34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Transformative Reality: Improving bionic vision with robotic sensing

Wen Lik Dennis Lui[1], Damien Browne[1], Lindsay Kleeman[1,2], Tom Drummond[2] and Wai Ho Li[1,2]

Abstract— **Implanted visual prostheses provide bionic vision with very low spatial and intensity resolution when compared against healthy human vision. Vision processing converts camera video to low resolution imagery for bionic vision with the aim of preserving salient features such as edges. Transformative Reality extends and improves upon traditional vision processing in three ways. Firstly, a combination of visual and non-visual sensors are used to provide multi-modal data of a person's surroundings. This enables the sensing of features that are difficult to sense with only a camera. Secondly, robotic sensing algorithms construct models of the world in real time. This enables the detection of complex features such as navigable empty ground or people. Thirdly, models are visually rendered so that visually complex entities such as people can be effectively represented in low resolution. Preliminary simulated prosthetic vision trials, where a head mounted display is used to constrain a subject's vision to 25x25 binary phosphenes, suggest that Transformative Reality provides functional bionic vision for tasks such as indoor navigation, object manipulation and people detection in scenes where traditional processing is unusable.**

I. BACKGROUND

In 1968, Brindley and Lewin discovered that electrical stimulation of the visual cortex caused patients to perceive bright dots of light called phosphenes, which occur in predictable locations within the visual field [1]. Subsequently, it was found that phosphenes can be elicited through electrical stimulation of other parts of the visual pathway. Visual prostheses such as retinal and cortical implants apply electrical stimulation to the visual pathway using an electrode array to generate a 2D grid of phosphenes similar to a low resolution dot image. An extensive survey of visual prostheses and their theory of operation can be found in [2].

Figure 1 provides a summary view of how visual prosthetic systems work. Video from a head mounted camera is down sampled into a low resolution image through vision processing. Neuromorphic coding converts pixels in the down sampled image into electrical signals that can be *understood* by the visual pathway. A 2D electrode array, usually implanted into the retina or Primary Visual Cortex (V1), communicates these signals to the visual pathway using electrical stimulation. Other stimulation locations, such as the optic nerve or Lateral Geniculate Nucleus (LGN), yields bionic vision but does not produce a grid of phosphenes conducive to vision processing.

Simulated Prosthetic Vision (SPV) is a non-invasive method used to evaluate the efficacy of visual prostheses.

This work was funded by the Australian Research Council Special Research Initiative in Bionic Vision and Sciences (SRI 1000006)

[1]Monash Vision Group, Monash University, Clayton, Australia {Wen.Lui,Damien.Browne}@monash.edu

[2]ECSE, Monash University, Clayton, Australia {Lindsay.Kleeman,Tom.Drummond,Wai.Ho.Li}@monash.edu

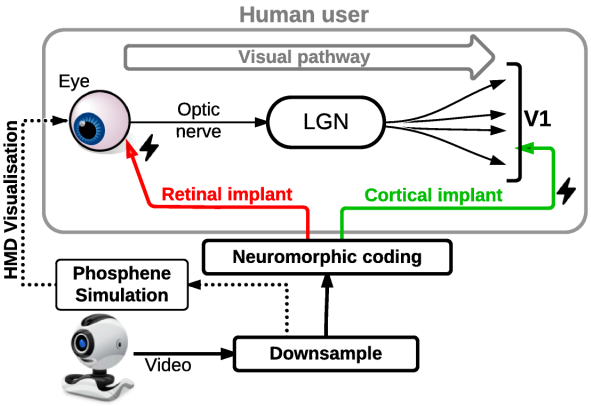


Fig. 1: System diagram summarizing the typical operation of implanted visual prostheses.

