SwarmVision: Autonomous Aesthetic Multi-Camera Interaction

George Legrady
Media Arts and Technology
UC Santa Barbara
Santa Barbara, CA, USA
legrady@mat.ucsb.edu

Danny Bazo
Media Arts and Technology
UC Santa Barbara
Santa Barbara, CA, USA
dannybazo@gmail.com

Marco Pinter
Media Arts and Technology
UC Santa Barbara
Santa Barbara, CA, USA
marco@marcopinter.com

Abstract

A platform of exploratory networked robotic cameras was created, informing new directions in computer vision engineering and utilizing an aesthetic approach to experimentation. Initiated by research in autonomous swarm robotic camera behavior, SwarmVision is an installation consisting of multiple Pan-Tilt-Zoom cameras on rails positioned above spectators in an exhibition space, where each camera behaves autonomously based on its own rules of computer vision and control. Each of the cameras is programmed to detect visual information of interest based on a different algorithm, and each negotiates with the other two, influencing what subject matter to study in a collective way. The emergent behaviors of the system suggest potential new approaches in scene reconstruction, video-based behavior analysis and other areas of vision and imaging research.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Arts and Humanities: Media Arts

Keywords

Human-centered computing, scene recognition, natural images, spatial layout, emergent behavior, aesthetic function, multi-robot interaction, computational photography.

1 Introduction

Through experimental testing, we examine the qualitative aspect of the machine-generated image’s visual and semiotic structure. How do machine generated images differ from those made by humans? And what may be areas of relevance for machine vision study for utilitarian and other applications? Images generated by humans and machines cover the full breadth from simple to complex, from easily recognizable to undecipherable, images that result from unintended disruptions, to studies of visual chaotic phenomena. There is no limit to the range of visual articulation, in particular when machines are activated to function on their own, as cultural values are not the guiding or restraining factors.

Useful image creation and its interpretation to a large extent involves cultural understanding [Goodman 1976]. It is a learned process following rules of visual ordering or structuring [Peters 2007]. We have far to go to arrive at methods for translating processes in our understanding of images that are based on emotional, perceptual, cognitive factors at the level of affect and the aesthetic, and additionally where the symbolic, cultural, syntactic biases shape our reading of the image.

1.1 Motivation & Aesthetic Research

Research in computational photography is dedicated to advancing the image-capture technology and extending the capabilities of digital photography, constraining itself to the technological development phase. We bring an experimental approach to examine the qualitative aspect of the machine-generated image’s visual and semiotic structure. Our project’s purpose is to develop new solutions through the study of such images, guided by an approach based in aesthetics and visual language, with the intent of discovering findings that may inspire further engineering development [Elkins 1999].

1.2 Related Work

There are a number of early artistic precedents that have explored machine-driven camera image-capture behavior driven by cultural or aesthetic interests. Michael Snow’s 1971 pioneering film “La Region Centrale” ceded aesthetic definition to the automated recording process of a programmed robotic-arm-mounted camera which received instructions from a cassette tape. Paul Garrin and David Rokeby’s 1998 “Border Patrol" implemented early face recognition computer vision algorithms to create an aggressively invasive experience, turning viewers into inescapable targets. Seiko Mikami’s “Desire of Codes” constructs on a large screen a hexagonal compounded image of spectators present in the exhibition space. Perhaps most relevant is the way in which the Microsoft’s Photosynth application’s positioning of online found photographs into a 3D defined point cloud of related images introduces a range of issues addressing how to configure images spatially based on angle, focus, zoom, and vantage point. However, whereas technologies such as Photosynth aim to arrive at smooth melding of images, our project is engaged in encouraging the noisy outcome of layered images resulting from autonomous investigations.

1.3 Description

Three cameras mounted on motorized rails scan the room for visual information of interest as defined by computer vision algorithms. Even though each camera functions autonomously, they compare and address each other’s results. When one of the cameras identifies subject matter of high saliency for its particular algorithm, the other cameras suspend their search and converge to

SIGGRAPH Asia 2013, November 19 – 22, 2013, Hong Kong.

examine the identified subject of interest from their different angles.

In the installation, four visualizations are featured on two screens/projections. The first screen features what each of the three cameras "see" - a depiction of what their vision algorithms are currently processing. The second screen shows an overview in a 3D reconstruction of the environment featuring a live video stream of the location of the cameras, and of the images they generate. Each camera continuously produces 10 still frames per second, and fills the 3D space with up to a hundred images per camera resulting in a volumetric form of layered stacked photographs that continuously changes as images fade away. The images' sizes and locations are determined by the locations and poses of the cameras, as well as their focal planes and focus locations at a given moment. The 4th visualization features the sum of activities situating all generated images and the three camera locations within a reduced virtual 3D spatial reconstruction of the exhibition space.

