CV-NICS: A LIGHTWEIGHT SOLUTION

TO THE

CORRESPONDENCE PROBLEM

by

GRAYLIN TREVOR JAY

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2009

ABSTRACT

In this dissertation, I present a novel approach for solving the correspondence problem using basic statistical classification techniques. While metrics such as Pearson's ρ or cosine similarity would not be powerful enough to solve the correspondence problem directly, their performance can be enhanced by augmenting the scene with random color static via a projector. Over time, this noise increases the statistical independence of imaged points **not** in correspondence. This allows the reduction of the correspondence problem to a simple similarity search of temporal features. Extensive experiments have shown the approach to be as effective as more complex structured light techniques at producing very dense correspondence data for a variety of scenes. The approach differentiates itself from traditional structured lighting by not relying on known camera or projector geometries, and by allowing relatively lax capturing conditions. Due to the statistically oriented nature of the approach and unlike more recognition focused techniques, the approach is naturally amenable to quality assessment and analysis. This dissertation provides a background on the correspodence problem, presents empirical and analytical results regarding the new techinque, and reviews the related work in the literature.

ACKNOWLEDGMENTS

My sincere thanks to the patient and professional staff at the University of Alabama, my friends, and my family. They have all provided the motivations and distractions I needed and usually with the appropriate timing.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The success---and increasing commodization---of dedicated three-dimensional (3D) hardware acceleration has created a growing gap between users' ability to process 3D data and their ability to capture or create it. A consumer can buy, for less than a few hundred dollars, a hand held device such as a Playstation Portable or iPod Touch that can be stowed in a pocket and is capable of processing rich “arcade quality" 3D scenes. However, for the same consumer the cost of the most rudimentary laser scanner for 3D capture can start in the thousands of dollars. This gap between content processing and creation closely resembles the large divide between consumers' video viewing and capturing abilities just a few years ago. The closing of the “video gap" saw not only the further democratization of media associated with the Internet, but also the creation of thousands of formerly unforeseen applications such as YouTube. Hoping to have a similar effect on the 3D community, simple and cost-effective acquisition of 3D data is one of the driving goals for researchers in computer vision. Due in part to the closing of the aforementioned video gap, 2D imaging capabilities are practically ubiquitous. If these capabilities could be adapted to 3D acquisition as well, it would go a long way towards addressing the asymmetry between 3D processing and capture.

Stereo vision is an attempt to emulate human vision in the sense that two horizontally separated 2D views of a scene are analyzed to infer 3D information [17]. It has proven useful in such diverse tasks as navigation, foreground background separation, and facial recognition. However, it has proven difficult thus far to use stereo vision and other highly analytic methods to yield the high resolution results needed for modeling and reverse

engineering. One of many reasons for this is the inherent difficulty of the so called *correspondence problem*. Stereo vision infers its 3D information through triangulation. By knowing where a given point appears in images captured from two slightly different viewpoints, the distance from those viewpoints can be calculated. The difficulty arises from the need to know which points within the images represent the same physical point in space or, in other words, which points within the images *correspond* to one another. This general recognition task is known as the correspondence problem. It is this problem that this dissertation attempts to address directly.

Generating correspondence data is made difficult by the fact that many points within a scene are too visually similar to discriminate between (i.e. they are mostly featureless) and that many distinguishing features do not appear identical when imaged from different viewpoints (i.e. are not viewpoint invariant).

In some applications such as motion tracking, use of sparse correspondence data is feasible. Either sparse data itself suffices for the application at hand or it can be utilized by making generalized assumptions regarding scene geometry. In these cases, the problems of featureless and view variant surface points are less prominent. Features that are identifiable and view invariant given certain conditions (e.g. after fast Fourier transform) such as edges and corners can be matched between views. This provides a small number of high quality correspondences. However, this dissertation is concerned with the correspondence problem itself and thus with *dense* high quality correspondence data. For collecting this kind of data, few approaches in the literature have proved as successful as structured light. In fact, one of the more prominent uses of structured light systems in the literature has been to assess the effectiveness of stereo vision algorithms [35].

Structured light mitigates the problems of featureless view varying surface points by physically projecting onto those surfaces patterns pre-constructed to have features view invariant under most of the transformations of both projection and imaging. With geometric knowledge of the single camera and projector, the traditional structured light techniques reconstruct a virtual view from the projector's perspective as if it were itself a camera. Both the actual camera image(s) and the virtual image(s) created from the projector's view contain elements of the pattern that was projected onto the scene. The elements being feature-rich and view invariant are easy to automatically recognize. It is this labeling of the image(s) with automatically recognizable elements that allows for a straight-forward solution to the correspondence problem. Points between images labeled by the same pattern elements are assumed to be in correspondence.

