tvst

Article

Perceptual Learning of Object Recognition in Simulated Retinal Implant Perception – The Effect of Video Training

Lihui Wang[1–4], Nico Marek[3], Johannes Steffen[5], and Stefan Pollmann[3,4,6]

[1] Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Key Laboratory of Psychotic Disorder, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3] Department of Psychology, Otto-von-Guericke University Magdeburg, Magdeburg, Germany
[4] Center for Behavioral Brain Sciences, Otto-von-Guericke University Magdeburg, Magdeburg, Germany
[5] Department of Simulation and Graphics, Otto-von-Guericke University Magdeburg, Magdeburg, Germany
[6] Beijing Key Laboratory of Learning and Cognition and School of Psychology, Capital Normal University, Beijing, China

Correspondence: Lihui Wang, Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, 1954 Huashan Road, 200030 Shanghai, China.
e-mail: lihui.wang@sjtu.edu.cn

Received: February 18, 2021
Accepted: July 22, 2021
Published: October 18, 2021

Keywords: retinal implants; perceptual learning; object recognition; video training; generalization

Citation: Wang L, Marek N, Steffen J, Pollmann S. Perceptual learning of object recognition in simulated retinal implant perception – The effect of video training. Transl Vis Sci Technol. 2021;10(12):22, https://doi.org/10.1167/tvst.10.12.22

Purpose: Retinal implants (RIs) provide new vision for patients suffering from photoreceptor degeneration in the retina. The limited vision gained by RI, however, leaves room for improvement by training regimes.

Methods: Two groups of normal-sighted participants were respectively trained with videos or still images of daily objects in a labeling task. Object appearance was simulated to resemble RI perception. In Experiment 1, the training effect was measured as the change in performance during the training, and the same labeling task was conducted after 1 week to test the retention. In Experiment 2 with a different pool of participants, a reverse labeling task was included before (pre-test) and after the training (post-test) to show if the training effect could be generalized into a different task context.

Results: Both groups showed improved object recognition through training that was maintained for a week, and the video group showed better improvement (Experiment 1). Both groups showed improved object recognition in a different task that was maintained for a week, but the video group did not show better retention than the image group (Experiment 2).

Conclusions: Training with video materials leads to more improvement than training with still images in simulated RI perception, but this better improvement was specific to the trained task.

Translational Relevance: We recommend videos as better training materials than still images for patients with RIs to improve object recognition when the task-goal is highly specific. We also propose here that achieving highly specific training goals runs the risk of limiting the generalization of the training effects.

Introduction

Photoreceptor degeneration in the retina characterizes eye diseases, such as retinitis pigmentosa (RP) and age-related macular degeneration (AMD), and often leads to blindness.[1] Although photoreceptor loss is irreversible and thus cannot be effectively treated, visual function can still be restored by appropriate electrical stimulation to the remaining intact visual pathways.[2] One of such visual prostheses is the retinal implant (RI), a photoelectric device which stimulates the remaining neurons in the retina to restore residual vision.[1,3–5]

Clinical studies have shown that, after the implantation, patients' with RIs visual performance can be improved from implant off to implant on in a number of tasks, such as light perception, motion detection, and object localization.[6–8] In contrast to the improvement in the basic visual functions, the restoration of more complex vision, such as shape discrimination and object recognition, was limited and variable.[9,10] For

instance, patients' performance in counting and localizing tableware significantly improved along with the use of the RI, whereas their performance in recognizing the identity of the same tableware barely improved.[8] The limited visual ability of object recognition in patients with RI contrasts to the usually rather high plasticity of high-level, compared to low-level vision.[11] This asks for a training regime through which object recognition can be improved and flexibly generalized.

In a recent study on simulated RI vision, we developed a training paradigm where object recognition was significantly improved and transferred across different contexts.[12] Specifically, we simulated object views mimicking the limited vision of patients with a subretinal implant[13,14] and used these simulated object images as training materials for normal-sighted participants. Although the simulation was based on physiological principles of retinal vision and technical features of RIs, it surely cannot be taken as completely equal to artificial visual perception. However, the simulation not only reduces testing burden for patients during the training paradigm development, but also enables investigators to conduct well-controlled and replicable experiments.

By training participants in a labeling task where a simulated object image was presented and participants had to choose the correct label among other alternative labels, we found that the recognition accuracy increased approximately 18% after a short period of training (approximately 1 hour). This improved object recognition persisted over a week and generalized to a different task context where one label was presented with different object images and participants had to choose the correct object to match the label, suggesting the potentiality of persistent and flexible object recognition in patients with RIs.[12]

Although our recent study showed that object recognition in patients with RIs can be improved and generalized by training, it remains to be answered if the trained RI vision can be further optimized. Here, we address this issue by including videos as the training material. It is proposed that the ambiguous information of an object activates multiple perceptual hypotheses that compete against each other, and object recognition is implemented by resolving the competition.[15] For instance, the object shown in Figure 1B may be recognized as a cup (handle on the right), a spoon (handle on the right), a pear (stalk on the right), or a stapler (head on the left) at first glance. To achieve a correct recognition of the spoon, the assumptions of it being a cup, a pear, or a stapler have to be excluded. To this end, relative to a still image, a video provides ecological and coherent information of an object and hence facilitates the resolution of the competitive perceptual

