location of the cursor was altered, such that the cursor movement was diverted by 30 degrees compared to the actual hand movement. In this condition, in order to have the cursor reach the target, they had to perform a skewed movement, 30 degrees away from the actual target location (for example, in order to reach a target that is 30 degrees away from the vertical, they would have to move their hand along the vertical). Importantly, the feedback suggesting that the motor command needs to be altered was provided only via the visual modality. No rotation cues were given via sensory substitution. They found that participants were quick to alter their hand movements appropriately, and that the adaptation to this rotation transferred rapidly from the visual trials to the SSD trials. After-effects (that is, movement of the hand and the cursor in the opposite once rotation was removed) were found independent of the sensory modality (vision or sensory substitution via audition) that presented the targets. This implies that the underlying neural representation of spatial information used for online action might be independent of the sensory modality with which the target is acquired. Interestingly, they found transfer between vision and sensory substitut-
tion with very little training, and with no conscious awareness of the presence of rotation, as reported by the participants.

The PSVA, prosthesis for substitution of vision by audition (Capelle, Truillemans, Arno, & Veraart, 1998), has also been used in localization studies. Recall that while The vOICe scans an image from left to right to create the sound, the PSVA provides a simultaneous coding of the entire image with frequency and as a result requires intentional movement to make sense of the image similar to “active sensing” such as using eye movements to perceive an image. The studies described thus far with The vOICe required participants to compute depth to accurately locate and grasp the objects. A study with the PSVA investigated the problem of depth perception from monocular cues. Renier et al. (2005a) had participants locate items in depth using natural corridor cues in a simple, simulated setting. The participants were able to report accurately the depth relations between the objects in the display, though it would be of great interest to extend this work to employ reaching and grasping as well. The authors reported that all early blind participants knew that relative size was a depth cue having heard how sighted people described daily observations. This knowledge is crucial for the task because the acquisition of object identity through haptics is size invariant because any object that can be touched is perceived as its three-dimensional size. This knowledge has also been reported in studies of visual-to-tactile sensory substitution (Segond, Weiss, Kauvala, & Signapaito, 2013). Early or congenitally blind participants still have more difficulty using this knowledge than those who had visual experience previously. Learning to “see” through the other senses might require not only active experience (Proulx, 2010), but perhaps visual experience as well for full familiarization or at least for the efficient acquisition of the few perceptual motor contingencies (Pasqualotto & Proulx, 2012). In fact, the initial studies with the Tactile Vision Sensory Substitution system found that participants needed to manipulate the camera themselves in order to experience distal attribution of the scene, and that the same did not occur if the camera was manipulated by the experimenter. Therefore they needed to experience perceptual-motor contingencies via active sensing themselves through an interaction with the object and this form of tactile sensation that represented the image (Bach-y-Rita, Collins, Saunders, White, & Scadden, 1969).

3. Spatial navigation

Among the other senses, the auditory system has superior temporal processing (David J. Brown & Proulx, 2013), and as a result most studies have focused on visual-to-auditory sensory substitution devices for representing object information with high acuity and fidelity in both the spatial and temporal domains (Haigh, Brown, Meiger, & Proulx, 2013). One study also examined the use of The vOICe for spatial navigation despite the fact that image acquisition with this device occurs infrequently. Navigation clearly requires online perception and correction to avoid obstructions, yet The vOICe samples the environment with an image every 1-2 s, such snapshots of the visual world might not be enough to allow for truly online corrections to movements, and this would be exacerbated with moving obstructions. Brown et al. (2011) had participants using The vOICe walk a short route that contained four obstacles. Participants required over five minutes to complete the task; they demonstrated improvement of over one minute with eight trials of practice.

