Running head: VISUAL PERCEPTION THROUGH SENSORY SUBSTITUTION

Exploring Visual Perception Through Auditory Sensory Substitution

Alex Storer
Department of Cognitive and Neural Systems
Boston University
storer@cns.bu.edu

In the style of the Journal of Visual Cognition: Full Article format. This is a Term Paper for CN730, a Graduate Level class in Vision, and the experiments herein were performed without supervision by or consultation of any human subjects committee.

Abstract

Sensory substitution is the presentation of the information normally associated with one sensory modality (e.g., visual information) through a distinct sensory modality (e.g., the auditory system). The vOICe is one such system, which converts images from a webcam into “soundscapes” (presented via headphones), which experienced users are able to interpret as visual in nature. This paper describes a series of experiments to test the limits of the vOICe both in natural environments and using controlled stimuli. These experiments show that a blindfolded user can learn how to use the vOICe through training, and that after such training the subject may process illusory contour stimuli differently from naive subjects. Although the naive subject reported no strong visual percepts, this paper shows that even simple sensory substitution systems may be a useful psychophysical tool to explore many aspects of sensory processing.

Exploring Visual Perception Through Auditory Sensory Substitution

Introduction

Sensory substitution is a paradigm by which one sensory modality may be presented through the use of another. Sensory substitution paradigms in the past have focused on a large number of possible combinations, including converting visual information into both tactile and auditory stimuli (Bach-y-Rita, 2004; Meijer, 1992). Thusfar, this type of technology has been explored by many different research groups, most of which focusing on the development of assistive technologies. Both blind and sighted users have been trained to use sensory substitution systems successfully, and after training are able to identify objects in the world, assess their distance, and even see effects such as the Ponzo illusion (Renier, Laloyaux, et al., 2005).

Of course, many researchers do not consider these findings to be representative of visual experience, and thus should not be classified as "vision" (Lenay, Gapenne, Hanneton, Marque, & Genouelle, 2003). This particular argument, however, is a very difficult one to have, because vision is not often clearly defined, and even if it were, it would most likely rely heavily on subjective experience ("what it's like") about which comparisons across subjects are challenging. The most stirring evidence to the contrary is found in the testimonials of users of these devices, many of whom attest to visual sensations, as documented online (Meijer, 2006). Many philosophers, on the other hand, hold that because many sensory-motor contingencies exist in the substituted modality, they should be considered as the same sense (O'Regan, Myin, & Noë, 2005). In particular, this refers to how the sensory input is altered as a result of, for example, moving the head. When a camera is mounted on the head, the input to the camera behaves much the same way the input to the eyes would behave, and thus, these many of contingencies are preserved.

Regardless, the phenomenon of sensory substitution is an interesting one, and can be used to address a number of fundamental perceptual questions. In particular, it allows the assessment of a developing sensory modality, which cannot commonly be assessed in normal or disabled adults. By having control of the factors that lead to a qualitatively sensory experience, one can modify them and observe how this affects the speed at which the sensation is acquired, or alternatively, whether it alters the quality of the sensation. Similarly, comparisons can be made across input modalities to isolate the effect of, for example, vision being processed by the eyes and lateral geniculate nucleus as opposed to alternative pathways.

While most successful work in visual substitution has focused on tactile inputs, notably the work of Paul Bach-y-Rita, working in the domain of audition yields a number of advantages (Sampaio, Maris., & Bach-y-Rita, 2001). In particular, it precludes the necessity of a tactile array to convey visual information - these arrays are expensive devices that are not widely available. The alternative in the auditory domain is simply a set of headphones, which is essentially free by comparison. Headphones are also less bulky and require less power, making them more portable, which aids versatile scientific study and use as an assistive device.

Auditory substitution systems have been relatively few, and have been developed in quite distinct ways. Some systems use sonar input as opposed to visual input, but more recently, most auditory sensory substitution systems have used cameras to provide visual input. Some groups have made efforts to simulate vision as accurately as possible - in particular, Christian Capelle has made efforts to provide inputs with both lateral inhibition and foveation (Capelle, Trullemans, Arno, & Veraart, 1998). In his system, the center of the visual field is represented with more pixels, and each pixel is associated with a particular sinusoidal auditory frequency. Of course, this system, much like any auditory system, cannot present the visual world continuously, and must sample it somehow. In this auditory coding scheme, neighboring pixels are not necessarily represented by neighboring sinusoids, which makes such a system initially unintuitive.