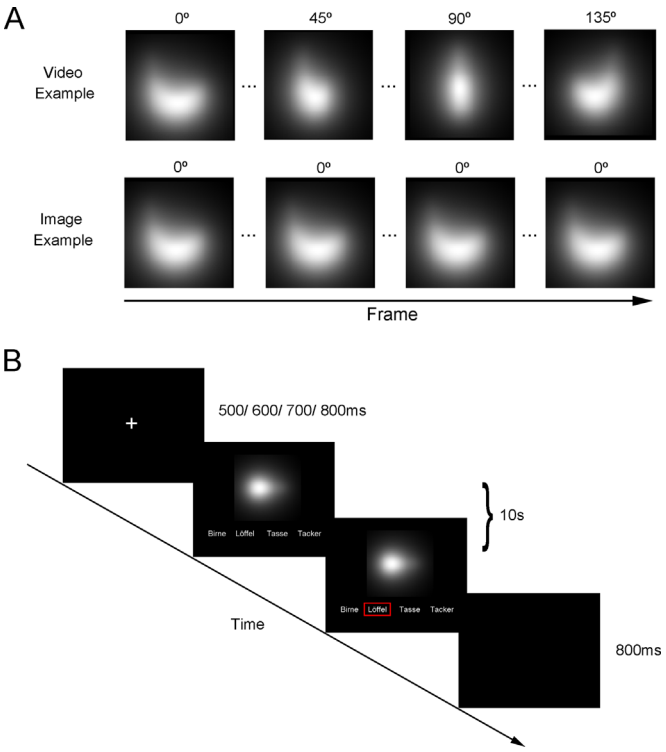

Figure 1. (**A**) A video example (*upper panel*) and an image example (*lower panel*) of a simulated banana. For space restrictions, only the 0, 45, 90, and 135 degree of the rotation are illustrated in the video example. (**B**) An example trial of the task. Participants were asked to choose the correct label for the simulated video/ picture by mouse click (a spoon is shown in this example). The correct label was marked with a red box after the mouse click irrespective of the correctness of the response.

hypotheses of the object. It has already been shown in previous studies that motion facilitates the recognition of blurred objects.[16] We therefore expected that training with videos would lead to better object recognition than training with still images.

An important goal in perceptual training is to achieve generalization. In our previous study,[12] we have shown that training effects by the labeling task can be generalized into untrained viewpoints of the same object, and into another task context. Here, we also tested if training with video could lead to the same, or even better generalizations, than training with still images. In Experiment 1, we first trained two groups with the labeling task, with one group trained with video materials and the other group with still images. We investigated if and to which extent training with videos leads to better improvement than training with still images. In Experiment 2, we trained another two groups in the labeling task but tested the training effects in the reverse labeling task and in different viewpoints of the trained objects to investigate the transfer of learning.

Experiment 1

Methods

Participants

Thirty-two healthy students participated in the experiment, with 16 of them (8 and 8, age = 20–38 years old) randomly assigned to the video group, and the other 16 (8 and 8, age = 21–34 years old) assigned to the image group. All participants had normal or corrected-to-normal vision, and all of them were German native speakers. This experiment was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics review board. A written consent form was obtained from each of the participants prior to the experiment. None of them had been exposed to the simulated videos or pictures before the experiment.

Stimuli and Design

In a first step, the 3-D computer graphics toolset Blender version 2.79 (https://www.blender.org/) was used to create 8 video clips using 8 3-D models (banana, cup, hourglass, pear, rose, scissors, spoon, and stapler) from the Free3D database (https://free3d.com/). Each original clip had a resolution of 1920*1080 pixels and the background color was kept constant black (Hex code: #000000). To ensure equal lighting of all 3-D objects, an invisible light source was placed above all models at the same coordinates. To avoid pixel edges in the original (not simulated) video clips, an anti-aliasing Mitchell-Netravali reconstruction filter was used during the rendering process. The frame rate was kept constant at 30 frames per second and each video lasted 10 seconds with a constant rotating speed. Eight clips were made for each of the 8 objects where a 180-degree rotation was recorded. The 8 videos were generated using the same 360-degree rotation of a certain object and differed only in the starting point and the corresponding end point of the rotation. Videos could start at 0/ 22.5/ 45/ 67.5/ 90/ 112.5/ 135/ and 157.5 degrees rotation, and the rotation hence ended at 180/ 202.5/ 225/ 247.5/ 270/ 292.5/ 315/ and 337.5 degrees of rotation.

In a second step, the 64 videos (8 objects with 8 starting points each) were then simulated with a custom implementation of the pulse2percept software.[13] Each picture had a resolution of 1369 stimulating electrodes in an area covering a visual field of 8 degrees visual angle. The technical parameters were chosen to reflect the subretinal Alpha IMS cochlear implant (Retina Implant AG, Tübingen).[17] A very low likelihood of axonal stimulation was chosen ($\lambda = 0.1$), as would be appropriate for a subretinal implant whose electrodes do not reach the axonal layer on the retinal surface. See Supplementary Materials for the full set of the simulated videos.

For the still image group, the first frame of each video starting point was taken. Therefore, there were 64 simulated images (8 objects with 8 viewpoints each), which had the same size on the screen as the video (Fig. 1A). The same technical parameters were used for the video and still image simulation. The still image in each trial was presented for 10 seconds or until a response was made.

Procedures

Participants were tested in a sound-attenuated and dimly lighted room. They were seated in front of a monitor screen with their head positioned on a chin-rest, and were required to fixate at the central cross throughout each trial. The eye-to-monitor distance was fixed at 65 cm.

Each trial began with a white fixation cross at the center of a black screen for a duration randomly selected from 500/ 600/ 700/ 800 ms (see Fig. 1B). Then the task frame, which contained the video or picture, was presented until the time limit (10 seconds to allow a full rotation) was reached. Four labels, with one of which being the correct label of the presented video/picture, were presented below the video/picture. Participants were asked to choose the correct label for the video/picture by clicking the left button of the mouse. In cases where more than one click was given, only the first mouse click was taken into account as the response. As a feedback, the correct label was marked with a red box immediately after the mouse click irrespective of the correctness of the mouse click. As in our previous study,[12] the feedback was presented in each trial in both the training and the post-test. The inter-trial-interval was a blank screen of 800 ms. Participants were asked to respond as fast and accurately as possible.