Does navigation require the higher resolution afforded by auditory devices rather than tactile ones? Visual navigation certainly utilizes lower-resolution peripheral vision and perhaps a simpler representation better suits an online task such as full body navigation through a space. This is a domain where visual-to-tactile sensory substitution would excel. Devices like the Tongue Display Unit (TDU), a 10 x 10 matrix of 1-mm electrodes placed on the tongue, and conveying image information via tactile stimulation, provide constant stimulation and updating of the image input, and, like the PSVA, require movement for accurate localization and identification of objects (Matteau, Kupers, Ricciardi, Pietrini, & Pito, 2010). The resolution of tactile devices is lower than that of an auditory device like The vOICe. Yet, as noted, normal obstacle avoidance utilizes peripheral vision, with lower resolution compared to the fovea due to a combination of having fewer cone cells and the cortical magnification factor that favors foveal vision (Lovschky, McConkie, Yang, & Miller, 2005). Peripheral vision is also a primary contributor for magnocellular processing which, like the representation provided by the TDU, is selective for contrast (the TDU has only a black and white, not greyscale, representation) and motion (Livingstone & Hubel, 1988). Thus having a sensory substitution device that provides a better repre-
presentation of contrast and motion might better integrate
with the neural mechanisms used for spatial naviga-
tion. Moreover assessing the importance of contrast
and motion for navigation with sensory substitution
would reveal whether these are the crucial computa-
tional aspects independent of sensory modality, and
thus metamodal.

Second, Weiss, and Sampaio (2005) assessed spa-
tial navigation by having the participants use the TDU
while sitting, but operating a camera-carrying robot
in a three-dimensional maze with a remote control.
Although the accuracy and speed of the maze com-
pletion were assessed and demonstrated success from
the very first attempt, it is difficult to generalize from
this to real-world navigation. A more recent study with
the TDU (Chebat, Schneider, Kupers, & Ptito, 2011)
used a large scale navigation task in a corridor with
obstacles, similar to the navigation employed with The
vOICe (Brown et al., 2011). Participants were able to
successfully navigate the course accurately and con-
genitally blind individuals were able to out-perform
sighted participants, supporting other work that found
route or egocentric spatial knowledge is preferred in the
absence of visual experience (Pasqualotto & Proulx,
2012; Pasqualotto et al., 2013b). The participants were
successful in terms of accuracy, but this study did not
provide as much information about the time required
to complete the task.

Navigation is a major challenge for those who have
acquired blindness later in life (Sampaio & Duter,
1988), primarily due to feelings of insecurity in rela-
tion to the environment, such as obstacles. Certainly
one can learn to depend on different kind of cues to
form representations of space for successful naviga-
tion (Fortin et al., 2008) and utilize white canes and
guide dogs. A recent study assessed two echoloca-
tion sensory substitution for visual devices that provide
online spatial information through sound, using ultra-
sound emitters to detect distance from objects (Kolarik,
Timmis, Cirstea, & Pardhan, 2014). They required
blindfolded participants to walk through an aperture
similar to a doorway, but with varied widths unknown
to the user. With vision, movements were rapid and
accurate, with no collisions. Two pulse-echo sen-
or devices were used, similar in function to sonar.
The Miniguide translates obstacle distance to hand
vibrations. The K-Sonar, due to having a narrower
beam, translates the distance of an individual object,
ethe existence of multiple objects, and also provides
information that can support identification of objects.

There were no significant differences between the
two devices used, the K-Sonar and the Miniguide.
Although performance was less accurate, with some
 collisions, and slower, as expected by walking veloc-
ity, the study provided a proof of concept that such
devices could be of use with further training, consistent
with past work with similar devices (Hughes, 2001).

For example with an aperture 18% greater in width than
the participant, participants had one collision every
trial with K-Sonar, or every two trials with Miniguide;
movement times were less than 2 seconds with vision
but 5 to 8 seconds with Miniguide and K-Sonar respec-
tively. With vision, participants rotated their shoulders
far less to move through the aperture than those
using the devices did. Furthermore the results implied
that the echolocation devices rely on multimodal areas
for depth processing in the occipito-parietal and
occipito-temporal areas (Renier et al., 2005a). As
demonstrated by Kolarik and colleagues, sensory sub-
stitution devices can provide another means of obstacle
detection, and although some studies have exam-
ined virtual navigation (Maidenbaum, Levy-Tzedek,
Chebat, & Amiati, 2013), further research examining
the control of actual action based on information pro-
vided via such devices will be crucial for this form of
visual rehabilitation.