The system developed by Peter Meijer takes a different approach, as the image is scanned from left to right as opposed to being presented instantaneously (Meijer, 1992). In his system, the frequency presented is a function of the row number, and the amplitude is a function of the pixel intensity, or brightness. One column consists of all available sinusoids, and is played for a short duration before moving to the next column. The left side of an image is presented primarily to the left ear, and the right side is presented to the right ear, with central locations of the image presented binaurally. Such a system contains a very intuitive mapping, but because of the scan time from left to right, motion is not effectively encoded. Because of the system's simplicity and because the necessary software is available online (as the "vOICe"), it is a popular choice for home users with relatively few resources (Meijer, 2006). As an application, the vOICe gives the user a high degree of control over parameters, including filtering and number of available sinusoids, which makes the parameter space optimizable for the needs of individual users.

Although the vOICe is certainly not the most effective system to use in the study of sensory substitution, it is sufficiently simple to both use without difficulty and demonstrate simple results as a proof-of-concept. One area of visual perception that has not yet been explored with a vision substitution system is illusory contour formation. Illusory contours pose an interesting question to investigate, as their is still debate as to which areas of the brain are necessary and sufficient for their formation (Lee, 2002). While studies have reported that illusory contours cause firing in V1 and V2, others have established that damage to higher regions such as IT can prevent perception of illusory contours altogether.

If illusory contours are perceived through use of the vOICe this could lead to a number of interesting conclusions. To begin with, it is unclear whether large degrees of training with a system such as the vOICe with sighted subjects will lead to activation of visual cortex. If not, then successful perception of illusory contours through use of the vOICe would imply that either multimodal or purely auditory areas are involved in illusory contour perception. This result would provide interesting evidence both with regard to the formation of illusory contours, but also with respect to the nature visual perception itself.

Thus, this study seeks to explore the perception of illusory contours by first training a subject to use the vOICe, and then testing on stimuli with or without illusory contours. In particular, if the subject differentially processes stimuli with illusory contours as opposed to without, this will suggest that perhaps illusory contours are being formed.

Methods

Throughout all training and testing, the vOICe Learning Edition (Meijer, 2006) was utilized on a laptop PC running Windows XP (Microsoft Corp., Redmond, WA). The vOICe is initialized with a set of defaults which were not generally not altered. Input to the vOICe system was provided by a USB webcam (Logitech Quickcam, Apples, Switzerland) and interfaced directly with the vOICe. Apple iPod (Sunnyvale, CA) headphones were utilized by the user of the system, who controlled the volume to his liking. The webcam was secured under the brim of the hat, and adjusted before each session to ensure that it was pointing in a direction that was satisfactory to the user. Full details of the functionality of the vOICe are available in the above citation, and will not be discussed presently.

Two subjects were used in this study, one male and one female, both aged 23, with uncorrected vision and hearing. The male participated in training and testing segments, while the female only participated in selected testing segments as a control.

Training

The trained user spent 25 hours using the vOICe spread across 11 weeks, including training, testing with feedback and testing without feedback. During this entire course of time the user was blindfolded, and reported no ability to perceive light. The training period consisted of a period of mobile exploration in a familiar environment for at least 30 minutes at a time. During these training periods, the author accompanied the user to prevent unnecessary damages to person and property, to answer any questions the user had about the present environment (e.g., object identification, navigation, etc.) and design tasks that necessitate the use of the vOICe. During these sessions, the subject had use of an inflexible stick of length roughly 1m to utilize as he saw fit. While the tasks set to the user were not exactly specified, they consisted primarily of identifying and collecting novel objects, navigating both familiar and unfamiliar spaces, and drawing images on both paper and whiteboard. The purpose of this time was to present a set of sensory-motor contingencies necessary to learn a visual mapping (O'Regan et al., 2005). In addition to this, in the early stages of training (the initial 4h), the user was encouraged to use a blindfold but not required to do so, and was presented shapes directly through the vOICe program without using a webcam. These shapes were quadrilaterals of all types, starts of varying point numbers, circles and ovals, and other common shapes, generated by hand with no underlying pattern.