While effective, traditional structured light techniques have some features that make them inconvenient for some applications. The need for geometric camera and projector knowledge as well as, often, a large number of input images make many structured light techniques too heavyweight for situations requiring more casual data collection. As described, most structured light techniques rely on the automatic recognition of specific pattern elements. This recognition sub-task is often accomplished by treating captured images as the input to a decoder. For example, a temporal pattern element might take the form of black and white dots that can be decoded as a unique binary number. The quality and accuracy of this kind of discrete decoding process can be difficult to assess, rendering the entire process difficult to analyze.

In this dissertation, I present a novel structured light technique which attempts to address these issues. I call the technique *correspondence via noise informed corollary statistics* or *CV-NICS*. Traditional structured light techniques rely on the recognition of

embedded elements such as m-arrays or binary codes. CV-NICS is novel in being a purely statistical approach.

I have re-framed the decode and label sub-task of traditional structured lighting as a statistical classification problem to derive a new multi-camera approach that has more lax capturing requirements and is open to a traditional machine learning analysis.

Extensive experiments have shown the approach's effectiveness to be on par with much more complex structured light techniques, despite its straight-forward implementation. figure 1.1 compares CV-NICS output to that of several traditional structured light techniques in the literature. CV-NICS produces data of comparable density and quality.

Fig. 1.1 Point clouds of hands as generated by several traditional structured light techniques. The first seven images are from the survey paper by Salvi et. Al. They were generated by: time-multiplexed binary coding, time-multiplexed phase shifted gray code, time-multiplexed n-ary code, two forms of De Bruijin spatial neighborhood coding, m-array spatial neighborhood coding, and direct color coding techniques respectively. The last image is the author’s hand and was generated via CV-NICS.

In the traditional approach, points within views are joined by a third entity, an a priori known pattern element. In the CV-NICS approach, points within views are instead joined *directly* by their statistical properties. Specifically, they are joined by their temporal dependence as measured by the cosine similarity of their colors over time. Rather than using

projected patterns to inject specific features, they are now used to increase this temporal independence between different physical surface points. This is accomplished by projecting random color noise or static.

This focus on independence instills a number of advantages. The ability of the color static to increase independence between points is mostly unaffected by the transformations of projection/imaging and is less effected by the frequency shifts inherent in being overlaid onto color surfaces than traditional patterns. This allows CV-NICS to use fewer sample images than many other structured light systems and completely eliminates the need for known camera or projector geometry.

This dissertation will introduce and evaluate CV-NICS in the context of the correspondence problem. Chapter 2 will introduce the correspondence problem and this dissertation's formalization of it. Chapter 3 will review the most closely related work in stereo vision and structured light. Chapter 4 will introduce CV-NICS itself. Chapter 5 will present an analytical and empirical analysis of the technique and my current results. Chapter 6 is a brief diversion into the 3D reconstruction methodology used throughout this document. Chapter 7 concludes the dissertation with a high level discussion of CV-NICS. All of the code associated with this disseration is available on-line at: http://code.google.com/p/cvnics .

CHAPTER 2

THE CORRESPONDENCE PROBLEM

To best describe the correspondence problem, some imagining terminology must be introduced. I thus begin by briefly introducing the pinhole camera model [28].Conceptually, a pinhole camera is simply an opaque box (see figure 2.1). The inside back of the box (relative to the scene or object of interest) is coated in some sort of photosensitive material. A pinhole is then punched through the front of the box, which faces the scene or object of interest. Because the box is opaque, any light striking the photosensitive material must necessarily have passed through the pinhole. It is this fact which creates the primary geometric property of the pinhole camera: light striking the back of the box had to have traveled along a straight line defined by the struck point and the pinhole.