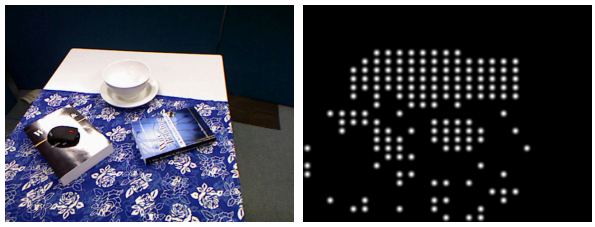
Modern SPV trials make use of a head mounted display with a forward facing camera, which allows a test subject to perform a variety of tasks while being visually constrained to a particular model of bionic vision (visual angle, acuity, levels of grayscale etc.) and a mode of vision processing (binary threshold, edges etc.). This is the approach used in our preliminary SPV trials comparing traditional vision processing and Transformative Reality. An extensive survey of SPV research is available from [3]. We model phosphenes using the canonical parameters recommended by the survey. Note that due to uncertainty regarding the ability to reliably produce multiple grayscale levels of phosphenes, especially for cortical implants, our SPV trials assume binary (on, off) phosphenes.

The spatial and intensity resolution of the phosphene grid produced by a visual prosthesis is constrained by biology, technology and safety. Next generation prostheses are aimed at providing several hundred (Second Sight, Argus III) to around a thousand (Bionic Vision Australia) phosphenes. Monash Vision Group, the author's research laboratory, is funded by the Australian Research Council to develop a 625-electrode cortical implant by 2013. With the above in mind, together with the SPV community's past [4] and present [5] consensuses of what visual representation best constitutes functional bionic vision, our SPV study uses a 25x25 linear grid of phosphenes subtending 10 degrees of vision.

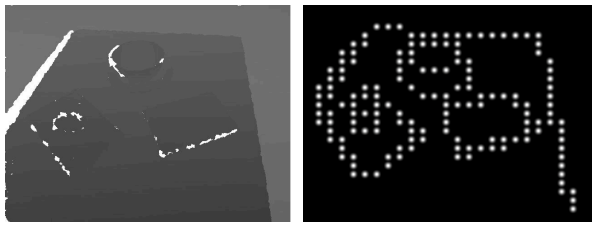
Traditionally, video images from a head-worn camera are converted into low resolution bionic vision using simple image processing techniques. This process is called vision processing. An example is shown in Figure 2a. The camera image on the left is converted to greyscale, down sampled by averaging patches of pixels and then adaptively thresholded into binary, resulting in the bionic vision output on the right.

From here on in, this canonical approach shall be referred to as *traditional vision processing*.

Traditional vision processing result no longer contains the location of objects. It also contains phosphene *noise* due to the textured table cloth. Such severe truncation of sensory information maybe avoidable for simple high contrast scenes (often used in SPV trials), but it is intractable in visually cluttered real world scenes. Better vision processing, such as the ground plane segmentation approach in [6], is needed to improve the saliency of information presented through bionic vision.



(a) Traditional vision processing



(b) Transformative Reality - Visual rendering of structural edges

Fig. 2: Comparison of traditional vision processing and Transformative Reality rendering of structural edges.

II. TRANSFORMATIVE REALITY

Vision processing has been used to enhance imagery for the vision impaired [7] and suggested as a means to improve the quality of low resolution bionic vision [8]. Current research into vision processing for visual protheses generally assumes the following [2]:

1) The input sensor is a camera or a stereo camera pair.
2) Simple image processing is used to produce a low resolution image for subsequent neuromorphic coding.
3) The visual world is represented directly to the subject.

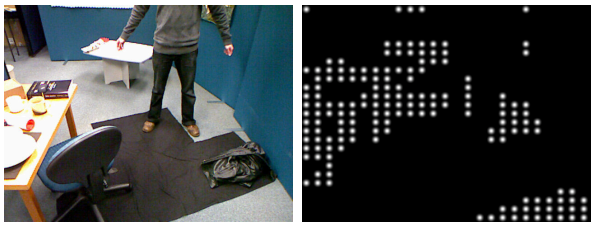
Many of these assumptions were formed in the early history of visual prostheses; Digital sensors were primitive, sensor processing required large computers and robotic sensing was in its infancy. Transformative Reality violates all three assumptions in order to dramatically improve the saliency of information provided through low resolution bionic vision.