Testing With Feedback

After training as described above had transpired for 10 hours, the user was presented with a formulaic set of inputs directly to the vOICe, consisting of squares or circles. These shapes were placed on a grid (which was itself hidden to the user) and the user was asked to identify the shape presented, and the location of the shape in the coordinates of the grid, which contained five spots on the x-axis and four on the y-axis.

These inputs were presented in groups of 32, a set which took approximately one half hour to complete. After giving a response, the author presented the correct response, and engaged the user in dialog regarding his perception of the prior shape. Before each such testing with feedback trial, one exploratory trial was undertaken.

After five such trials, a similar set of inputs was presented, with the shapes randomly rotated (by a factor of 10 degrees). The user was then asked to identify not only the shape and location, but also the rotation of the given figure when relevant.

The purpose of these experiments with feedback was to train the subject to identify simple shapes, and also to track his learning over time in a quantifiable fashion.

Testing Without Feedback

After ten total trials in the above system, five with rotation and five without, the user was presented with a novel set of inputs to which he had never been exposed. The naive subject was tested in an identical fashion, given as little information regarding the structure of the system as was possible. In this task, again sets of 32 trials were presented, each trial consisting of a shape, a three second waiting period, and another shape. The shapes consisted of "pac-men" as illustrated in Figure 1, arranged in the Kanisza square formation. The users' task was to identify whether the first configuration was the same as the second configuration, which was true 50% of the time.

Half of the images consisted not of pac-men aligned to a square, but rather, randomly rotated, and all images in total were rotated randomly and placed randomly on the screen. For the shapes that changed from the first presentation to the next, two changes were possible. The first is consistent with the illusory square ("consistent"), such that one of the circles appears to be subsumed underneath the illusory square. The second is inconsistent with the illusory square ("inconsistent"), such that the entire pac-man figure is translated inwards. These cases are presented in Figure 1, but it is important to recall that all individual elements are randomly rotated half the time. Subjects had control of the volume, and the time of presentation for each image, but were not able to return to an image that had been presented previously. In addition, no feedback was given regarding performance, and similarly, no descriptions of the present stimuli were given to the subjects. Two of these sessions of 32 trials were administered.

Thus, this experiment investigates whether it changes are more readily detectable if the illusory contour is preserved, or if the illusory contour is disturbed. If present, this would constitute differential processing based on illusory contour presence or absence. As a control, randomly rotated shapes make exactly the same changes, but are not in a "square" configuration. If the configuration itself is not important, then there should be no difference between the randomly rotated and aligned conditions.

Results

The results of this experiment contain two components, one being the outcome of the testing phase, and one being an anecdotal description of the sensation and efficacy of using the vOICe in mobile tasks. The results are summarized in Figure 2, the result of identifying and locating rotated squares and circles, and Figure 3, the results from both trained and naive subjects testing on the same or different task described above.

In the training phase, the subject had a clear upward trend in terms of percent correct, considering shape identification and localization. While the data for angle identification are ambiguous, one must recall that this data reflects performance after training for five sessions for shape identification and localization, with no training for angle identification. The testing phase demonstrates the difference between the naive and trained subjects, and none of the differences are significant according to a paired t-test ($p >> 0.05$). Within the trained subject, however, there are statistically significant differences between the "consistent-random" and "consistent-aligned" tasks

$(t(14) = 2.2193, p = 0.0380)$, and also between the "inconsistent-aligned" and "consistent-aligned" tasks ($(t(20) = 2.5, p = 0.0212)$ using an unpaired t-test. To recapitulate, this is a difference between accuracy in identifying changes when they are present, and comparing across conditions. There is a difference between when illusory contours are preserved versus destroyed (consistent-aligned v. inconsistent-algined), but also a difference between configuration in a square versus random configuration (consistent-aligned v. consistent-random).