In the exhibition setting, visual segments of spectators who enter the viewing space populate the images, leaving an imprint of their presence that is later erased as the images sequentially fade away. On the two screens, viewers can perceive both individual camera behaviors (microcosmic) and their relationships to each other (macrocosmic) through communication. The project explores the transformative condition of the photographic process as it transitions from a still image to one that is reified within physical space.

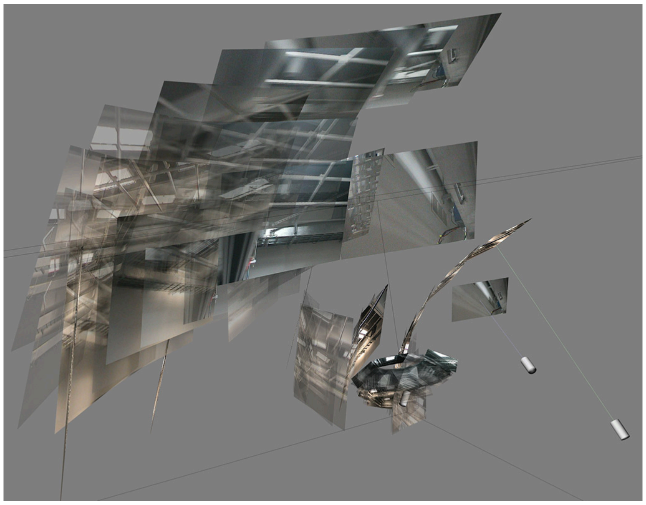

Figure 1: *Image created by an autonomous swarm consisting of three robotic cameras. The cameras each seek out features of interest using separate computer vision algorithms, and together they generate a spatially reconstructed representation of their visual environment. Image © George Legrady.*

2 Technical Description

2.1 Hardware

2.1.1 PTZ Ethernet Cameras

SwarmVision's three cameras are commercial IP cameras of the type typically used for security applications. They have sufficient range of motion to cover approximately a hemisphere with their camera's view. Their video images are sent to the computer through ethernet as streaming videos, and control commands are sent from the computer via serial communication.

2.1.2 Motorized Rails

The custom-built, belt-driven rails are driven by a DC motor with a built-in rotary encoder, and are controlled by an Arduino with a motor controller board. Software on the Arduino includes a proportional-integral-derivative controller with gain scheduling that communicates with the computer using serial protocol. The rails have a range of motion of approximately six feet.

2.2 Software

2.2.1 Computer Vision

Each robot has a built-in computer vision algorithm based on either: a contrast-detecting Pyramid of Gaussians distance in pixel space, a straight-line-detecting Hough line filter, or a saturated-pixel detector (These three particular filters were chosen for their diversity--the swarm system has many common feature detectors available for potential use.)

The computer vision algorithms each process incoming video frames and determine the locations of features of interest. The sub-region within the image with the largest concentration of detected features is found. The center of this sub-region is then designated the target gaze location and used in the camera control computations.

2.2.2 Control of Camera PTZ

For each robot, given a target location in its current video frame, along with its own position along the rail and pan/tilt/zoom state, a movement command is calculated such that the robot's gaze will center on the target gaze location. In this way, the robot will proceed to "seek out" features in the space.

Zoom control follows a maximization scheme as well. The robots adjust their zoom level between maximum wide-angle and maximum telephoto in increments. As they adjust their zoom level, the number of features found in the space is compared from moment to moment. The robots attempt to maximize this number, and move towards gradually achieving a zoom level that results in a large number of detected features.

Together, these control mechanisms turn visual features in the room into attractors for the gaze of the robots.

2.2.3 Control of Rail Movement

A rotary quadrature encoder allows for centimeter-precision in camera position at speeds over one meter per second. Control of the motor speed is handled by a custom Proportional-Integral-Derivative (PID) controller. The PID controller uses gain scheduling to vary the proportional feedback constant in order to achieve very fast convergence to incoming position commands.

Commands from the main computer drive the robots in the direction of their gaze. They dolly towards the subject of their attention in order to improve resolution of their image, if possible.

2.2.4 Communication Between Cameras

Over time, the robots converge on and explore the features of regions of high visual saliency in the environment together. This is enabled by a simple target acquisition and communication protocol. When a robot detects a local maximum in the saliency of its visual environment (indicated by repeated zero-crossings in the derivative of its saliency-driven zoom level), it tells the other robots the location of its current gaze. All robots then converge on that location and proceed with their saliency-driven behavior.