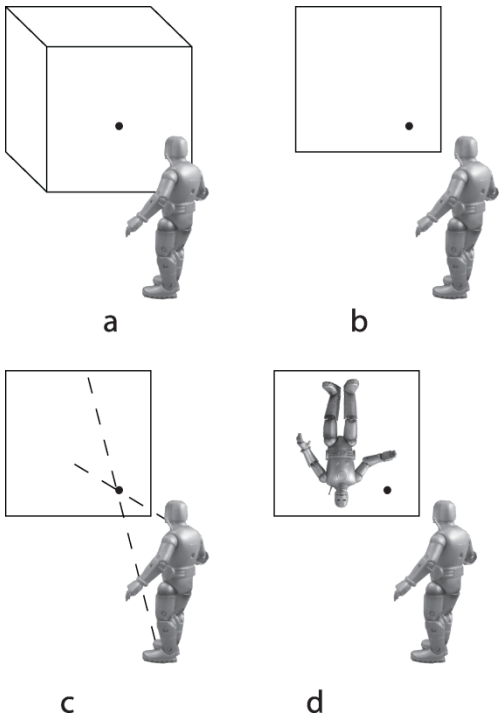

Fig. 2.1 Pinhole camera model. a) the camera consist of an opaque box with photosensitive material coating the inside back and a pinhole punched through the front. b) For illustration purposes, all but the back inside and pinhole have been made invisible. c) Light coming into the box and striking the photosensitive back *must* come through the pinhole, forming a straight line. d) The colors of the first surfaces encountered along these lines are recorded on the photosensitive material.

If we ignore light sources and consider only light reflected off non-transparent surfaces, we can use the pinhole camera model to construct a rather unorthodox definition of a 2D image: a set of 3D vectors augmented with the color of the first surface encountered along each vector.

This definition suggest both the challenge of 3D reconstruction from images and a possible solution. The challenge arises from the infinite nature of the vectors. If a given color

associated with a vector is blue, it simply indicates that there was a blue surface **somewhere** along that particular vector. Where exactly (and hence where in 3D space a surface point occurred) is not captured by the imaging process. However, the nature of the 3D vectors suggest a possible solution to this lack of information: multiple images.

Two vectors will intersect at a single point at most. If we employ a second pinhole camera to image the same scene, it may capture some of the same surface points as the first camera. If it does, we can reconstruct where in 3D space those surface points are because the camera-one vectors capturing those points intersect with any corresponding camera-two vectors at a single point in calculable point in 3D space. Obviously, this reconstruction relies on a knowledge of the geometry at hand. The geometries of the two cameras must be known as must their positions in so called *world coordinates*. In the literature, discovery of this information is known as: *camera calibration* and cameras for which this information is defined are known as *calibrated cameras* [43]. Besides camera calibration, this reconstruction process relies on another more subtle piece of information: a mapping function between the vector sets of the two images.

Lots of inter-camera vectors intersect with each other. For 3D reconstruction, vector intersections are only interesting if the colors associated with their involved vectors happen to originate from the exact same surface point. We call such pairs of vectors corresponding vectors. Finding such pairs is one form of the correspondence problem but such a formulation is of limited use, as we are interested in the correspondence problem as an image analysis problem. Therefore, we now relate these vectors to points within 2D images.

Any camera (including a pinhole camera) is a 3D to 2D mapping function, $c(.)$, that takes in a viewpoint and a real world coordinate (x, y, z in $\mathbb{R}^3$). A viewpoint, $\boldsymbol{v}$, is a real world location (x, y, z) and an associated orientation ($\boldsymbol{o}$). For a set viewpoint, $c(.)$ acts as a

mapping between $\mathbb{R}^3$ and an image space, $\mathbb{I}_v^2$, uniquely identified by the said viewpoint. This mapping is destructive. Multiple x, y, z's result in the same u, v. When we map from $\mathbb{R}^3$ to a $\mathbb{I}_v^2$ we call the x, y, z in $\mathbb{R}^3$ the *originating* point. As in the vector description, a projector can be considered an inverse case of a camera. Thus a projector can be considered a function $p(.)$, that maps from $\mathbb{I}_v^2$ to $\mathbb{R}^3$.

An image is a finite subset of an image space. Assume two images A and B resulting from the same camera function, $c(.)$, being applied to the same world space $\mathbb{R}^3$ from two different viewpoints $\boldsymbol{a}$ and $\boldsymbol{b}$. It may be the case that some point in A has the same originating point as a point in B. That is, there exist some point, p, in $\mathbb{R}^3$ for which: $c(\boldsymbol{a}, p) = A_{u,v}$ and $c(\boldsymbol{b}, p) = B_{u',v'}$. We say that two such image points with the same originating point are in *correspondence*. For a given pair of images, finding all such mappings ($A_{u,v} \rightarrow B_{u',v'}$) is solving the *correspondence problem*. We describe an incomplete solution as *dense* if it covers a large percentage of the mappings and the mappings it covers are clustered together in the image space. Conversely, an incomplete solution is *sparse* if it covers only a few mappings separated by large areas of image space.