4. Active sensing

Perception is necessary for action (Goodale, 2008);
likewise, action plays a pivotal role in perception,
what has become known as active sensing. Active
sensing has been shown to play a vital role in per-
ception of various modalities, including the visual,
the auditory, the tactile and the olfactory systems
(Schroeder, Wilson, Radman, Scharfman, & Lakatos,
2010; Wachowiak, 2011), in humans as well as a vari-
ety of other species (Nelson & MacIver, 2006). Active
sensing is the movement of the sensors for optimization
of sensing performance (Mitchinson & Prescott, 2013),
and it plays a central role in the process of perception
(Ahissar & Arieli, 2012).

The studies reviewed here report relatively high
success rates for the generation of action under SSD
guidance. These results indicate that the SSD users
are able to maintain the appropriate calibration of
their movements based on SSD feedback, and suggest
they are making use of the SSD feedback to fine-tune
their movements. Movements of sighted individuals
in the absence of visual feedback, have been shown to be miscalibrated, and not corrected based on proprioception alone (Levy-Tzedek, Ben Tov, & Karniel, 2011; Levy-Tzedek, Krebs, Arle, Shils, & Poizner, 2011; Soechting & Flanders, 1989; Wolspert, Ghabramani, & Jordan, 1995). It can therefore be concluded that the effective generation of movement is a direct result of successfully deciphering cues from the SSD, rather than relying on proprioceptive cues, or on memory (e.g., from visually experiencing the experimental setup, where applicable).

Perception of depth cues plays an important role in generation of a fuller representation of spatial layout, an ability which has been demonstrated when 2D images were conveyed via an SSD, such as the PSVA (Renier et al., 2005a). We estimate it would therefore be beneficial to combine input from SSDs that provide image information (e.g., the vOICe) with distance-estimation devices (e.g., the K-Sonar), while actively exploring a scene. Exploratory attempts in this direction have been made using newly available depth sensors such as the Kinect (Gomez, Mohammed, Bolognani, & Pun, 2011).

5. Sensory substitution as cross-modal “translators”

Training on how to use SSDs is a major issue that should be considered, as the choice of a particular training paradigm may affect the ultimate performance. For example, active training has emerged as having particular benefits in recent accounts (Proulx et al., 2014; Ricciardi, Bonino, Pellegrini, & Pigazzi, 2014); this is particularly important given that most research in this area falls within the domain of invasive approaches. The study of sensory substitution and visual impairment can also reveal the neural mechanisms underlying perception for action that supersedes just one sense (Proulx et al., 2014; Ricciardi, Bonino, Pellegrini, & Pigazzi, 2014). It can therefore be concluded that concurrent sensory input arriving from different modalities has been shown to improve perception (for a review, see Proulx et al., 2014). Therefore, once the users learn how to interpret the signals arriving from the SSD, and make visual sense of them, they can then use the SSD in combination with other rehabilitation approaches, e.g., with a visual prosthesis, and serve as an interpreter of sorts, which can guide and assist the rehabilitation process of the more invasive approaches.

6. Using sensory substitution to reveal the neural mechanisms underlying perception for action: Ventral & dorsal streams in the blind