For both groups, there were 64 trials (8 object videos * 8 starting viewpoints for the video group, respectively 8 object pictures * 8 viewpoints for the image group) in each of the 3 blocks. In each block, the 64 trials were presented in a pseudorandom order in the way that 8 objects from 4 different (starting) viewpoints were followed by the same 8 objects from 4 additional different (starting) viewpoints. This arrangement was to follow the design in our previous study[12] so that the results observed here can be attributed to the inclusion of the video materials. Within each block, the distracting labels in each trial was randomly selected from other labels of the eight objects. There was a 1 minute break after each block.

Table 1. The Performances of Starting Point During Training (Mean Accuracy in % With Standard Errors) in Recognizing the Eight Objects for the Video Group and the Image Group in Experiment 1 (Upper Rows) and the Initial Performances From the Pre-Test in Experiment 2 (Lower Rows)

Exp.	Group	Banana	Cup	Hourglass	Pear	Rose	Scissors	Spoon	Stapler
Exp. 1	Video	69.6 (7.8)	61.6 (5.1)	53.6 (6.2)	81.3 (4.4)	46.4 (5.6)	58.9 (5.2)	72.3 (4.6)	40.2 (4.8)
	Image	65.6 (4.8)	50.8 (3.9)	78.1 (4.9)	75.0 (3.4)	60.2 (3.8)	57.0 (4.4)	74.2 (2.7)	53.9 (4.1)
Exp. 2	Video	51.4 (4.5)	52.1 (5.6)	50.0 (5.2)	54.9 (5.8)	36.1 (3.5)	24.3 (3.1)	66.0 (4.0)	31.3 (3.4)
	Image	47.9 (3.4)	56.3 (5.9)	45.8 (4.7)	50.7 (5.2)	32.6 (3.4)	38.2 (4.3)	55.6 (4.4)	18.1 (3.5)

Exp., experiment.

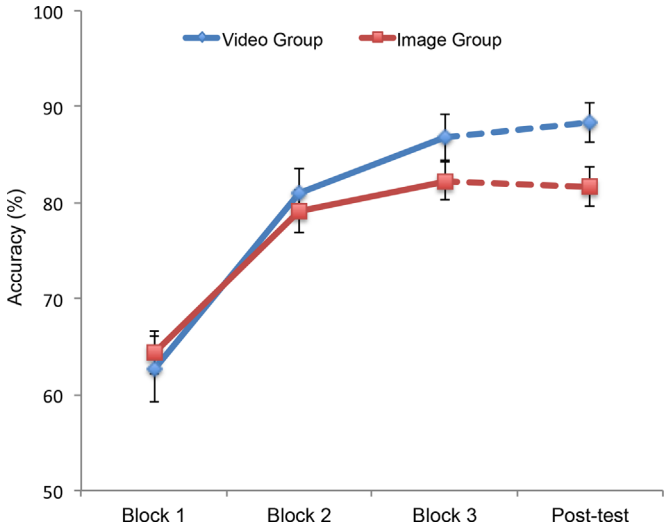

Figure 2. Accuracies with standard errors are shown as a function of block order and group in Experiment 1. The dashed line indicates a week between block 3 and the post-test.

An additional block was presented as a post-test, which included the same 64 trials (8 objects * 8 starting viewpoints/viewpoints) as in each block of the training. The post-test took place 1 week after the training.

Statistical Data Analysis

For each participant, the accuracy was calculated as the percentage of trials with a correct response. To show the starting point of the training in recognizing the simulated videos/pictures, the accuracy in block 1 where each of the simulated videos/pictures was presented for the first time, was calculated for each object (Table 1).

The mean accuracy in each block for each group is shown in Figure 2. To test whether there was an improvement in object recognition during training, and whether the training effect differed between the two groups, a mixed logit model[18] was carried out to model the binary choice (correct versus incorrect response) in each trial using the *glmer* function in *R*.[19] Specifically, group (video training versus image training), block order (1, 2, vs. 3), and the interaction between group and block order were included as fixed factors, whereas the individual subjects and the simulated objects were included as random factors. Following a significant interaction, post hoc analyses were performed with the package *phia* in *R*.[20] The variances of different objects were included as random effects because our research question focused on which training regime led to better training effects. The random effects that contributed by different objects in each linear mixed model are shown in Table 2.

To show if the improvement persisted 1 week after the training, we calculated the retention score by subtracting the accuracy of block 3 from the accuracy of the post-test. Given that the persistence of the training effect may be based on a null effect (i.e. no difference between block 3 and the post-test) whereas a null effect cannot be confirmed by either logit mixed modelling or paired *t*-test, we calculated the Bayes Factor (BF)[21,22] using JASP[22] to quantify the extent to which the null hypothesis was supported. Here, the BF describes the relative probability of the data under the null hypothesis that the accuracy in the post-test was equal or higher than the accuracy in the last training block, relative to the alternative hypothesis that the accuracy in the post-test was lower than the accuracy in the last training block. Per convention, a BF >3 and <10 is taken as moderate evidence for the tested hypothesis while a BF >1 and <3 yields anecdotal evidence.[23]

Next, for both groups, we calculated the retention score by subtracting the accuracy of block 3 from the accuracy of the post-test. An independent *t*-test was performed on the retention scores to test if one group had better retention after training than the other group.

