Transforming the Commonplace through Machine Perception: Light Field Synthesis and Audio Feature Extraction in the *Rover* Project

Best Paper Award

Robert Twomey and Michael McCrea

Robert Twomey
Department of Art
Youngstown State University
1 University Plaza
Youngstown, OH 44555, U.S.A.
rtwomey@ysu.edu

Michael McCrea
DXARTS
University of Washington
207 Raitt Hall
Seattle, WA 98195, U.S.A.
mtm5@uw.edu

See <www.mitpressjournals.org/toc/leon/50/4> for supplemental files associated with this issue.

ABSTRACT

***Rover* is a mechatronic imaging device inserted into quotidian space, transforming the sights and sounds of the everyday through its peculiar modes of machine perception. Using computational light field photography and machine listening, it creates a kind of cinema following the logic of dreams: suspended but mobile, familiar yet infinitely variable in detail. *Rover* draws on diverse traditions of robotic exploration, landscape and still-life depiction, and audio field recording to create a hybrid form between photography and cinema. This paper describes the mechatronic, machine perception, and audio-visual synthesis techniques developed for the piece.**

Rover synthesizes three areas of emergent technology: computational light field photography, machine listening, and low-cost embedded motion control. The project engages these techniques to create a hybrid, variable representation of place. *Rover*'s method is similar to that of the landscape painter, travelogue writer, or field recording artist: traveling to a variety of locations gathering audio and visual documentation of its experience. However, rather than producing fixed representations, it captures dense sets of data to be computationally explored, refocusing through scenes in an endless process of machine reflection. Figure 1 shows an example of this searching focus within a scene.

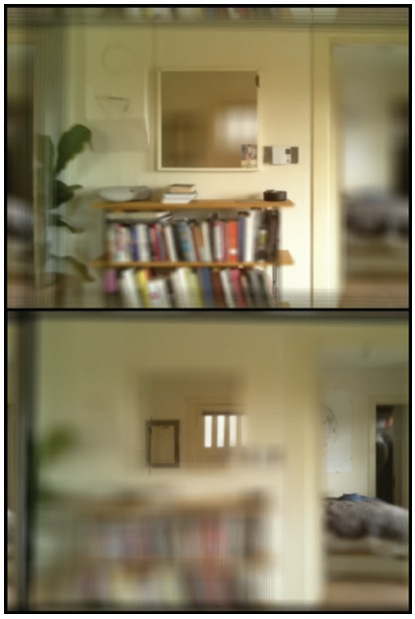

Figure 1. Refocusing a domestic scene to shift the focal plane from near (top) to far (bottom). (© Robert Twomey and Michael McCrea)

Rover has three main components: a mechatronic imaging apparatus, an analysis back-end, and synthesis front-end (Figure 2). The imaging apparatus is a single camera with a computer-controlled positioning system. Hundreds of images are gathered in a structured way, recording light from multiple vantage points within a scene. Sound is also recorded for the duration of the image sampling.

Upon completion of this on-site engagement, the audio and visual samples are processed to extract salient features. Multiview geometry techniques are used to recover camera positions, producing a light field data set. In a parallel process, sonic moments are identified with machine-listening techniques and classified for later use.

Finally, a real-time synthesis engine takes the products of these analyses—light field data and classified audio—and creates a 30-minute audio-visual composition. In a single cycle of the piece, viewers are taken on a journey