The study of sensory substitution and visual impairment can also reveal the neural mechanisms underlying perception for action that supersedes just one sense (Proulx et al., 2014; Ricciardi, Bonino, Pellegrini, & Pigazzi, 2014). This is particularly important given that most research in this area falls within the domain of vision. Visual information is segregated into different features early in processing, even at the level of the retina, and later into two parallel cortical pathways, the ventral stream and the dorsal stream, which appear to be functionally segregated, with the former discriminating object features and the latter handling information about spatial location (Goodale, 2008). The ventral occipitotemporal “what” pathway extends from the primary visual cortex (V1) through visual area 4 (V4) to inferotemporal areas and then up to the prefrontal cortex. The dorsal occipitoparietal “where/how” pathway from V1 leads through the posterior parietal cortex towards the premotor cortex (Sterns-Amiot, Dukwar, Reich, & Amedi, 2012). The dorsal stream is critical for real-time control of action, providing visual guidance for movement (Milner, 2012). The ventral stream supports object, event, and action identification (Goodale, 2008) as well as feature analysis for object identity independent of variance in other features for which identity should be invariant, such as position and orientation (Freeman & Ziemba, 2011). The role of the ventral stream in action, then, is to provide visual information to enable the identification of a specific object and activate the cognitive systems to plan the action performed towards the object. These streams were pri-
A growing body of research examining perception, action, and cognition in the visually impaired suggests that the brain is not necessarily segregated into sensory-specific areas but instead is metamodal and organized in terms of the tasks, functions, and computations that the areas carry out (Pascual-Leone & Hamilton, 2001; Proulx et al., 2012); this is sometimes referred to as supramodal as a synonym for metamodal in the literature (Kupers, Pietrini, Ricciardi, & Ptito, 2011; Kupers & Ptito, 2011; Ricciardi & Pietrini, 2011). This perspective, in contrast with the classical one, classifies brain areas by the form of information processing carried out (e.g., shape recognition) independent of the sensory modality that appears to be employed.

First we will review the evidence for the metamodal organization of the brain. Then we will note the natural extension of this form of functional organization to perception for action, in much the same way that the dorsal stream has been re-characterized from processing “where” information (Mishkin & Ungerleider, 1982), to being more task-specified as the processing of “how” (Milner & Goodale, 2008) such information can be acted upon.

Two examples of cortical regions, first thought to be visual in nature, but later discovered to be non-modal, are the fusiform face area (Kanwisher, McDermott, & Chun, 1997) and the parahippocampal place area (Epstein & Kanwisher, 1998) that are now known to respond to the same categories even when only perceived by touch or sound (Ricciardi & Pietrini, 2011). An extensive review by Ricciardi and Pietrini (2011) details many sources of evidence for the supramodal, or metamodal, functional architecture of the brain.

Striem-Amit et al. (2012) conducted a brain imaging study to examine the functional segregation of the dorsal and ventral streams in the congenitally blind. Using the vOICe, congenitally blind participants were presented with different simple novel shapes in different areas of the “visual” field, and these were transformed into sound. The participants were asked either to identify or locate the shape. This elicited clear differentiation between ventral and dorsal streams, indicating that even with no visual experience, these two streams still encode the tasks they would be expected to perform had it been visual information. To be more specific, shape identity tasks activated the ventral occipital temporal sulcus, which is located in the midst of the ventral stream. In contrast, the localization task preferentially activated auditory regions, such as the supramarginal gyrus located in the inferior parietal lobe, as well as the precuneus, in higher order area of the dorsal stream.

Critics of research at studies using SSDs are that they usually study highly proficient users of the system, who may have already undergone a plastic change, such as Pito, Moesgaard, Gjedde, and Kupers (2005) who showed increased occipital cortex activation following SSD training. Striem-Amit et al.’s (2012) study circumvents this limitation by only training participants to use the devices for a maximum of one and a half hours and participants were completely naïve to The vOICe before this training, thus avoiding the problem that brain imaging might be measuring an artificial plastic change rather than what is inherent in the congenitally blind.