In terms of more anecdotal results, the trained subject reported an increasing awareness of what was present in the visual field, but never felt as though the experience was visual in nature. In addition, his performance on tasks such as identifying objects via head-mounted web cam was not nearly as accurate as for those using vision. In particular, some salient objects were easily identifiable, in particular the stick when it had fallen, but many objects were difficult to identify using the vOICe alone. Combining it with, for example, reaching or probing with the stick helped performance greatly, but was still not anywhere approaching skillful. Navigation was generally successful in familiar environments - anomalous objects, people or unexpected boundaries had a large impact on navigation skill. Also noteworthy is the skill that the subject developed in navigation tasks, that is to look for prominent features and track them over time. Bright signs on dark backgrounds, angled objects and texture transitions were used as signposts and also markers. One prominent example is a dark stripe where the carpet joins the wall that can the subject used as the "curb" in the hallway, to help stay centered and identify corners.

Discussion

Overall system performance is quite complicated and perhaps analyzed in smaller components.

Discussion of Anecdotal Results

As is mentioned in the results, the most interesting point is that a blindfolded user utilizing the system for over 20 hours was not able to develop a percept that is visual in nature. There are several reasons why this could be the case, many of which probably play a role in this. First, it could simply be that 20 hours is not a long enough time to train, especially discontinuously, to begin to associate vision with any substitution system. An optimal situation for generating visual percepts would probably involve continued use of a given system for days or perhaps longer. A training regime such as this one is clearly not a sufficient mechanism for the development of a new sensory mapping.

Another problem that should be discussed are the flaws in the vOICe itself. Because it scans from left to right at a rate on the order of a second per scan, it is very nearly impossible to develop any sense of motion, which thus eliminates instantaneous visual feedback. Consider this in particular in the context of head movement - oculomotor control consists of sophisticated mechanisms to compensate for the movements that human beings make on an everyday basis. Additionally, when images are passed across the webcam, they are seen as blurred even before being compressed into drastically fewer pixels for conversion into sound. While many discuss the sensory-motor contingencies of vision, such as how the world moves on our retina as a result of our actions, and their relationship to sensory substitution (O'Regan et al., 2005), it is worth noting that these contingencies are not as widespread with a system such as the vOICe, as is evidenced by its treatment of motion.

Other difficulties include the very small visual angle that is attainable with a webcam. Using the voice is equivalent to only being able to foveate, and to this fovea being extraordinarily weak, in terms of resolution. The lack of any peripheral vision make tasks such as visual search nearly impossible. Even fixating on something can prove challenging, because small movements of the head translate into large changes in terms of where the webcam is pointing. Because of the slow refresh rate, it is nearly impossible to correctly for these small changes, and keep objects in the field of view. Moreover, identifying an object is not only quite difficult, but very difficult to interact with. In an attempt to grasp an object, one must make a blind reach, because the webcam will invariably move, and there is no way to monitor the hand as it moves due to the slow refresh rate. Reaching and missing an object by only a centimeter is as bad as having never found it, because it is then not obvious as to whether the object was present or absent, and after moving it is no longer visible. These and other problems make sense, but are not obvious prior to extensive system use over a variety of tasks.

Finally, it is important to emphasize the top-down nature of the use of the vOICe. In different contexts, a white square can mean an exit sign, a window, a piece of paper or any infinite number of things. Correctly interpreting a scene requires a good deal of knowledge about context - knowing that an object in view is a person can help a user of the vOICe interpret its shape, but without this prior knowledge, it is difficult to deduce that the object is a person.

Discussion of Numerical Results

While the results indicate that there is a statistically significant difference in processing between cases in which illusory contours were preserved as opposed to broken, there are several caveats to drawing conclusions from this. The first is that the trained subject was not able to correctly report any qualities of the input. He reported that the Kanizsa figures were groups of smaller objects, and could not identify a global structure of any kind, but rather identified them as amorphous blobs. This ties in to the above discussion of the importance of top-down information - without knowing what the possible set of inputs is, it was impossible to identify any particular input.