Transformative Reality (TR) assumes the following:

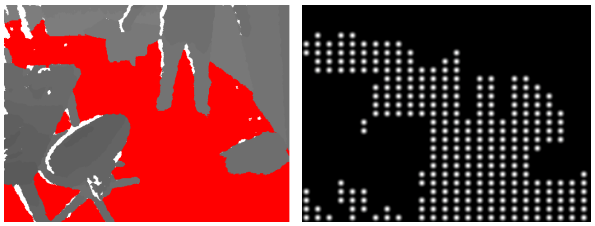
1) No restriction on the input sensors, including the use of multiple sensors and non-visual sensors.
2) Real time robotic sensing algorithms are used to model the world around the subject.
3) Models are visually rendered to allow symbolic representations such as using avatars for a person's face.

These new assumptions allows TR great freedoms in how it senses, models and visually represents the world. Figure 2b shows the TR rendering of *structural edges* for the scene in Figure 2a. The result is achieved using a lightweight range camera that senses the world in 3D (left image, darker pixels are nearer). An output phosphene is lit when the range data in the corresponding region is non-planar. Such structural edges [9] are commonly used in robotic sensing to segment objects in visually cluttered scenes. The TR structural edges clearly improve upon traditional vision processing for tasks dealing with objects.

Traditional vision processing also implicitly assumes visual scenes with high contrast to allow reliable sensing using a camera. Apart from laboratory environments, the real world rarely provides sufficient contrast. Figure 3 shows traditional vision processing and the *empty ground* TR mode output for a cluttered indoor scene. For indoor navigation, traditional vision processing fails to highlight obstacles such as the legs of the computer chair due to the lack of visual contrast.



(a) Traditional vision processing



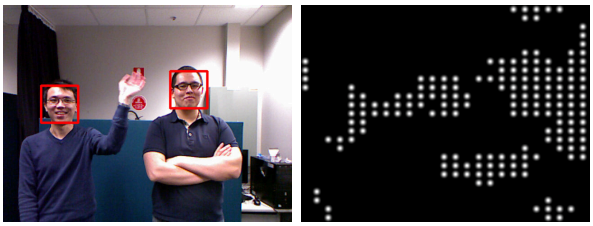
(b) Transformative Reality - Visual rendering of empty ground

Fig. 3: Traditional vision processing compared to empty ground TR mode for indoor navigation.

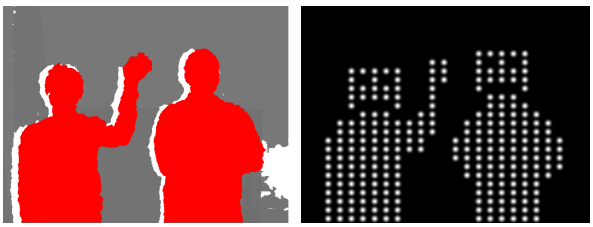
The empty ground TR mode operates by first generating a ground plane estimate using a range image (left image of Figure 3b) and the direction of gravity sensed using an accelerometer. The use of an accelerometer allows the algorithm to run in real time by restricting the search space of potential ground planes to those with normals that point upwards against gravity. The ground plane is estimated using a RANSAC-based inverse depth approach [10]. Plane inliers in the range image are shown in red on the left of Figure 3b). Phosphenes are rendered at the corresponding locations in the TR output. This results in a clear representation of navigable ground as lit phosphenes and potential obstacles as dark regions. Note that the computer chair and leather jacket (bottom right of scene) are correctly represented as obstacles.

TR can also render entities that are visually too complicated to represent directly. The *people detection* TR mode, shown in Figure 4, uses a combination of colour and range

camera data. Frontal faces are detected using the Viola-Jones boosted classifier [11]. A person's body is found using the range camera by searching below detected faces (red regions in bottom left of Figure 4). The resulting TR output combines a low resolution face avatar with a filled region for the body to represent a person. Again, TR provides an improvement over traditional vision processing.



(a) Traditional vision processing



(b) Transformative Reality - Visual rendering of people

Fig. 4: TR rendering of people. An avatar represents frontal faces and a contiguous filled region represents a human body.

III. PRELIMINARY SPV TRIALS

Section II described three TR modes: structural edges, empty ground and people detection. These modes were fully implemented in a prototype as described by the system diagram in Figure 5. Robotic sensing algorithms were implemented using C++ on a consumer laptop (Intel i5) running Ubuntu. A Microsoft Kinect provides multi-modal sensing (range camera, colour camera and accelerometer).