   doi:10.1162/LEON_a_01458

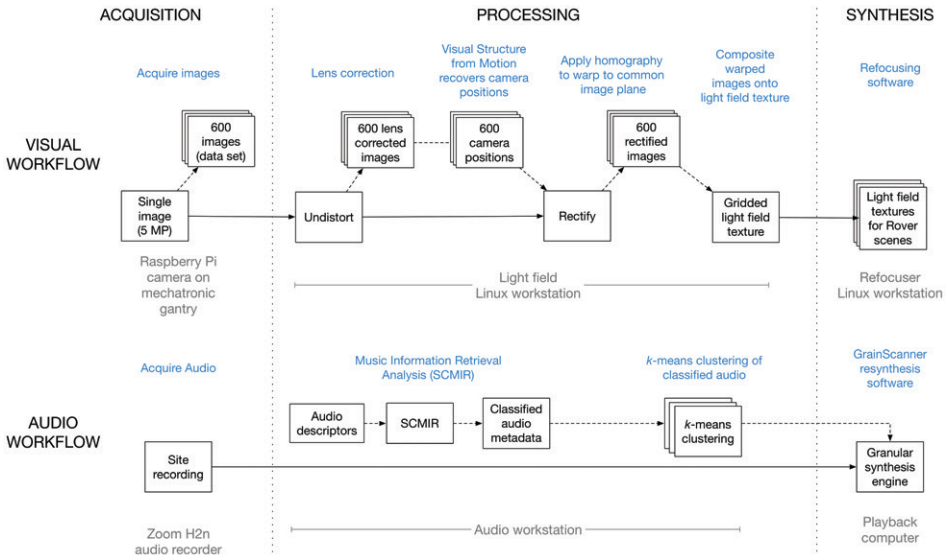

Figure 2. Audio and visual processing workflow. (© Robert Twomey and Michael McCrea)

through the sites *Rover* has visited. The viewer peers through a dreamlike vantage onto spaces that emerge and dissolve as *Rover* churns through recorded images, varying its synthesis parameters to manipulate its own perception. Sounds are similarly revisited and reshaped until they are no longer commonplace.

The result is a hybrid between photography and video, a kind of cinema that follows the logic of dreams: suspended but mobile, familiar yet constantly shifting in detail. Indeed, the places visited through *Rover*'s motility are the kinds of places found in dreams: cliffside, seaside, bedside, trapped in the corners of homes, or adrift and unable to return home.

Background: Light Field Imaging

In traditional photography, the camera captures an image from a single vantage point, losing all information about the directionality of light rays through the scene. In light field imaging, a camera or imaging device samples a scene from multiple views, capturing both positional and directional information about incident light. The additional information in a light field image allows for the synthesis of new views of the scene, well beyond the physical limitations of traditional lenses, resulting in ex post facto control over rendered depth of field, vantage point, and focal plane [1].

Light field imaging has existed in theory and practice for more than a century. Nobel Laureate Gabriel Lippmann described and created a multilens plenoptic camera in 1911, shooting 12 coplanar images on photographic film using 12 lenses [2]. In the past 20 years, advances in digital image sensors and computing capabilities have enabled breakthroughs in light field techniques: dynamic light field–based rendering [3], light field acquisition with custom camera arrays [4], moving gantries [5], hand-held light field cameras [6], and even consumer products [7].

Light field synthesis, when augmented by programmable robotic movement, computer vision and machine listening techniques, exists somewhere between photography, photogrammetry, and cinema. Mature, open-source computer vision libraries, cheap imaging hardware, and affordable motion control systems have put light field techniques within the reach of independent artists and researchers.

Formal Strategies

This work draws on the literary tradition of the travelogue, documenting a subjective journey through unfamiliar places. The sites *Rover* visits are familiar to us: homes, forests, bedsides. At the same time, *Rover*'s gaze is mechanical and exacting—that of a robotic explorer. This intersection of the travelogue form with a nonhuman gaze creates an aesthetic situation wherein the viewer sees the commonplace transformed—familiar scenes observed through an alien subjectivity.

The compositional structure for *Rover* produces a looping orbit through multiple interior and exterior sites, inspired by W.G. Sebald's *The Rings of Saturn* [8]. *Rover* evokes the digressive and at times bizarre aspects of Sebald's peripatetic journey: captivation in quotidian details found in photographic artifacts and encountered throughout the landscape. The narrative is one of constant departure; each scene, upon persistent observation, grows increasingly remote.

Rover's unique modality is born of a technology modeled after our own vision—the camera. While artists have developed a stable understanding of the photograph as form, its reinvention as a shifting, manipulable image through light field photography complicates its identity as fixed media. *Rover*'s dynamic reworking of photographic space invites the viewer along as it actively manufactures synthetic images. Moment to moment, *Rover* ultimately produces hybrid photographs, maintaining the grip on nostalgia and affective communication advanced in Roland Barthes' *Camera Lucida* [9], but they reveal their own contingent instability. Viewers' understandings of apparently inert moments or places are given new dimension through *Rover*'s computational search. In one scene, a wooden frame hangs above a bookshelf, holding a floating, dimensionless volume of light. In the next moment, as focus shifts, the volume resolves into an image of *Rover* seeing itself. The frame, however, has disintegrated and the optical play is revealed to be the inverted space of a mirror.