Another issue is that due to the small number of subjects and the relatively small number of trials, each response was treated as being drawn independently from a given distribution, which is in this case discrete. A t-test also makes underlying assumptions of Gaussian distributions of data, which is not possible in a True/False paradigm. Even so, these results point towards possible differences in processing as a result of training, in particular, that colinear inputs are processed differently than inputs that are not colinear. The "consistent-aligned" case is the only case that preserves all illusory contours, and it was identified at a very low rate, around 50%. The same exact change enacted on a "random" input (that is, not having a square shape) was identified over 80% of the time. Images that were aligned to begin with and then had an entire pac-man translated inwards did not preserve illusory contours, and were recognized on every trial.

While there were no significant differences between subjects, the trained subject outperformed the control in the nonlinear cases, and underperformed in the colinear case. It is as though the trained subject learned that colinearity is a salient feature, and that two images with the same colinearities are roughly the same, while images with no colinearity are more easily distinguishable. Gestalt principles of good completion hold, however, that continuity may exist over spectral change of constant slope (Masuda-Katsuse & Kawahara, 1999). Thus, it is entirely possible that the system itself, due to a left to right sweeping across the image, may support this salience of linearity, and the differences in subjects are simply random. That is to say, it may be overly optimistic to read this into the data. Testing this, however, would be trivial, because systems that use a modulated sinusoidal audio signal at each point rather than a binaural sweeping will not have gestalt continuity in the auditory domain, but it will still be present in the visual domain.

One reason to be hopeful, however, for the presence of learning in the subject is that the statistical properties of natural images support the notion of colinear salience. Work in the analysis of images shows that the probability of edge co-occurrence is very high along linear trajectories from a starting point, making a general "bipole" shape (Geisler, Perry, Super, & Gallogly, 2001). It is entirely possible that over the course of training to use the system to navigate, identify objects and manipulate the world in various ways, the trained subject implicitly learned statistical properties of the input set. A larger data set could help identify whether this effect is robust, as could training on different sensory-substitution systems.

Suggestions for Further Work

This study gives hope to the notion that sensory substitution may provide a useful technique to study aspects of vision that may not be assessable by other mechanisms. Fruitful further research directions include, perhaps most importantly, the development of a useful substitution system that can provide the largest degree of contingencies between the simulated sense. A space-variant system with a fast acting periphery is well-suited to natural biological vision, and is a clear next step from the current set of devices. Additionally, work involving the brain and the regions that are used to decode sensory substitution information must be further explored. Current studies have demonstrated the effects of TMS and plasticity, and compared them across both sighted as well as early and late-blind subjects, but much further research needs to be done (Renier, Collignon, et al., 2005; Collignon, Lassonde, Lepore, Bastien, & Veraart, 2006). For example, to what extent do different systems mapping the same substituted sense (vision, in this case) lead to the same types of cortical activation? Furthermore, is it possible to build functional models of these effects in a way that can help clarify mechanisms of plasticity in the human brain? These and further questions are important topics for further research regarding sensory substitution on the whole, but herein an argument has been made for sensory-substitution systems to help assess basic psychological issues. Even this short study has led to potentially interesting result, and it is hoped that such a proof-of-concept will inspire further research along similar methodological lines.

References

Bach-y-Rita, P. (2004). Tactile sensory substitution studies. *Annals of the New York Academy of Sciences*, *1013*.

Capelle, C., Trullemans, C., Arno, P., & Veraart, C. (1998). A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Trans Biomed Eng.*, *45*.

Collignon, O., Lassonde, M., Lepore, F., Bastien, D., & Veraart, C. (2006). Functional cerebral reoganization for auditory spatial processing and auditory substitution of vision in early blind subjects. *Cerebral Cortex*, *10*.

Geisler, W., Perry, J., Super, B., & Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*.

Lee, T. S. (2002). The nature of illusory contour computation. *Neuron*, *33*(5).

Lenay, C., Gapenne, O., Hanneton, S., Marque, C., & Genouelle, C. (2003). In touch for knowing. In (chap. Sensory Substitution, Limits and Perspectives). John Benjamins Publishers.