Results

Starting Point of Training in Recognizing Each Object

As shown in Table 1, the starting point (i.e. the accuracy for block 1), in recognizing each of the 8 objects was above the theoretical chance-level (25%), all $p < 0.001$, for both groups (one-sided, with Bonferroni corrections for the 8 comparisons). In addition,

Table 2. The Variance Contributed by the Random Effect of Object Type, and the Individual Random Effect of Each Object that Estimated From the Mixed Logit Model in Each of the Two Experiments

Exp.	Session	Variance	Banana	Cup	Hourglass	Pear	Rose	Scissors	Spoon	Stapler
Exp. 1	Training	0.29	0.90	−0.40	0.25	−0.21	−0.19	−0.26	0.62	−0.77
Exp. 2	Training	0.26	0.58	−0.19	0.72	0.47	−0.32	−0.37	−0.19	−0.78
	Pre-test vs. post-test 1	0.33	0.30	0.67	0.07	0.41	−0.31	−0.80	0.54	−0.89

The value of the individual random effect indicates the random effect of each object relative to the averaged random effect across all objects after the fixed effects (i.e., group and session) have been excluded.

Exp., experiment.

the independent *t*-tests showed that the starting accuracies for each of the 8 objects did not differ between the two groups, all $p > 0.1$ (two-sided, with Bonferroni corrections for the 8 comparisons).

Training Effects

The mixed logit model revealed a significant effect of block order, $z = 2.18$, $p = 0.029$, and a significant interaction between group and block order, $z = 3.17$, $p = 0.002$, whereas the effect of group did not reach significance, $z = 1.34$, $p = 0.180$. Further post hoc analyses on the interaction showed the effect of block (improved performance from block 1 to block 2, and from block 2 to block 3) for both the video group, $\chi^2(2) = 188.19$, $p < 0.001$ (mean accuracy = 62.7%, 81.1%, and 86.8%), and the image group, $\chi^2(2) = 104.82$, $p < 0.001$ (mean accuracy = 64.4%, 79.1%, and 82.2%). Moreover, the video group outperformed the image group only in the third block, $\chi^2(1) = 3.80$, $p = 0.026$ (one-sided), but not in the first or the second block, both $\chi^2 < 1$.

Retention

For the video group, the BF analysis comparing the accuracy in block 3 and the accuracy in the post-test showed that BF = 6.803, indicating that the null hypothesis ("the accuracy in the post-test was equal to or higher than the accuracy in block 3") was 6.803 times more likely to be true than the alternative hypothesis ("the accuracy in the post-test was lower than the accuracy in block 3"). Thus, we have moderate evidence that training with the simulated videos improved object recognition, and this improved performance can last at least for 1 week.

For the image group, the BF analysis comparing the accuracy in block 3 and the accuracy in the post-test showed that BF = 2.778, indicating that the null hypothesis ("the accuracy in the post-test was equal to or higher than the accuracy in the last training block") was 2.778 times more likely to be true than the alternative hypothesis ("the accuracy in the post-test was lower than the accuracy in block 3"). These results suggested that training with still images also improved object recognition and lasted for 1 week, but the improved performance and the persistence seemed not as robust as the video group. This hypothesis would be further tested by between-groups comparisons below.

The independent *t*-test showed that the retention scores (accuracy at post-test - accuracy at block 3) were comparable between the two groups (1.6% in the video group versus −0.6% in the image group), $t < 1$. However, the accuracy of the post-test was higher in the video group (88.4%) than in the image group (81.6%), $t(30) = 2.33$, $p = 0.027$. These results suggested that training with videos led to better and more robust performance than training with still images.

Discussion

In Experiment 1, the results from the 2 groups showed that training with still images and training with videos both led to improved object recognition, and the improved recognition persisted at least for a week. The improved object recognition through training in the image group (block 3 vs. block 1: 17.9%) was consistent and comparable with the improved performance shown in our recent studies where still images with simulated RI-vision were used as the training materials (18.4% in Wang et al.[12]; 16.0% in Nath et al. unpublished). By contrast, the improved recognition performance here in the video group (block 3 vs. block 1: 24.1%) was larger than the improved performance in the image group. Importantly, given that the starting point of training in the two groups was comparable, the larger training effect in the video group cannot be due to easier object recognition in videos but rather that videos were better suited for boosting the perceptual learning in RI vision.

It has been shown in clinical trials that, without training, object recognition in patients with RIs was limited and variable.[8,17] For instance, in an early study, only one of the three patients with RIs was able to name daily objects, such as tableware and fruits.[17]

Although the object naming in a real-world situation cannot be taken as equal to the labeling task in the present study, the poor real-world performance asks for a training regime to improve the newly restored vision of patients with RIs. In a recent study, object recognition was barely restored with the use of the implants, although object counting and localization was restored to a great extent.[8] As a training regime, the present results and our previous study[12] consistently showed above-chance recognition performance in the first block of training, and the recognition performance can be further improved and maintained after a short training process. These results suggest that the limited and variable object recognition in RI vision can be improved with lasting effects by optimized perceptual training.

An important finding in Experiment 1 was that video materials led to a stronger training effect than still images, which manifested at the late stage of the training process. When no prior knowledge of an object can be obtained (e.g. at the beginning of the present experiments), the information accumulation is subject to random fluctuations.[24,25] Therefore, although the different viewpoints of an object were shown in a video, the information afforded by a video could be as ambiguous as the information afforded by a still image, resulting in comparable performances between the two groups in the first block of the training process. As the training proceeded, the coherence of the information in the video boosted the predictive processing[26] of the object, which facilitated the accumulated information favoring one perceptual hypothesis over the other alternatives.[15] An alternative account for the stronger training effect in the video group might be that videos provided more information than still images, as a video affords all viewpoints along a specific axis in a full rotation. To this end, the training effects could have been simply determined by the available number of viewpoints of a specific object. However, if this were true, the video group should have shown better performance than the image group from the beginning of the training. This prediction could be rejected based on the observation that the stronger training effect only emerged as the training proceeded. Therefore, the improved training effect appears to rely more on the coherence rather than simply on the number of the viewpoints.