In contrast to *Rover*'s continuous motion, moments of rest in the image resynthesis acquire an unexpected stillness and distance. We find resonance and inspiration for this in the cinematography of Andrei Tarkovsky, enveloped in the effects of light, depth, and stilled time. In particular, his personal polaroids [10] pique longing and nostalgia through their traces of inhabitation yet lack of human presence. *Rover*'s interiors are saturated with signs of inhabitation, though in cyclically revisiting these places, those signs become fleeting. A floral centerpiece on a dining table, for example, is rendered ephemeral as *Rover*'s focus wanders away, dissipating the object.

Rover departs from domestic interiors into expansive exteriors—seaside, cliffside, and pastoral trails. Viewers witness the dissolution of the photographic eye as light fields are shaped to reveal and obscure features of the landscape: the trunk of a tree dissolves to reveal the surrounding forest, only to further dissolve, giving way to the distant sea. Through the perception of the viewer this selective attention is attributed with subjective intent, akin to the controlled focus and erasure in Gerhard Richter's paintings of reappropriated landscape photographs [11].

While the refocusing system can perform "strict" light field synthesis towards synthetic photography, the parameters for resynthesis can also push the process into abstraction (Figure 3). For example, wildflowers on a cliffside, when faithfully rendered, serve as a pointillistic texture surrounding an empty pathway leading into the distance. When the synthesis bounds are intentionally crossed, these yellow flowers fragment into a color field that merges with evening light caught by the clouds in the background.

Figure 3. Refocusing parameters can push the image synthesis into abstraction. (© Robert Twomey and Michael McCrea)

Figure 4. The CNC cable-driven camera gantry, shooting an exterior scene (top) and an interior scene (bottom). (© Robert Twomey and Michael McCrea)

This abstraction violates the presumed fidelity of machine perception. In moments of abstraction—when *Rover* appears lost, when depth and sense are seemingly destroyed and *Rover* fails to spatially resolve its light data—its vision, while imperfect, is no less "true." *Rover*'s inability to faithfully recount its travels paradoxically finds resonance with our own imperfect—or impressionistic, one might even say creative—memory.

The various degrees of plasticity in the light field synthesis technique highlighted here hold great appeal to artists seeking to recruit the photographic image back in both time and space.

Image Acquisition

Light field photography requires dense sets of spatially rectified images. For each of the 15 scenes included in the initial iteration of the *Rover* project, approximately 600 images were acquired on-site. To achieve this structured sampling, we developed a computer-controlled positioning system with a single camera moving within a two-dimensional plane (Figure 4). This apparatus is a cable-driven computer numeric control (CNC) system, suspending an embedded processor (Raspberry Pi) with image sensor (pi camera) on spectra line between two portable C-Stands. Camera movements are directed by an external computer sending GCODE commands and interpreted by an Arduino stepper motor controller running GRBL, an open-source, embedded GCODE-parser [12].

This single moving camera approach offers several advantages. It is economical and portable, a key constraint given the exploratory, site-based nature of the project. With a cable-based configuration, the synthetic aperture can be expanded or condensed to match the desired spatial resolution: an outdoor scene may require a large aperture with a deep focal plane (Figure 5), while an indoor scene may be better suited to a smaller aperture and denser image set (Figure 1). In the present configuration,

Rover's imaging plane can span 40 feet. Finally, the quantity and positioning of images within the capture sequence can be modified on-site.

Figure 5. Outdoor scenes are well-suited for a large synthetic aperture, accommodated by the configurable capture system. (© Robert Twomey and Michael McCrea)

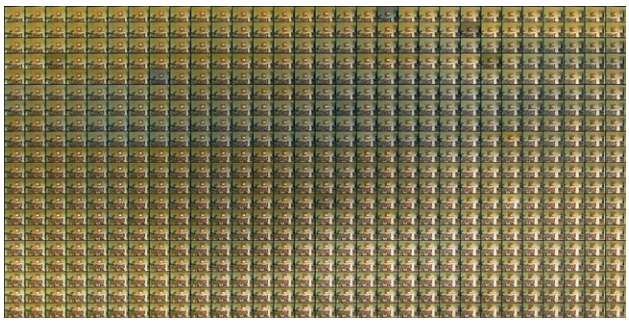

Figure 6. Temporal effects in the light field data: sunlight in the room shifts between the upper left and lower right of this image set. This dataset "contact sheet" is used to synthesize Figure 1. (© Robert Twomey and Michael McCrea)