Furthermore, many studies exploring the use of SSDs in the congenitally blind use familiar, well-practiced stimuli, such as Kim and Zatorre (2011) who trained their participants for five days with the stimuli used for the experiment. This means that potential findings were conflated by participants using memory of the stimuli presented in training, rather than shape processing in the occipital region. Striem-Amit et al. (2012) used novel stimuli to ensure that memory was not playing a role in the activation of dorsal and ventral streams, just shape and location processing. The neural correlates of action control and representation has not been only studied using sensory substitution, but there is extensive evidence using standard paradigms in sensory-deprived individuals, further giving support to the metamodal nature of the underlying neural architecture for action (Renzi et al., 2013; Emiliano Ricciardi et al., 2009; Röder, Kusmiercz, Spence, & Schicke, 2007).}

Qin and Yu (2013) postulated that the metamodal nature of the visual cortex could be due to a plastic change in the brain that is caused by the deprivation of visual stimuli. Thus, rather than being inherently metamodal, the visual cortex has been shown to be recruited by other higher functioning processes, such as memory and cognition in the blind (Borton, Sinclair, & Agato, 2012; Rentier et al., 2010), and therefore might not be metamodal naturally. Striem-Amit et al. (2012) refute this suggestion by demonstrating this metamodality in the blindfolded sighted as well as the blind, indicating that the meta-
modal nature of the two streams is not due to sensory deprivation.

This study seems to suggest that a metamodal approach to understanding the brain structure is a more valid approach compared with the standard unimodal model. If the metamodal theory is to be substantiated, more areas of the brain must be demonstrated to be metamodal. Reich, Sewed, Cohen, and Amedi (2011) studied the “visual” word form area (VWFA), a component of the visual cortex, to try to establish whether this area is metamodal. The VWFA in the sighted is activated across writing systems and scripts and encodes letter strings irrespective of case, font or location in the visual field. In this study blind participants were presented with Braille words and nonsense Braille letter strings. The authors used fMRI to investigate whether the VWFA would be activated in response to the tactile letters. The investigators found a striking differentiation between the real Braille words and the nonsense words in the congenitally blind, indicating that the VWFA is activated for reading words regardless of input, plus develops specialization for reading regardless of visual experience, supporting the metamodal theory. Braille reading activating the primary visual cortex in the blind has also been shown by Cohen et al. (1997), thus strengthening the idea that the visual cortex is a metamodal area of the brain that can be activated by any sensory modality.

A further brain area that has been put forward as metamodal is hMT+ (also called the fifth visual area, V5), which encodes for motion. A study by Montanu et al. (2010) found this area was activated when perceiving motion using an SSD that rests on the tongue in both the sighted and the blind, again indicating that this area encodes information about motion regardless of sensory input. So far in this review only visual areas have been discussed as being metamodal. For the theory to be substantiated fully, other areas of the brain that have traditionally been considered to be unimodal must be demonstrated to be activated by more than one modality. For example, the auditory cortex would be expected to be activated if a non-audio input was presented that required the same computation as an audio signal. Calvert et al. (1997) used an fMRI with participants with normal hearing. The authors first identified the auditory cortex by playing participants speech through headphones and then asked them to repeat the words back to themselves in their heads. Then they played video of a face silently mouthing numbers, which the participants had to repeat back to themselves in their heads. This lip-reading led to activation in the auditory cortex, despite there being no auditory signal present, indicating that the auditory cortex is also metamodal (Calvert & Thesen, 2004). This is further corroborated by Okada et al. (2010) who presented 20 participants audio input alone compared with audio and an accompanying congruent visual speech input. Audio plus visual led to more activation in the auditory cortex than audio input alone, indicating that areas of the so-called auditory cortex are similarly responsive to visual stimuli, again supporting the idea of the brain as a series of modules that deal with specific computations, rather than being described by the input they receive. Other research suggests similar findings for other “primary” sensory areas, such as the primary olfactory cortex, which has been shown to be responsive to other associative information (Chapuis & Wilson, 2012; Weiss & Sobel, 2012). Note that the metamodal, or supramodal, nature of a brain area can be further demonstrated by studies comparing sighted and blind individuals, as this would exclude visual imagery as a potential confound. Indeed many of these areas described have been tested with such an experimental design, and a recent review (Ricciardi et al., 2014) provides examples of such evidence.