Masuda-Katsuse, I., & Kawahara, H. (1999). Dynamic sound stream formation based on continuity of spectral change. *Speech Communication*, *27*.

Meijer, P. (1992). Beyond sensory substitution - learning the sixth sense. *IEEE Trans Biomed Eng.*, *39*.

Meijer, P. (2006). *Seeing with sound.* ([Online at http://www.seeingwithsound.com; accessed 29-January-2006])

O'Regan, J. K., Myin, E., & Noë, A. (2005). Skill, corporality and alerting capacity in an account of sensory consciousness. *Progress in Brain Research.*

Renier, L., Collignon, O., Poirier, C., Tranduy, D., Vanlierde, A., Bol, A., et al. (2005). Cross-modal activation of visual cortex during depth perception using auditory substitution of vision. *NeuroImage*, *26*.

Renier, L., Laloyaux, C., Collignon, O., Tranduy, D., Vanlierde, A., Breuyer, R., et al. (2005). The Ponzo illusion with auditory substitution of vision in sighted and early-blind subjects.

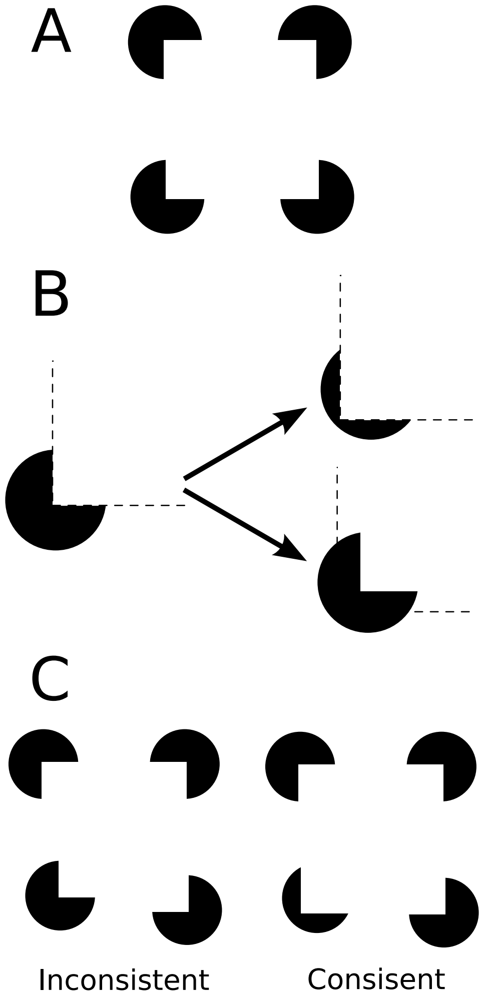
Sampaio, E., Maris., S., & Bach-y-Rita, P. (2001). Brain plasticity: 'visual' acuity of blind persons via the tongue. *Brain Research*, *908*.