Temporal parameters offer control over how *Rover* records the passage of time. For example, shifts in natural light are recorded into the light field data during a long-duration scene capture (Figure 6). This temporal dimension can be navigated in a nonlinear way during scene reconstruction. Furthermore, image acquisition can be ordered spatially in anticipation of particular moments or events witnessed by the system. For instance, in one scene of the final work, a subset of images was selected to emphasize the passage of a train through the landscape.

Visual Analysis

Once captured, image processing is a multistep workflow. Images are 1) corrected for lens distortion, 2) located in space using structure from motion algorithms, 3) rectified to a common virtual imaging plane, and 4) arranged in gridded textures to be sampled for resynthesis.

Lens distortion is corrected using a chessboard calibration procedure with Python/OpenCV [13]. Intrinsic and extrinsic lens parameters are applied with Python code to undistort the raw images. Since the Raspberry Pi Camera is a fixed-focus device, chessboard calibration is only done once and can be applied for all subsequent image sets.

Next, the undistorted images are used to recover camera positions with Visual Structure From Motion (VSFM), a comprehensive multiview geometry software [14]. Visual features are recognized within images [15]. Pairwise feature matches are created between images, used for a sparse scene reconstruction, and camera locations are refined through multicore bundle adjustment [16]. This results in 3D positions and rotations for all cameras relative to the visual scene (Figure 7).

To produce a light field dataset for each scene, these hundreds of camera views are rectified to a common virtual image plane. Figure 8 shows an input image, single projected image, and stack of aligned, rectified images. This rectification also compensates for slight error in the physical capture position. Finally, rectified images are stored as a maximum resolution "contact sheet" (as in Figure 6) to serve as the texture data for refocusing software. In this texture, subimage location corresponds to position in the acquisition grid. This texture is selectively

sampled for synthetic aperture manipulation and resynthesis effects as described below.

Audio Analysis

One goal for the audio synthesis was to use computational listening techniques to explore how a machine narrator might identify sonic qualities or moments of interest in a scene to then reimagine for the viewer. This process begins at the site of light field capture, where stereo field recordings are made over the approximate duration of the capture sessions. The recordings were continuous and unscripted, capturing many small events and shifts in environmental sound. This material presents a challenge that is familiar to artists working with durational recordings or, more generally, with large datasets: to distill large amounts of information into an experience that conveys the variety and scope of the source material, while also providing emphasis and interpretation.

To this end, we developed a framework of computational synthesis that begins with audio feature extraction. This stage translates the recordings from single-dimensional (temporal) material into a multidimensional space that is organized by perceptual, spectral, or temporal characteristics. This task was performed in SuperCollider, using the SCMIR library by Nick Collins [17]. After analysis, the multidimensional data is further ordered into *k*-means clusters, ordered by self-similarity. This creates a nonlinear structure that can be navigated algorithmically upon resynthesis.