Experiment 2

Whereas Experiment 1 showed training effects for both still images and videos as training stimuli, Experiment 1 did not contain a common test of training efficiency for both kinds of training. In Experiment 2, we added such a common test as common pre-test baseline and post-test measure for both still image and video training. Using still images as training materials, our previous study[12] showed that the training effect in the labeling task transferred to the reverse labeling task. Therefore, we chose the reverse labeling task as the common test for still image and video training in Experiment 2. We expected a replication of successful transfer for training with still images and we wanted to test to which extent training with videos could achieve the same transfer. To answer this question, in Experiment 2, we included the reverse labeling task both before (i.e. pre-test) and after the training session (i.e. post-test) and compared the performance in the reversed labeling task between pre-test and post-test.

Methods

Participants

The sample size of Experiment 2 was determined based on the effect size observed in Experiment 1. Calculated using G*Power,[27] the effect size of the group difference was Cohen's $d = 0.824$ given the between-group statistics $t(30) = 2.23$ and $n = 16$ per group. Sample size estimation showed that 19 participants were required for each group, given alpha = 0.05 and beta = 0.2. Forty participants were recruited, with half of them randomly assigned to the video group and the other half to the image group. None of them participated in Experiment 1. Due to incomplete data of 4 participants, the final data analysis was based on 18 participants (video group = 14 and 4, 19 to 25 years old; image group = 16 and 2, 18 to 27 years old) in each group.

Stimuli and Design

A set of the same eight objects was simulated along a reference system. Specifically, each object in the video rotated along the z axis of the 3-D reference system in Experiment 1 and each frame in the video clip represented a 2-D image on the x-z plane. In Experiment 2, each 3-D object was first rotated along the y axis for 45 degrees to get a new viewpoint model. Then the y axis was fixed, and the object rotated along the z axis. Each frame in the video clip represented a 2-D image on the x-z plane. As in Experiment 1, 8 viewpoints were extracted from the video for each object, rendering 8 objects * 8 viewpoints still images (see Supplementary Materials for the rotated images). These 64 still images were used as stimuli in the sessions of pre-test, post-test 1, and post-test 2. No video was used in these test sessions. During the training session, the video

and the image stimuli were the same as the stimuli in Experiment 1. Therefore, in Experiment 2, the 8 objects had different viewpoints during the training session compared to the test sessions.

Procedures

The procedures in Experiment 2 were the same as the procedures in Experiment 1 with the following exceptions. A pre-test and a post-test 1 session were conducted before and after the training session, respectively. In each of these 2 sessions, there were 64 trials of the reverse labeling task where 4 images were presented below a label and participants had to choose the correct image to match the label by mouse click.[12] This task frame was presented up to 3 seconds until the first mouse click or the 3 second time limit was reached. After the task frame, a feedback frame was presented for 2 seconds. The feedback frame was the same as the task frame except that the correct image was marked by a red box. In each of the two test sessions, each of the 8*8 images appeared as the target image only once. The distracting images in each trial were randomly selected from the non-target images, and each of the 8*8 images appeared as one of the distracting images with equal probability. The pre-test and post-test 1 sessions were conducted in the same day as the training session. The same test session was conducted one week after the training session as post-test 2.

Statistical Data Analysis

The initial performance in recognizing the simulated objects was calculated in terms of the accuracy in the pre-test (see Table 1). It should be noted that the initial performance in pre-test was calculated to show the starting point for the training/generalization effect. In Experiment 1, the starting point was calculated as the accuracy in block 1 because the training effect was measured as the difference between block 3 and block 1. In Experiment 2, the initial performance was calculated as the accuracy in pre-test because the generalization effect was measured as the difference between post-test 1 and pre-test.

The mean accuracy in each block/session for each group is shown in Figure 3. To replicate the effect during training in Experiment 1, a mixed logit model as in Experiment 1 was carried out to model the binary choice (correct versus incorrect response) in each trial during the training session. As in Experiment 1, group (video training versus image training), block order (1, 2, vs. 3), and the interaction between group and block order were included as fixed factors while the individual subjects and the simulated objects were included as random factors.

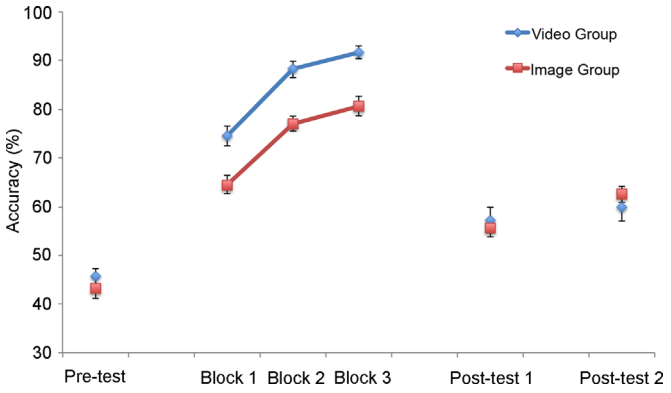

Figure 3. Accuracies with standard errors are shown as a function of session and group in Experiment 2. Post-test 2 was conducted a week after the other sessions.

To test whether the performance in a different task can be improved after training, a mixed logit model was performed with group (video training versus image training), test session (pre-test versus post-test 1), and the interaction between group and test session being included as fixed factors, whereas the individual subjects and the simulated objects being included as random factors.

To show if the persistence of the improvement after training, as in Experiment 1, the accuracy in post-test 1 and the accuracy in post-test 2 were compared using BF analysis for each of the two groups. The retention score of each group was obtained by subtracting the accuracy in post-test 1 from the accuracy in post-test 2. An independent *t*-test was then conducted to compare the retention scores of the two groups.