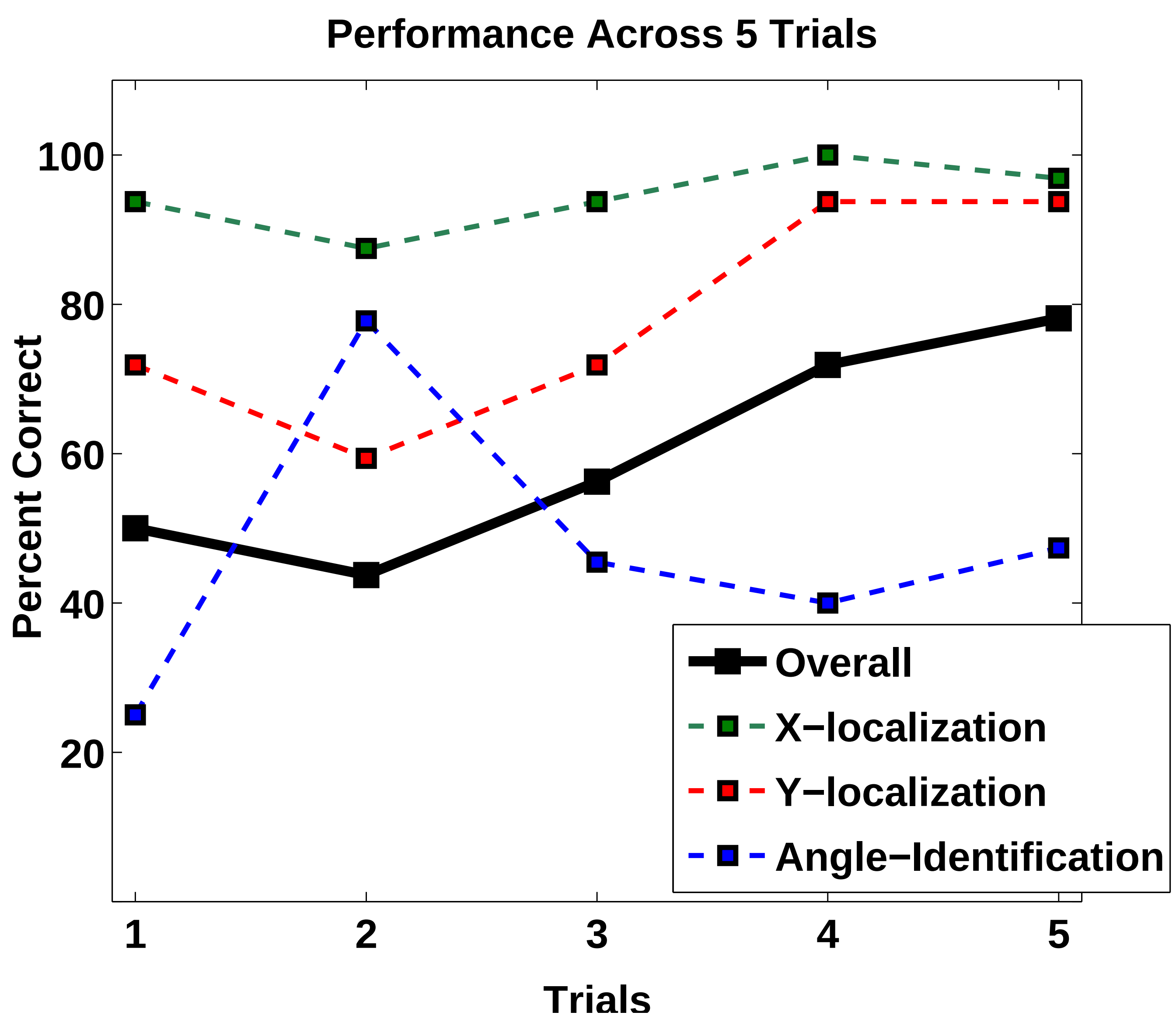
Figure Captions

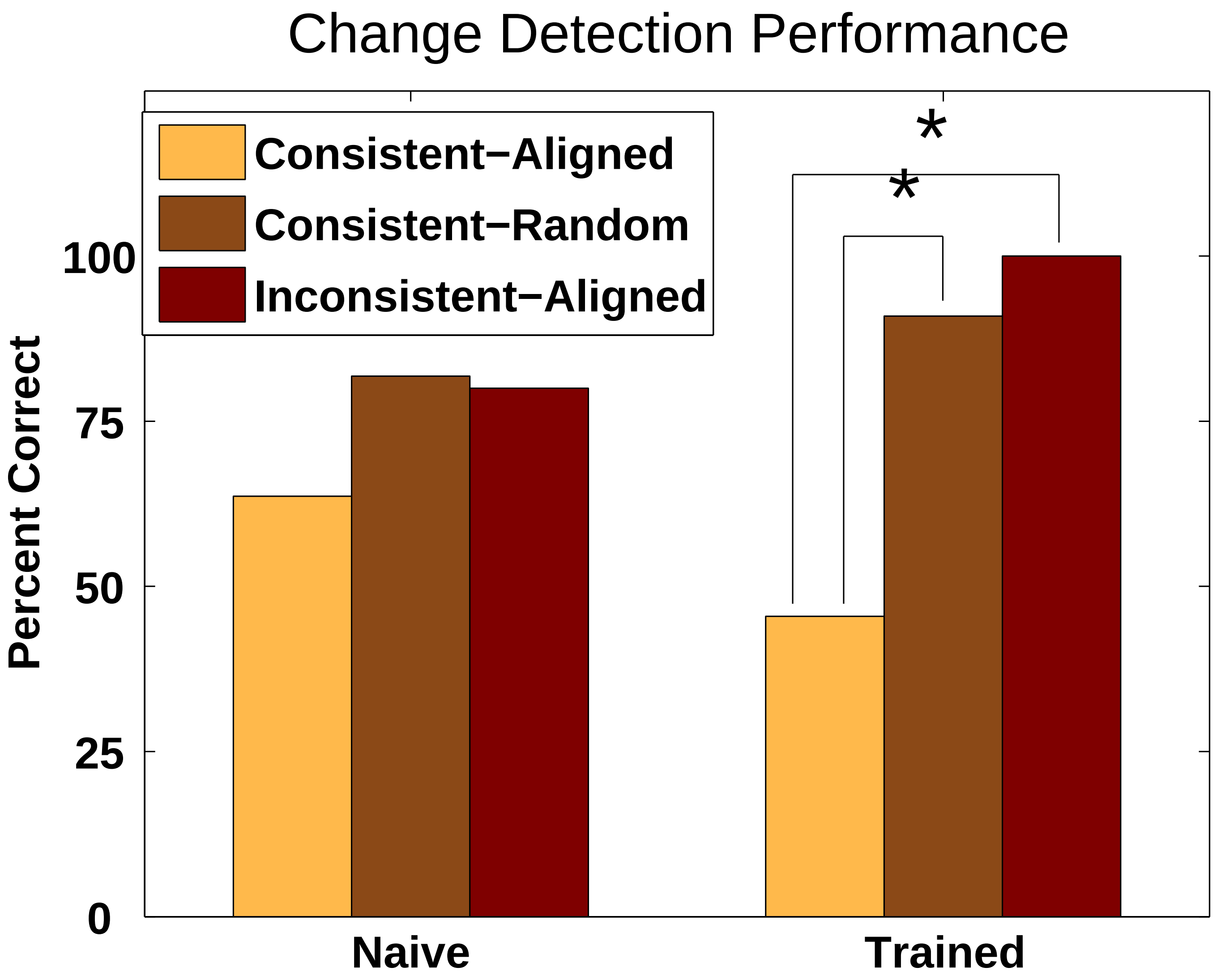
Figure 1. A. The general Kanizsa Square without any modifications. B. The two changes that are made during the testing without feedback phase of the experiment, consistent, in which the circle is subsumed under the square, and inconsistent, in which the pac-man occludes the square. C. The consistent and inconsistent shapes presented in aligned form. The images are also presented with random rotation of each pac-man during testing.

Figure 2. During training, the subject learned to identify and locate a shape on a grid. The black line represents the performance on identification and location, while the blue line represents the angle identification within 10 degrees of rotation. This data represents performance upon introduction of the rotational alteration. Prior to this, the subject received five training sets containing of the same task sans rotated squares.

Figure 3. Naive subject performance in a same-different task plotted against trained subject performance. A star indicates a significant difference ($p < 0.5$). Recall that "aligned" means that the pac-man shapes initially form a Kanizsa Square before being perturbed, while "random" signifies that each pac-man was initially rotated. In the "inconsistent" case, the change is a translation of a pac-man figure, while in the "consistent" case, the pac-man's edges do not shift, but the circular component does shift. For further clarification, refer to Figure 1.


A
B
C
Inconsistent
Consisent


Performance Across 5 Trials
Percent Correct
100
80
60
40
20
1
2
3
4
5
Trials
Overall
X−localization
Y−localization
Angle−Identification


Change Detection Performance
Consistent−Aligned
Consistent−Random
Inconsistent−Aligned
Percent Correct
100
75
50
25
0
*
*
Naive
Trained