How might these cortical regions, thought to be primary sensory areas, respond to information provided by other sensory modalities? For example, the visual cortex must be connected to the cochlear nuclei in the brain stem (Recanzone & Sutter, 2008) for there to be activation in this area when presented with images in sound via The vOICe. Two hypotheses have been put forward in explanation of how these, previously thought to be unlinked, peripheral sensors and brain regions are connected. According to the cortical reorganizational hypothesis, cross modal brain responses are mediated by the formation of new pathways in the sensory deprived brain (Bronchti et al., 2002). A more compelling argument is the unmasking hypothesis, which states that loss of sensory input induces unmasking and strengthening of already existing neural connections (Kupers & Pütö, 2011). The unmasking hypothesis is suggested here to be correct, as blind participants in the Striem-Amit et al. (2012) study showed activation in visual pathways when using an SSD with only minimal training, precluding the idea that the brain had a chance to
Furthermore, Falchier, Clavagnier, Barone, and Kennedy (2002), using retrograde tracers on cynomolgus monkeys, showed that the primary visual cortex (V1) receives input not only from the visual thalamus but also receives projections from the auditory cortex as well as polysensory areas of the temporal lobe. Further anatomical evidence of multiple other connections projecting to the visual cortex comes from Rockland and Ojima (2003), who showed that areas V1 and secondary visual cortex (V2) receive projections from several parietal and auditory association areas in macaque monkeys. These anatomical explorations in monkeys show that the previously thought of as unimodal areas of the brain are actually connected to other sensory areas directly, rather than just via upstream multimodal association areas, and these data support the idea that the connections are extant in all and are unmasked due to sensory substitution tasks rather than created by the brain following sensory deprivation.

However, the debate continues, with evidence coming from Kupers et al. (2006) who used TMS to blind and sighted participants trained to use the tongue display unit (TDU). The TDU is a sensory substitution device that places a 3 × 3 cm electrode array consisting of 144 gold plates on the tongue, and an electrical pulse is generated by each electrode to correspond to a pixel on the image that is being assessed (Bach-y-Rita & Kercel, 2003; Pits et al., 2005). When TMS over the occipital cortex was used, blind participants reported somatotopically organized tactile sensations on the tongue, whereas sighted individuals who had also been trained with the TDU did not report this (Kupers et al., 2006). This indicates that a few alternative explanations might be possible. First, that there are differences in the blind that are not found in the sighted, suggesting that instead of there being masked connections between peripheral tactile sensors and the occipital cortex in everyone, there is actually a new connection created between cortex and sensor in the sensory deprived, supporting the cortical reorganization hypothesis. Another option is that the unmasking occurs to a different extent for blind and sighted individuals. That is, perhaps in the blind brain there is the potential for faster, more efficient unmasking than in the sighted brain.

Based on the evidence presented here, areas of the visual cortex have been demonstrated to be metamodal in nature, such as the dorsal and ventral streams, the VWFA and V5. Further the auditory cortex has been shown to become active when an appropriate computational task is presented to it, despite this task coming in the form of visual input (Calvert et al., 1997). Therefore a system which describes brain structure based on the task or computations that an area performs, rather than the sensory modality that is most often encoded by an area, seems more accurate taxonomy. Previously the brain areas were thought to only be connected to the specific sensory modalities that they computed. However, evidence from lesion studies in monkeys and brain imaging studies in humans have shown that these metamodal areas are in fact linked up to peripheral sensors that were previously not thought to project to them. This leads to the idea that the brain is split into metamodal, computational based modules, that are linked to all sensory inputs and that can process information that is requisite to their computational task, regardless of the modality that this information originates from.

The results derived from blind participants supporting a metamodal organization of the brain extend to sighted participants as well. Long-term blindfolding of sighted participants with training in using The vOICe (Proulx et al., 2008) resulted in brain activation similar to that of blind participants in visual cortex but with auditory stimulation (Boroojerdi et al., 2000; Facchini & Aglioti, 2003; Lewald, 2007; Merabet et al., 2008; Pollok, Schnitzler, Stoerig, Mierendorf, & Schnitzler, 2005). Such effects can occur in a matter of days, too, as in a study with five days of visual deprivation and Braille training (Kauffman, Theoret, & Pascual-Leone, 2002; Pascual-Leone & Hamilton, 2001). There is mounting evidence that the visual cortex is functionally involved in auditory and tactile tasks in sighted subjects as well (Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006; Liang, Mouraux, Hu, & Iannetti, 2013; Sathian & Zangaladze, 2002; Zangaladze, Epstein, Grafton, & Sathian, 1999).