2.2.5 Visualizations

Two visualizations display the information gathered by the robots. In the first visualization, a three-part screen shows the live stream from each robot's video camera. The three images have been processed by each robot's computer vision algorithm, and so each one highlights a different set of features in the visual environment.

The second visualization is a real-time, 3-D computer generated environment representing the installation space. Within this virtual environment, the robots' video streams are placed into space at a location determined by the robots' poses, focal planes, and zoom levels. The individual images making up the video streams are thus distributed throughout in the room, in their actual locations as "perceived" by the robots. The result is a constantly-evolving, fragmented visual database of the installation environment, incorporating the objects and people present within it.

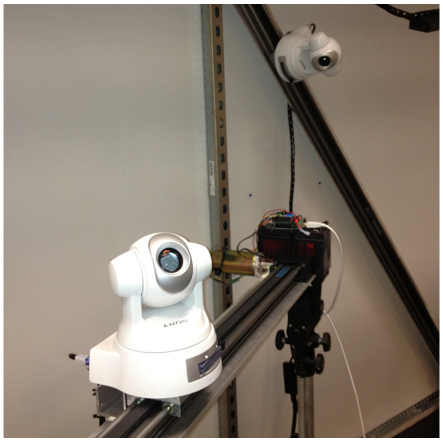

Figure 2: *A close-up taken during development showing a robotic camera in the foreground and the motor driving the rail in the background.*

3 Discussion

3.1 Platform Development

The platform layout is flexible and has been tested in different configurations. Figure 4 illustrates one possible configuration, where all camera rails are horizontal and near the ceiling, looking down upon the subjects of the scene. Rails can also be positioned on the diagonal along a wall. Custom motor controls were developed to allow arbitrary angles and to compensate for the gravity on the camera. These varying configurations create visual interest for viewers in the space, and increase the potential angles of view for the system. The system has been accepted for installation at multiple international venues during summer and fall 2013. The ongoing research follows simultaneous tracks of engineering investigation and artistic exploration, with the guiding premise that artistic and aesthetic approaches to scene analysis can yield unexpected directions into computer vision and graphics research.

3.2 Engineering Approach

It was found that the key variables in creating and evaluating emergent behaviors of the system include:

- Types and variability of computer vision techniques in evaluating features of interest;
- Methods of control and communication between the camera agents; and
- Visualization of the resultant imaging on the screen/projection.

Early work with the platform led to exploration in the construction of a “reference image” by all three robot cameras. An image was input into the system, e.g. a drawing or a photograph, and the three cameras worked cooperatively to find subjects in the viewing field with features that matched those of the reference image, such that when the cameras’ three live images were merged, there would be a clear match to the reference image. This early exploration was useful in providing an initial and defined set of goals for the robots, with a (qualitative) measure of evaluating success. This investigation was shelved for potential future refinement, but was important in informing the ongoing development of the system. It was determined that a more open-ended exploration by the robots could yield a higher chance of success in allowing interesting behaviors to emerge.

The visualization also evolved over the research period with the system. The current implementation displays a non-literal 3D reconstruction of the space, where the details and emphasis are based on the content of the subject versus simply the literal spatial layout of the room. Unexpected patterns emerge naturally in the three-dimensional representation. The ongoing work suggests that this type of content-influenced spatial visualization may provide, to humans studying the system, a more intuitive understanding of the scene analysis. This may suggest new approaches in the exploration of computer vision problems such as, for example:

- Human behavior analysis (through multi-camera communication),
- Conventional 3D scene reconstruction.

Finally, studying the impact on humans interacting in the space yielded unexpected results. Viewers watching the cameras became instantly aware that the robots were studying and analyzing the scene. Looking at the large lens of a camera, a viewer would know exactly where in the room a robot was “looking.” Just as humans are drawn to look at an area of space indicated by the direction of gaze of another human, so were viewers drawn to look where the robots were watching. Viewers in the space also desired attention from the robots, and would often move in front of the camera eye, and experience satisfaction when having been designated a subject of interest. Further, during moments when the robot cameras appeared to be looking at each other, viewers experienced a sense of non-verbal communication between those autonomous agents.

3.3 Artistic Approach

Machine learning methods based on human approaches to compositional framing can be significantly enhanced by methods from the field of fine arts photography, as those trained in visual language are able to recognize high-level visual characteristics and relationships in images and paintings that may go unnoticed to an untrained eye. A skilled photographer or visual artist mentally assembles or dissects complex images based on an understanding of the syntax of visual language. Acquired through repeated, disciplined, and conscious efforts in image creation, visual creators develop the means to analyze and identify how depth, spatial organization, form, color motion, texture, noise, uncertainty, etc (aesthetic primitives) function to articulate a visual experience, and once created, to analyze and evaluate how these elements form or fail to form an artistically coherent expression.