Results

Initial Performance in Recognizing Each Object

For the video group, the initial performances in recognizing 6 (banana, cup, hourglass, pear, rose, and spoon) of the 8 objects were above the theoretical chance-level (25%), all $p < 0.05$, (one-sided, with Bonferroni corrections for the 8 comparisons), whereas the initial performance in recognizing the scissors and stapler did not exceed the theoretical chance-level, both $p > 0.1$. For the image group, the initial performance in recognizing 6 (banana, cup, hourglass, pear, and scissors) of 8 objects were above the theoretical chance-level (25%), all $p < 0.05$ (one-sided, with Bonferroni corrections for the 8 comparisons), whereas the initial performance in recognizing the rose and stapler did not exceed the theoretical chance-level, both $p > 0.1$. In addition, the independent *t*-tests showed that the initial accuracies for each of the 8 objects did not differ

between the 2 groups, all $p > 0.09$ (two-sided, with Bonferroni corrections for the 8 comparisons).

Training Effects

During training, the mixed logit model revealed a significant effect of Block order, $z = 7.36$, $p < 0.001$, and a significant interaction between group and block order, $z = 3.45$, $p < 0.001$, whereas the effect of group did not reach significance, $z = 1.17$, $p = 0.244$. Further post hoc analyses on the interaction showed the effect of block (improved performance from block 1 to block 2, and from block 2 to block 3) for both the video group, $\chi^2(2) = 152.20$, $p < 0.001$ (mean accuracy = 74.6%, 88.2%, and 91.7%), and the image group, $\chi^2(2) = 90.33$, $p < 0.001$ (mean accuracy = 64.5%, 77.1%, and 80.6%). Moreover, the video group outperformed the image group in all of the 3 blocks, block 1: $\chi^2(1) = 11.51$, $p < 0.001$, block 2: $\chi^2(1) = 24.99$, $p < 0.001$, and block 3: $\chi^2(1) = 33.53$, $p < 0.001$. The significant interaction between group and block order shown by the mixed logit model, together with the increasing χ^2 values shown by the post hoc analysis, indicated that the video group improved more than the image group as the training proceeded (accuracy difference between groups: 10.1% at block 1, 11.1% at block 2, and 11.1% at block 3). These results are consistent with the results in Experiment 1 that the training effect increased as a function of the block order.

The mixed logit model that focused on pre-test (mean accuracy = 44.4%) and post-test 1 (mean accuracy = 56.4%) showed only a significant effect of test session, $z = 3.20$, $p = 0.001$, whereas the group effect and the interaction did not reach significance, both $|z| < 1$. These results suggested that the object recognition in the reverse labeling task was improved after training for both groups, and the improvement did not differ between the two groups.

Given the consistent patterns during training in the two experiments, data of block 3 in both experiments were pooled to show the effect size of the achieved training effect, such that it can be helpful to evaluate how much the recognition accuracy can be improved after video training more than after image training. Here, only the data in block 3 but not data from all 3 blocks were pooled for the following reasons: (1) block 3 was the final block to show the achieved effect size of the training block; and (2) although post-test can be used to show the achieved effect size, the task types and viewpoints in the post-tests were different between the two experiments and it is thus inappropriate to collapse the data. The mixed logit model on the performance in block 3 showed that the performance in the video group (89.4%) was better than the performance in the image group (81.4%), $z = 2.61$, $p = 0.009$.

Retention

For the video group, the BF analysis comparing the accuracy in post-test 1 (57.2%) and the accuracy in post-test 2 (59.9%) showed that BF = 9.662, indicating that the null hypothesis ("the accuracy in post-test 2 was equal to or higher than the accuracy in post-test 1") was 9.662 times more likely to be true than the alternative hypothesis ("the accuracy in post-test 2 was lower than the accuracy in post-test 1"). These results suggested that the improved performance by video training can last at least for 1 week.

For the image group, the BF analysis comparing the accuracy in post-test1 (55.6%) and the accuracy in post-test 2 (62.6%) showed that BF = 14.265, indicating that the null hypothesis ("the accuracy in post-test 2 was equal to or higher than the accuracy in post-test 1") was 14.265 times more likely to be true than the alternative hypothesis ("the accuracy in post-test 2 was lower than the accuracy in post-test 1").

The independent t-test shows a trend that the retention score (post-test 2 versus post-test 1) in the image group (7.0%) was higher than the retention scores (2.2%) in the video group, $t(34) = 1.80$, $p = 0.081$.

Discussion

In Experiment 2, we first replicated the findings in Experiment 1 that training with videos led to better improvement in object recognition than training with still images, as shown by the significant interaction between group and block order during the training session. However, in contrast to Experiment 1, the better performance in the video group emerged earlier (i.e. from the first block). This pattern may be due to that the experience of the pre-test - despite different viewpoints - functioned as "training" and helped to build up the prior knowledge of the simulated objects. This prior knowledge could further boost the utilization of the coherent information in the videos, leading to a stronger training effect in the video group.

An alternative account could be that the starting point of training of the two groups was not equal in the labeling task used for training, although they were comparable in the reverse labeling task used in the pre-test. Based on this account, the better performance of the video group in block 1 was due to higher starting point rather than due to stronger training effect. To test this alternative account, the trials in block 1 were sorted into 4 bins according to the trial order (1–16th trials as bin 1, 17–32th trials as bin 2, 33–48th trials as bin 3, and 49–64th trials as bin 4), and the recognition accuracies between the two groups in each bin were compared with BF analysis (Fig. 4). If this alternative account were true, the video group would have outperformed