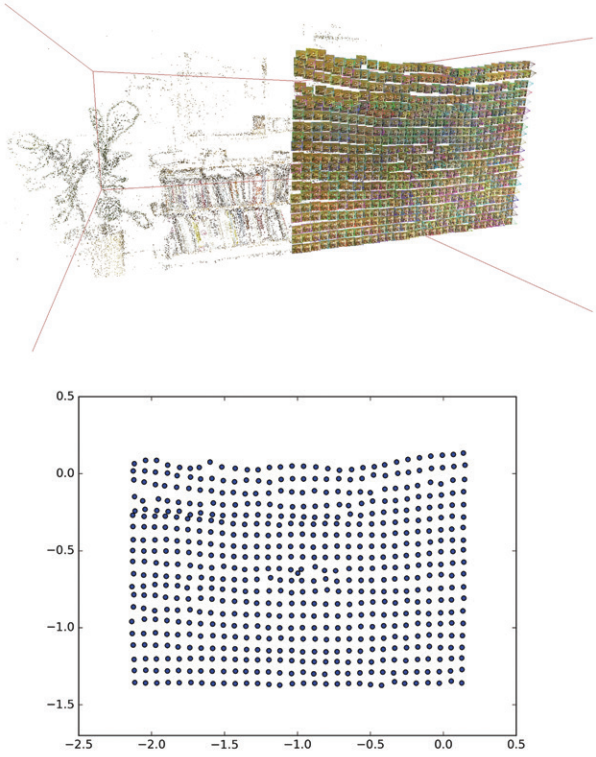

Figure 7. Recovered camera positions and rotations relative to a scene (top). Same cameras plotted on virtual image plane (bottom). Note the slight irregularity in the sampling grid due to the nonrigid acquisition platform. (© Robert Twomey and Michael McCrea)

Figure 8. A test image (top left) projected onto an image plane (top right) and aligned with other rectified images (bottom). (© Robert Twomey and Michael McCrea)

Resynthesis

The front end of *Rover* is a live synthesis system. It has two parts: an audio synthesis engine written in SuperCollider, and a light field refocusing engine written in openFrameworks and OpenGL Shading Language (GLSL). These systems are controlled in real time with Open Sound Control (OSC) messages generated from a unified compositional control system written in SuperCollider.

The scale of our light field data requires a high-performance program for resynthesis. Initially written in C++, the refocusing algorithm has been ported to GLSL and optimized to run on a high-end GPU (NVIDIA Titan X), employing its full compute capability and usable memory (12GB of pixel data). It loads high-resolution, uncompressed light field textures and enacts a broad range of transformations inherent to synthetic aperture rendering of light fields. OSC signals modulate the depth of field, parallax, focus, crops, pans, zooms, contrast, brightness, and hue shifts.

The texture is comprised of many rectified images, forming a "contact sheet" as described in the previous section. Pixel data from all, or only a subset, of the images are sampled to generate the projected scene. The positions of the subimages used determine the particular vantage point of the synthesized view. Therefore, the perspective of the viewer can be made to traverse the image plane by dynamically sampling different regions of the contact sheet.

For the audio synthesis, a compositional process brings clustered audio feature data back into a temporal dimension for each scene. We wrote a software instrument in SuperCollider that allows high-level parametric control over granular synthesis—a process wherein fragments of sound are arranged in time with varying duration, amplitude, pitch, and density. The instrument, GrainScanner, uses *k*-means clusters as stochastic anchors, scanning each cluster for sounds of high similarity in the center and more anomalous sounds toward the outer radius. Figure 9 shows examples of clustered audio data points and the synthesis control GUI.

The GrainScanner offers a wide degree of control over the perception of synthesized sound as diegetic (a faithful reproduction of the original) or purely gestural (by stretching, freezing, layering, cycling, or phasing moments in time). This diversity of sonic character serves many ends: reinforcing the sense of place shown on screen, enhancing the motility of the machine narrator as it navigates the space of the image, and augmenting the memorious quality of the visual system by recounting, echoing, and realigning sounds in real time.

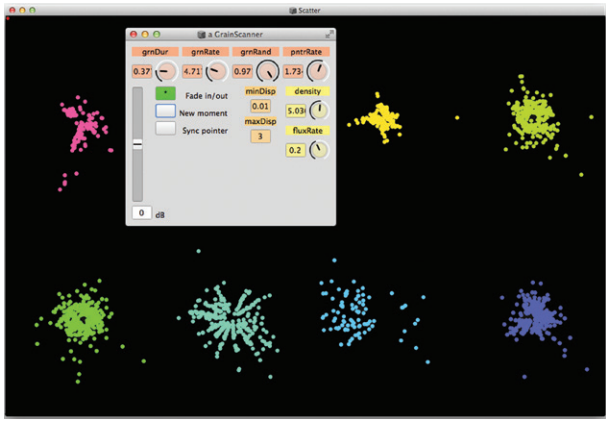

Figure 9. Clusters of feature-classified audio samples based on site recordings with an instance of a GrainScanner GUI.

This framework of "generative computational synthesis"—the aggregate of offline analysis and real-time synthesis—used for both sound and image, is analogous. This allows the algorithmic controls for sound synthesis to be composed in concert with those used for visual synthesis. For example, the drifting focus of an image is accompanied by a modulation in granular density of the sound field, obscuring or revealing the source material in concert with the shifting clarity of the image.