While much of this analysis often involves intuition, and functions at the implicit level (not unlike any expert performance such as playing tennis, chess, etc.), we argue that it is possible to identify systematic elements or rules of visual construction that are dependent on general cultural conventions. Artists arrange these elements in unconventional ways to continuously verify and expand the boundaries of the rules of visual language. These rules may serve to automate visual recognition increasing complexity in the performance of computer vision machine behavior.

While engineering research directs itself to enhancing signal quality, artistic practice tends to embrace noisy systems. For instance, whereas Photosynth will normalize and blend images towards seamlessness, our system is based on stacking images according to focus and zoom parameters, which results in a clustered form that requires and trains the viewer and the system to make sense of the mapped visual space. Through this process, we dwell into a perceptional and machine performance situation that sidesteps standard engineering work, but addresses a messy and realistic scenario that systems will generate. We initiate the process by giving the robots a set of rules such as search and capturing of frames for most salient regions. Our system then aims to center, constantly adjust zoom to maximize saliency, resulting in the dense layering of continuous images.

Public engagement offers rich opportunities for study. The robotic camera system will be installed in various exhibitions, which will become the occasion to study how viewers will interact with the system. As the system creates visual recording of its activities, participants will be seen to interact with it, and this will lead to next step developments based on the recorded data.

4 Conclusions and Future Work

The system of exploratory networked robotic cameras, guided by aesthetic techniques and rules of visual language, has begun to yield emergent behaviors. We aim to inform scientific research in computer vision and computational photography through an aesthetics- and semiotics-based approach to camera systems engineering.

The camera agents utilized in the current research project have multiple degrees of freedom in orientation (via pan/tilt) but are limited to a single degree of spatial motion along a track. Continued research will incorporate wheel-based robots on the ground, adding an additional degree of freedom to spatial motion.

Airborne drones suggest a promising direction, both in terms of features and affordability in number, despite limitations in maintaining certain orientations (pitch and roll) and overall imaging stability. Ground-based robots with vertically telescoping cameras, while expensive, allow multiple degrees of freedom in space and in orientation for highly stable imaging.

Ongoing research will investigate the integration of different types of agents with different abilities. For example, airborne drones may quickly search for potential areas of interest; then suggest, to the more stable wall-based and ground-based agents, that they position themselves appropriately and zoom in for more detailed examination. Integration of cameras with differing imaging systems will also be investigated, including thermal imaging, thus allowing investigation of a subject at both varying angles and varying spectral characteristics.

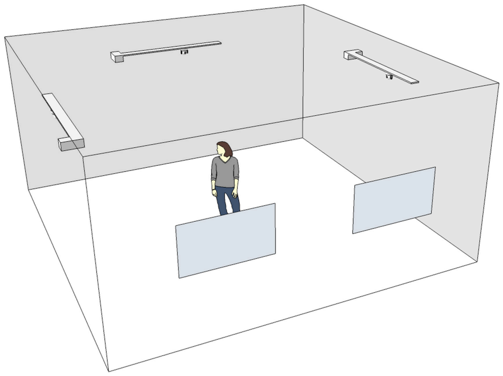

Figure 3: *Installation layout with cameras surrounding the space.*

Throughout these explorations, a primary focus of ongoing research is the study of the emergent behaviors of the system. A formal approach to the characterization of emergent behaviors will be developed. This will involve the establishment of goals for the system which are subject to quantitative analysis. Additionally, the behavioral elements of the system will be designed flexibly to allow for rapid exploration of different variables. These include: the ability of the agents to act cooperatively versus antagonistically; the ability to exert control over one another in master/slave relationships; and the amount and type of information shared between agents. As the relationships are altered through ongoing iterations, the resultant emergent behaviors can be quantified for effectiveness. With the ability to automatically quantify the effectiveness of behavioral patterns, there is the possibility of applying genetic algorithms to the behavioral variables, thus allowing the system to autonomously explore and determine the most effective methods of communication and control for imaging tasks.

5 Acknowledgements

Special thanks to our colleagues Daniel Vaquero, Ahana Reddy, Yuan-Yi Fan for contributions of computer vision and image processing algorithms, literature review, and for the development of a stable controller for robotic diagonal movement. Research on this project has been funded by the Robert W. Deutsch Foundation and National Science Foundation grant for ongoing financial support this research.

References

ELKINS, J. 1999. Interpreting Non-Art Images. In *The Domain of Images*, Cornell University Press, 31–51.

GOODMAN, N. 1976. "Notation" in the Structure of Art. Indianapolis, IN.

PETERS, G. 2007. Aesthetic Primitives. In *Information Visualization (IV'07)*, 316–325.