Thus far, research using sensory substitution has revealed the metamodal properties of brain areas once thought to be visual in nature. Although such research has targeted both ventral and dorsal streams of processing, those studies examining the dorsal
stream have focused on “where” information in the perceptual sense rather than “how” information in the perception for action sense (Striem-Amit et al., 2012). Much of what is known about the neural basis of perception for action comes from studies of vision. For example, a large body of work has detailed the role of the anterior intra-parietal area (AIP) in macaques for the processing of object shape, size and orientation to guide grasping (Taïa, Mine, Georgopoulos, Murata, & Sakata, 1990). A similar region has been found in humans as well (Binkofski et al., 1998), with a number of studies detailing the role of the posterior parietal cortex for the coordination of eye and hand movements when reaching for targets (Desmurget et al., 1999; Van Donkelaar, Lee, & Drew, 2000). Are these vision-specific perception for action areas, or might non-visual information guide hand movements and grasping in these cortical regions? The metamodal hypothesis would have one predict that the task (grasping and hand movements) of AIP define its function, and that any sensory input that provides the necessary spatial information and object features would also evoke such functional recruitment. Sensory substitution would be an ideal approach to examine whether AIP and other areas involved in perception for action are not specific to one sensory modality, vision, but rather allow for precise hand and grasping movements on the basis of any informative sensory modality that provides input.

7. Shared spatial representation across the senses

Spatial information is shared across sensory modalities, such that information obtained via one modality (e.g., vision) is utilized when making movements based on input from another sensory modality (e.g., audition) (Levy-Tzckdek et al., 2012b). These findings are in line with evidence from neural recordings showing that the dorsal stream participates in mediation of sensory-motor control, even when the sensory modality is not vision (Fiehler, Burke, Engel, Bien, & Rösler, 2008), suggesting a modality-independent representation of action control. Several studies have demonstrated (Rossetti, Stelmach, Desmurget, Prablanc, & Jeannerod, 1994) that direct vision of the target immediately prior to movement significantly increased movement accuracy, whereas a delay between target presentation and movement resulted in deteriorated accuracy; that is to say, the longer participants had to use their cognitive representation of the target, the less accurate were their movements; this line of results was the basis for the suggestion that vision is needed to calibrate the proprioceptive map; and yet, participants in the studies reported here were able to use the auditory or tactile information (possibly to create a ‘visual’ like representation which is the ultimate goal of SSDs for the blind), in some cases without the possibility to perform vision-based calibration between trials, in both plan the movements and improve the accuracy on subsequent trials.

Proulx and colleagues (2014) reviewed these and other results to suggest that the representation of space might not be dependent on the modality through which the spatial information is received. This is supported by brain-imaging studies indicating that areas within the visual-dorsal stream, such as the precuneus (Kübler & De Volder, 2010) and the right anterior middle occipital gyrus (Collignon, Champoux, Voss, & Lepore, 2011), showed a functional preference for spatial information delivered using the auditory or tactile modalities in congenitally blind and in sighted individuals. These brain areas are also activated when sound encodes spatial information, rather than arriving from different spatial locations, in both sighted and congenitally blind individuals (Striem-Amit et al., 2012). This hypothesis, of a task-specific, rather than a modality-specific structure of brain areas was similarly suggested by research indicating that, during a spatial navigation task with the TDU, congenitally blind individuals were found to recruit the posterior parahippocampus and posterior parietal and ventromedial occipito-temporal cortices, areas that are involved in spatial navigation under full vision (Chebat et al., 2011); similarly, the dorsal pathway has been shown to mediate movement control when only kinesthetic, and no visual information was used to guide actions even when tested in congenitally blind individuals, with no prior visual experience (Fiehler et al., 2008). It has been suggested that blind individuals tend to rely on a more egocentric, rather than an allocentric frame of reference (Cattaneo, Bhatt, Merabet, Pece, & Vecchi, 2008; Pasqualotto & Proulx, 2012; Pasqualotto et al., 2013b; Rösler et al., 2007). As data are accumulated, supporting the hypothesis that the representation of space in the brain is modality-independent, or that very little training is required to create a ‘visual’-
like representation of space using sounds, we suggest that SSDs, by providing detailed spatial information for people with impaired vision, might allow them to employ an allocentric spatial frame of reference when interacting with their environment.