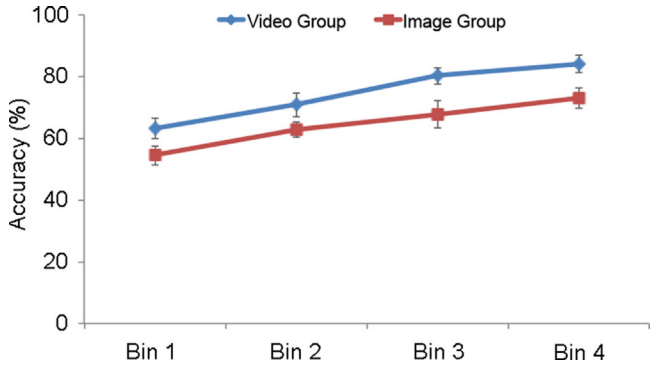

Figure 4. Accuracies with standard errors are shown as a function of the trial bin in block 1 of Experiment 2.

the image group from the very first bin. In violation of this prediction, the video group showed reliably higher accuracy (BF >3) than the image group only in bin 3 and bin 4. Specifically, based on the BF analysis, the hypothesis "the accuracy of the video group was higher than the accuracy of the image group" was 2.65 times more likely to be true than the hypothesis "the accuracy of the video group was equal or lower than the accuracy of the video group" in bin 1, 1.97 times more likely to be true in bin 2, 6.71 times more likely to be true in bin 3, and 9.44 times more likely to be true in bin 4. Therefore, the better performance of the video group in block 1 is unlikely due to a better starting point than for the image group. Moreover, the better performance of the video group that emerged as the training proceeded was also consistent with the pattern in Experiment 1, pointing to the role of information coherence in leading to the improvement.

The performance in post-test 1 showed an overall improvement over the performance in the pre-test and this improvement was maintained 1 week after training. Importantly, given that the viewpoints of the objects and the task context in these test sessions were different from the viewpoints and task context in the training session, these results suggested that perceptual training leads to persistent and flexible improvement in object recognition that transfers to new viewpoints and new task contexts.[12]

The improved performance in the post-test might be simply due to the experience of the pre-test rather than the experience of the training. However, the main purpose of Experiment 2 was to assess differential improvement due to the training tasks measured on a common task that is different from both training tasks. In answering this question, the current results did show slightly better retention of the generalization by image training than by video training to the reverse labeling task in the post-test (see general discussion for the potential mechanism), although the two groups went through the same amount of training trials.

One may note that the initial recognition performance in Experiment 2 did not exceed chance-level for all of the 8 objects, as compared with the all-above-chance starting point in Experiment 1. One potential explanation of the poorer performance is that the viewpoints of the object models that were used to simulate RI vision in Experiment 1 are common in daily life. After the rotation, the viewpoints in Experiment 2 might be less common, and hence more difficult to recognize.

General Discussion

In two experiments, we consistently showed that both training with videos and training with still images led to better object recognition. Importantly, in both experiments, the group difference (performance of the video group versus performance of the image group) increased as a function of block order during training, as shown by the significant interaction between group and block order, suggesting that training with videos led to more improvement than training with still images. These results suggested that video materials can be used to optimize the object recognition of patients with RIs.

Although the persistent and flexible improvement can be achieved by both videos and still images, the video group did not show better generalization than the image group. Instead, the image group even showed slightly better retention of the generalization after 1 week than the video group. This counterintuitive pattern resembles the model "over-fitting" in machine learning where maximizing the predictive power of a model on the trained dataset limits the generalization power of this model on untrained dataset.[28] In the present study, the video materials provided more specific information about a particular object than the image materials.[16] Relative to the still images, the videos not only contained coherent information of a specific object, but also coherent information of the visual noise that went along with the object. This specific information was beneficial for the performance when the tested objects (i.e. post-test in Experiment 1) were the same as the trained objects (i.e. training session in Experiment 1), leading to higher accuracy of the post-test in the video group than in the image group. By contrast, when the tested objects and context changed (post-test in Experiment 2), the object-specific information of the trained object was not beneficial anymore and could even constrain the recognition of the new stimuli. As a result, the video group did not show better performance in the post-test sessions. Therefore, it is

crucial to balance achieving specific training effects and the generalization of the training effects when applying the training regime in clinical interventions. Another factor, however, could be that the task in the common test was more similar to the training task in the image group. It is unknown which group would have shown better generalization if a video-based task had been used as the common test.

In both experiments, the video group and the image group showed comparable starting points in recognizing the simulated objects. This equivalent performance, however, was inconsistent with findings by Pan and Bingham[16] who showed that the performance in recognizing blurred objects improved immediately when the objects moved. Similar findings were also observed for reading texts.[29] It should be noted that in the present study, all simulated objects were presented against a blank background. By contrast, in the study by Pan and Bingham,[16] each moving object was embedded in a natural context to render a meaningful event (i.e. optic flow[30]), providing more prior knowledge about the presented object. Such prior knowledge thus can be helpful to the recognition of the blurred object.

In the present study, we aimed to test if and to which extent training with video could optimize the object recognition in simulated RI vision. We used the same design (i.e. feedback was always presented after a response) and the same task (labelling task and reverse labeling task) as in our previous study[12] so that the results of our present study can be attributed to the inclusion of the video materials. Future experiments with different design and training context are expected to advance the understanding of the training effect and to develop a more ecological training regime. For instance, in addition to the video group and still image group, multiple still images of the same object with different viewpoints could be simultaneously presented for a new group of participants during training. For this third group, multiple viewpoints are available in the same trial as the video training but the object information shown is not as coherent as the video. The inclusion of such a group could further test whether the training effect by video was due to the accumulated information of object viewpoints or to the coherence of an object in the video.

Although the effect size of group difference was small (7% difference in accuracy based on the pooled data in block 3 from the two experiments), this should not be taken to be the ultimate improvement that patients with RIs can achieve through video training in real situations. It should be noted that the training sessions in our study were short (3 blocks of 64 trials). It is yet to be investigated whether the improvement can be further enhanced by optimizing the training regime, for instance, by including longer training sessions or repeated training.