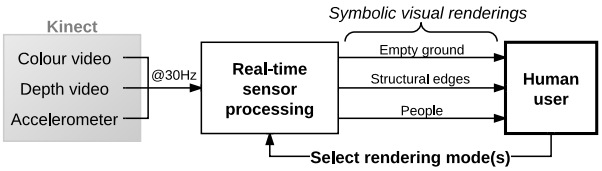


Fig. 5: TR system used in SPV trials.

As shown in Figure 6, the Kinect was attached to a NVIS SX-60 head mounted display (HMD) for our Simulated Prosthetic Vision (SPV) trials. The system operated in real time at roughly 25Hz for all trials. Four preliminary trials were conducted where the first author wore the HMD while being monitored by the last author. Images taken from the subject's point of view are shown at the end of the paper in Figures 7 to 10. In all trials the TR output allowed the subject to complete the task described in the caption independently.

SPV trials were also conducted using traditional vision processing but resulted in early termination as the subject was unable to proceed without consistent external help

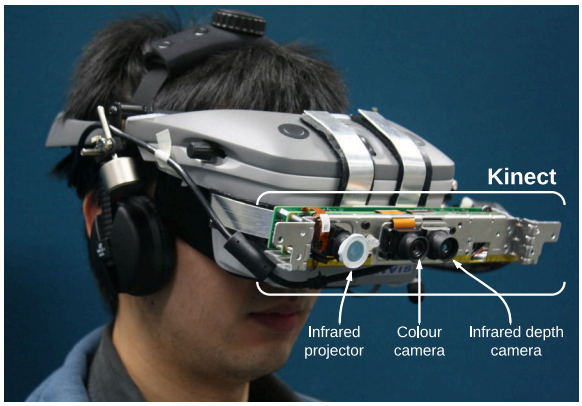


Fig. 6: Customized head mounted display used in SPV trials.

including interventions to prevent collisions with obstacles. The subject reported that the vision processing output had arguable value over simply using his sense of touch alone. Feedback obtained from 10 naive subjects viewing TR and traditional vision processing sequences from the SPV trials on a computer monitor echo the trial results with all subjects preferring TR over traditional processing.

IV. CONCLUSIONS AND FUTURE WORK

Additional SPV trials will be conducted in collaboration with medical researchers at Monash University and Vision Australia (advocacy group for vision impaired Australians) to assess the improvements made by TR in a more quantitative manner. Preliminary SPV trials suggest that Transformative Reality is able to generate useful low resolution bionic vision from real world scenes. TR can operate in less structured scenes with low visual contrast, which poses a problem to traditional vision processing.

REFERENCES

[1] G. S. Brindley and W. S. Lewin, "The sensations produced by electrical stimulation of the visual cortex," *The Journal of Physiology*, vol. 196, pp. 479–493, 1968.

[2] G. Dagnelie, *Visual Prosthetics*, G. Dagnelie, Ed. Boston, MA: Springer US, 2011.

[3] S. C. Chen, G. J. Suaning, J. W. Morley, and N. H. Lovell, "Simulating prosthetic vision: I. Visual models of phosphenes," *Vision Research*, vol. 49, no. 12, pp. 1493–1506, Jun. 2009.

[4] K. Cha, K. Horch, and R. A. Normann, "Simulation of a phosphene-based visual field: visual acuity in a pixelized vision system." *Annals of biomedical engineering*, vol. 20, no. 4, pp. 439–49, Jan. 1992.

[5] M. P. Barry and G. Dagnelie, "Simulations of Prosthetic Vision," in *Visual Prosthetics: Physiology,*, G. Dagnelie, Ed. Boston, MA: Springer, 2011, pp. 319–341.

[6] N. B. Chris McCarthy and P. Lieby, "Ground surface segmentation for navigation with a low resolution visual prosthesis," in *EMBC*, 2011.

[7] E. Peli and T. Peli, "Image enhancement for the visually impaired," *Optical Engineering*, vol. 23, no. 1, pp. 047–051, 1984.

[8] G. Dagnelie, "Virtual technologies aid in restoring sight to the blind," in *Communications Through Virtual Technology*. IOS Press, 2001, ch. 15, pp. 247–271.