Presentation

Rover premiered at the Black Box 2.0 Festival in Seattle, Washington in 2015 [18]. It was installed in a 20-foot shipping container as a single-channel video projection with four-channel audio (Figure 10). A rectangular projection screen was suspended toward the rear of the container. Two speakers were installed behind the screen and two speakers hung just inside the container doors, behind a viewer that had entered the viewing space. Installed in this long box and lit only by the floating image plane, *Rover* echoes both its own imaging process and the form of the *camera obscura*, its artistic and technical antecedent, creating the impression for the viewer of stepping into a large imaging device.

Figure 10. *Rover* installed at the Black Box 2.0 Festival. (© Robert Twomey and Michael McCrea)

Conclusion and Future Directions

Over the course of development, we have identified a number of improvements and new directions for *Rover*. Technically, the GLSL shader code can be optimized to work with full-resolution light field datasets beyond the memory limits of the graphics card, using dynamic texture loading as is typical in the game industry. The image and audio analysis pipeline can be ported to cloud infrastructure [19]. Variations on the image acquisition apparatus, for instance a rigid Cartesian CoreXY gantry [20], have been developed to enable new capture geometries.

There are also a number of desirable additions to *Rover*'s operation; for instance, dynamically updating the light field data to include new images acquired during the course of an installation or exhibition. The resynthesis engine can be integrated into other responsive or immersive (AR/VR) display paradigms. Nonuniform refocusing parameters can be implemented using control masks, for instance varying focus across regions of an image. The mechatronic capture process itself has a performative quality that could be used for performance-based project iterations.

Most importantly, as the project is iterative by nature, we continue to develop content for the *Rover* system, experimenting with new subject matter and seeking new sites for *Rover* to visit.

Acknowledgments

We would like to acknowledge the support of Julia Fryett and Anne Couillaud of Aktionsart and the University of Washington's Center for Digital Arts and Experimental Media.

References and Notes

1. V. Vash, "Synthetic Aperture Imaging Using Dense Camera Arrays," PhD Thesis, Stanford University (2007).
2. "Integral Photography," *Scientific American*, 165 (1911).
3. M. Levoy and P. Hanrahan, "Light field rendering," *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH (1996).
4. The Stanford Multi-Camera Array, <https://graphics.stanford.edu/projects/array/>.
5. "Light Field Gantry," <http://lightfield.stanford.edu/acq.html>.
6. R. Ng, "Light Field Photography with a Hand-Held Plenoptic Camera," Stanford University Computer Science Tech Report CSTR 2005-02 (2005).
7. Lytro Illum, <https://www.lytro.com/imaging>.
8. W.G. Sebald, *Rings of Saturn* (New York: New Directions, 1995).
9. R. Barthes, *Camera Lucida* (New York: Hill & Wang, 1980).
10. A. Tarkovsky, *Instant Light: Tarkovsky Polaroids* (London: Thames & Hudson, 2006).
11. G. Richter, *Gerhard Richter: Landscapes* (Ostfildern: Cantz Verlag, 1998) pp. 84–87, 97–99.
12. GRBL, an open-source, embedded, high-performance g-code-parser and CNC milling controller written in optimized C, <https://github.com/grbl/grbl>.
13. Python / OpenCV Camera Calibration Example, <http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_calib3d/py_calibration/py_calibration.html>.
14. C. Wu, "VisualSFM: A Visual Structure from Motion System" (2011), <http://ccwu.me/vsfm/>.
15. C. Wu, "SiftGPU: A GPU implementation of Scale Invariant Feature Transform (SIFT)" (2007), <http://cs.unc.edu/~ccwu/siftgpu>.
16. C. Wu, et al., "Multicore Bundle Adjustment," Proceedings of *IEEE CVPR,* pp. 3057–3064 (2011).
17. N. Collins, SuperCollider Music Information Retrieval Library (SCMIR), <https://composerprogrammer.com/code.html>.
18. Black Box 2.0 Festival (6 May–7 June 2015), <http://www.aktionsart.org/allprojects/2015/5/6/black-box-2>.
19. Supported by an Amazon Web Services Cloud Credits for Research Grant, awarded December 2016.
20. CoreXY Cartesian Motion Platform, <http://corexy.com/theory.html>.