8. Future directions for the online visual control of action through sensory substitution

Here we have reviewed the few studies that have begun to explore how sensory substitution might allow for the online control of action via visual information perceived through sound or touch. One limitation of many of these studies so far is that performance has often been assessed solely in terms of accuracy, with only a few that have also reported response times so that the full efficiency of the behaviors could be assessed.

There are a number of areas where the classic approaches to studying the online control of action can be adopted to address interesting issues with sensory substitution. One fascinating example would be to examine how size illusions via sensory substitution influence perception and action, and whether visual experience is necessary for an illusion to be effective. A crucial test to assess whether there is a clear distinction between perception and action has been whether a size illusion, such as the Müller-Lyer illusion, impacts both perceived size and grasping. Recent research (Biegstraaten et al., 2007), and a comprehensive review (Bruno & Franz, 2009), suggest that there is not necessarily a dissociation between perception and action. This has also been supported by work in other related areas using illusions and other approaches (Bruno, Knox, & de Grave, 2012; de Brouwer, Brenner, Medendorp, & Smeets, 2004; Foster, Blankeboer, Leitheit, & Franz, 2012; Franz, Hesse, & Kollath, 2009; Mendoza, Elliott, Meegan, Lyons, & Welsh, 2006). This would be an interesting approach to assess the mechanisms of online control with sensory substitution because the perception of size using other size illusions has also been conducted with sensory substitution. Laurent Renier and colleagues used the PSVA sensory substitution device to assess whether blind and sighted participants would perceive the illusion through interpretation of the sound created by the device (Renier et al., 2005b; Laurent Renier, Bruyer, & Volder, 2006).

In these studies they found that although the sighted and late blind participants perceived the Ponzo illusion and the Vertical-Horizontal illusions just as the sighted do, the early blind participants did not. This has the fascinating implication that visual experience is necessary to have the expectations required to perceive size in a way that is influenced by context. They did not assess perception versus action, and it would be of great interest to assess whether these illusions, perceived through sensory substitution, affect grasping differentially depending on visual experience. Another interesting extension would be to assess the impact of using another sensory substitution approach, such as tactile substitution, on the influence of size illusions. Given the fascinating finding in other work that the online control of action might be sensory modality independent as seen from converging work looking at learning and adaptation with sensory substitution devices (Levy-Tzedeck et al., 2012b), there is a lot of potential to use sensory substitution as a means for revealing the underlying mechanisms for the perceptual control of motor action. The next step would be to move beyond the limited data available on online control with full reporting of both response time and accuracy to allow a full assessment of sensory substitution performance. Future work looking at having shifting targets at the time of response initiation and double-step paradigms for reaching to targets (Goodale et al., 1986; Gritsenko et al., 2009) would best allow work on perception for action with sensory substitution to contribute to the fundamental understanding of online control.

Acknowledgments

This work was supported by a grant from the EPSRC to Michael J. Proulx (EP/J017205/1). Partial support was provided by the Helmsley Charitable Trust through the Agricultural, Biological and Cognitive Robotics Center of Ben-Gurion University of the Negev.

References


Levy-Tardec, S., Ben Tov, M., & Karniel, A. (2011). Rhythmic movements are larger and faster but with the same frequency on removal of visual feedback. J Neurophysiol, 105(2), 2120-2126. doi: 10.1152/jn.00626.2010