[9] R. Krishnapuram and S. Gupta, "Morphological methods for detection and classification of edges in range images," *Journal of Mathematical Imaging and Vision*, vol. 2, no. 4, pp. 351–375, Dec. 1992.

[10] T. J. J. Tang, W. L. D. Lui, and W. H. Li, "A lightweight approach to 6-DOF plane-based egomotion estimation using inverse depth," in *ACRA*, 2011, p. Online. [Online]. Available: http://youtu.be/qZbsuwzkZp4

[11] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *IJCV*, vol. 57, no. 2, pp. 137–154, May 2004.

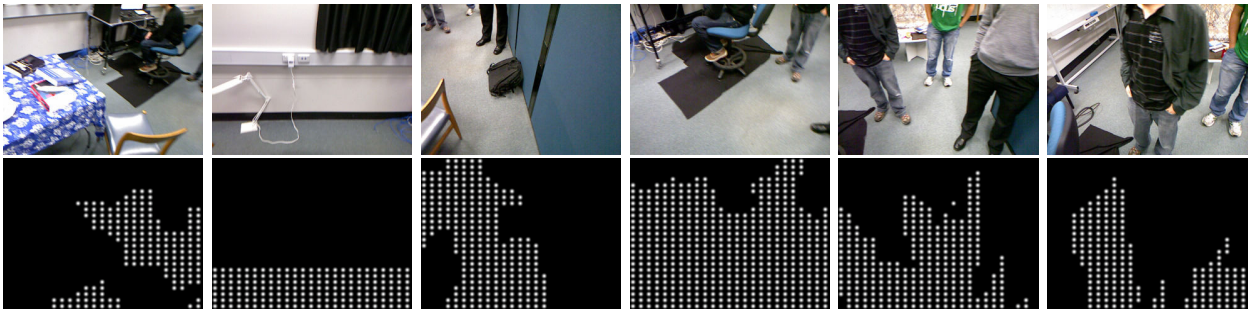


Fig. 7: Visual navigation trial: Subject told to navigate around obstacles and people. Top: Kinect colour camera. Bottom: User's POV within HMD (empty ground TR mode). Every 50th frame shown.

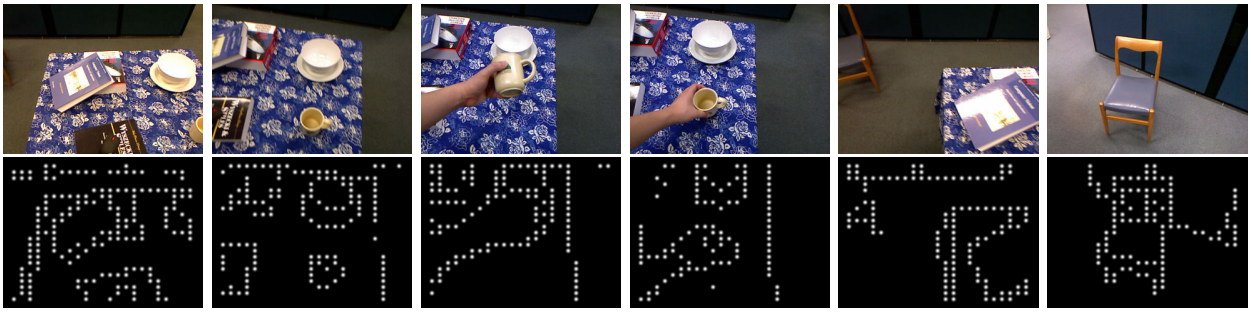


Fig. 8: Object detection and manipulation trial: Subject told to pick up a cup then navigate to a chair (both placed randomly in test area). Top: Kinect colour camera. Bottom: User's POV within HMD (structural edge TR mode). Every 50th frame shown.

Fig. 9: Human interaction trial: Subject asked to count the number of people and point to the person who is waving. Top: Kinect colour camera. Bottom: User's POV within HMD (face and body detection). Every 15th frame shown. Note that the TR system automatically switched modes once a frontal face is detected (green rectangle).

Fig. 10: Free form trial: Subject asked to organise a domestic environment. Top: Kinect colour camera. Bottom: User's POV within HMD (ground plane and structural edge modes). Full Video: `http://youtu.be/iK5ddJqNuxY`