In summary, we extend previous studies on the training regimes for patients with RIs by showing that video materials led to a better and more robust training effect than still images. However, this better training effect was specific to the trained viewpoints and task context. When it came to new viewpoints and new task context, although both materials led to generalization, the video materials were not superior to still images. Currently, we recommend video as superior training material only when the training goal is highly specific. However, future studies with actual patients may investigate if video training is more efficient for object recognition in real-life situations, where objects can be viewed with changing perspectives.

Acknowledgments

The authors thank Markus Donig for assisting in data collection.

Supported by the European Fund for Regional Development (EFRE), ZS/2016/04/78113, Project: Center for Behavioral and Brain Sciences (CBBS) and by the Deutsche Forschungsgemeinschaft (grant PO548/14-2). L.W. is supported by the Shanghai Sailing Program (No. 20YF1422100).

Disclosure: **L. Wang,** None; **N. Marek,** None; **J. Steffen,** None; **S. Pollmann,** None

References

1. Shepherd RK, Shivdasani MN, Nayagam DAX, Williams CE, Blamey PJ. Visual prostheses for the blind. *Trends Biotechnol*. 2013;31(10):562–571.
2. Merabet LB, Rizzo JF, Amedi A, Somers DC, Pascual-Leone A. What blindness can tell us about seeing again: merging neuroplasticity and neuroprostheses. *Nat Rev Neurosci*. 2005;6(1):71–77.
3. Chader GJ, Weiland J, Humayun MS. Artificial vision: Needs, functioning, and testing of a retinal electronic prosthesis. *Prog Brain Res*. 2009;175:317–332.
4. Fine I, Boynton GM. Pulse trains to percepts: the challenge of creating a perceptually intelligible world with sight recovery technologies. *Philos Trans R Soc Lond B Biol Sci*. 2015;370(1677):20140208.
5. Zrenner E. Will retinal implants restore vision? *Science*. 2002;295(5557):1022–1025.

6. Dorn JD, Ahuja AK, Caspi A, da Cruz L, Dagnelie G, Sahel J. Argus II Study Group. The detection of motion by blind subjects with the epiretinal 60-electrode (Argus II) retinal Prosthesis. *JAMA Ophthalmol*. 2013;131(2):183–189.
7. Ho AC, Humayun MS, Dorn JD, et al. Long-Term results from an epiretinal prosthesis to restore sight to the blind. *Ophthalmology*. 2015;122(8):1547–1554.
8. Stingl K, Schippert R, Bartz-Schmidt KU, et al. Interim results of a multicenter trial with the new electronic subretinal implant Alpha AMS in 15 patients blind from inherited retinal degenerations. *Front Neurosci*. 2017;11:445.
9. Stingl K, Bartz-Schmidt KU, Besch D, et al. Artificial vision with wirelessly powered subretinal electronic implant alpha-IMS. *Proc Bio Sci.* 2013;280(1757):20130077.
10. Stingl K, Bartz-Schmidt KU, Gekeler F, Kusnyerik A, Sachs H, Zrenner E. Functional outcome in subretinal electronic implants depends on foveal eccentricity. *Invest Ophthalmol Vis Sci.* 2015;54(12):7658–7665.
11. Ahissar M, Hochstein S. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn Sci.* 2004;8(10):457–464.
12. Wang L, Sharifian F, Napp J, Nath C, Pollmann S. Cross-task perceptual learning of object recognition in simulated retinal implant perception. *J Vis.* 2018;18(13):1–14.
13. Beyeler M, Boynton GM, Fine I, Rokem A. Pulse2percept: A python-based simulation framework for bionic vision. *Proceedings of the 16th Python in Science Conferences*, 2017;81–88.
14. Perez Fornos A, Sommerhalder J, Pittard A, Safran AB, Pelizzone M. Simulation of artificial vision: IV. Visual information required to achieve simple pointing and manipulation tasks. *Vis Res.* 2008;48(16):1705–1718.
15. Trapp S, Bar M. Prediction, context, and competition in visual recognition. *Ann N Y Acad Sci*. 2015;1339:190–198.
16. Pan JS, Bingham GP. With an eye to low vision: Optic flow enables perception despite image blur. *Optom Vis Sci*. 2013;90(10):1119–1127.
17. Zrenner E, Bartz-Schmidt KU, Benav H, et al. Subretinal electronic chips allow blind patients to read letters and combine them to words. *Philos Trans R Soc Lond B Biol Sci*. 2011;278(1711):1489–1497.
18. Jager TF. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *J Mem Lang*. 2008;59(4):434–446.
19. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1–48.
20. Martinez HDR. Analysing interactions of fitted models. 2015. Available at: https://cran.r-project.org/web/packages/phia/vignettes/phia.pdf.
21. Rouder JN, Morey RD, Speckman PL, Province JM. Default Bayes factors for ANOVA designs. *J Math Psychol*. 2012;56:356–374.
22. Wagenmakers E, Marsmann M, Jamil T, et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev.* 2018;25:35–57.
23. Wagenmakers E, Love J, Marsmann M, et al. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychon Bull Rev.* 2018;25:58–76.
24. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*. 2006;113(4):700–765.
25. Link SW, Heath RA. A sequential theory of psychological discrimination. *Psychometrika*. 1975;40(1):77–105.
26. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. 1999;2(1):79–87.
27. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175–191.
28. Webb GI. Overfitting. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning*. Boston, MA: Springer; 2011.
29. Fornos AP, Sommerhalder J, Rappaz B, Safran AB, Pelizzone M. Simulation of artificial vision, III: Do the spatial or temporal characteristics of stimulus pixelization really matter? *Invest Ophthalmol Vis Sci*. 2005;46(10):3906–3912.
30. Gibson JJ. *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin; 1979.