However, to complete the introduced model it must be extended to account for the optical properties of the surfaces involved. For the purposes of analysis, the scene's global illumination is broken down into sub-functions associated with a given point x, y, z in $\mathbb{R}^3$. Each x, y, z thus has an associated function $l_{x,y,z}(.)$ that describes how it interacts with light. Each $l(.)$ takes in: a viewpoint $\boldsymbol{v}$ (consisting of a point in $\mathbb{R}^3$ and an orientation vector), an incoming intensity i_i, a wavelength λ, and an incoming orientation o_i. $l(.)$'s return the perceived intensity i_p of the wavelength in question at the given viewpoint. To construct a complete picture of the intensities a given viewpoint perceives, the appropriate $l(.)$ must be

evaluated for all wavelengths, orientations, and incoming intensities involved (i.e. for all light sources). This characterization of the $l(.)$'s ignores prismatic effects, but these are rare in non-transparent surfaces.

For points in two images to be in correspondence means that they image the same physical point. However, the imaging process makes it difficult to find correspondences in the naive manner of mapping image points directly back to their unique originating physical points. Physical points in $\mathbb{R}^3$ are always unique even if only because of their unique location, but this very location information is destroyed by the mapping process of $c(.)$. Perceived surface properties are of limited use in distinguishing between points, because distinct points may have identical physical properties. Augmenting a scene with projected patterns must be carefully attempted. Such patterns are distorted by the projection *and* imaging processes of both $p(.)$ and $c(.)$.

While these and other factors certainly make the correspondence problem difficult, it is by no means insoluble. In the next chapter we will briefly review a family of techniques that have had particular success: structured light techniques.

CHAPTER 3

RELATED WORK

One solution to the dilemma of the correspondence problem is a class of traditional structured light techniques I call *labeling* or *code and recognize* techniques [14]. In them, a projector (now often an LCD projector) is used to display a pattern on the surfaces of the scene. The scene is then imaged by two or more cameras as usual. The pattern is designed to have automatically recognizable sub-elements. Being recognizable in the image created by any given camera, these sub-elements serve as labels that link corresponding points from different images together and allow for 3D reconstruction. Most structured light techniques ignore real surface features, focusing on finding and comparing the features known to have been introduced.

To server this purpose, the features introduced in the scene need to be useful post the projection and imaging transformations of the $p(.)$ and $c(.)$ functions described in chapter 2. While lossy, the $p(.)$ and $c(.)$ functions do preserve some geometric relationships, especially those in the x, y-plane relative to the camera. Much of the structured light literature assumes known $p(.)$ and $c(.)$ functions and concentrates on constructing patterns that can be recognized based only on these preserved relationships.

With no loss of generality, the image variables examined in Chapter 2 can be considered *sets* of images. This allows for accounting of temporal scene changes and the use of temporal pattern approaches such as time-multiplexing. These types of patterns have the advantage of being immune to any perturbation by $p(.)$ or $c(.)$, but have the disadvantage of requiring more images be taken as they consist of multiple pattern frames. Because of these

disadvantages, hybrid pattern approaches combining spatial and temporal approaches are common.

To know a priori a characterization of $p(.)$ or $c(.)$ is known within the structured light parlay as having a calibrated camera or projector [4]. Many structured light techniques take advantage of such calibrated knowledge by a-skewing a multi-camera (or camera position) approach for a virtual camera approach [42]. Here, correspondences are calculated not between different camera viewpoints, but between a camera viewpoint and a reconstruction of the view of the projector.

Regardless of pattern technique or whether a structured light technique uses a multi-camera or virtual camera approach, the mechanics of determining correspondences are quite similar. All of the recognizable pattern elements within a view are decoded into unique identifiers. The same decoding process is undertaken for another view. Treating the unique identifiers as spatial labels, those points between views sharing an identifier are considered to be corresponding points. This approach I call labeling. The CV-NICS process is very different from this labeling approach. The element of interest for CV-NICS is not a priori known unique identifiers of the traditional approach but the surface properties of the scene itself.

In general, the literature makes little effort either to exploit or account for the surface properties of the items in a scene subsumed in our model by the function family $l(.)$ (see Chapter 2). Most pattern approaches simply use extreme color values combined with thresholding to mitigate the issue.

3.1 Overview of Pattern Techniques

With the exception of some work we will cover concerning dynamic and multi-stage analysis, most of the literature on structured light is concerned with the optimal encoding/decoding strategy for creating the pattern with automatically recognizable sub-elements. An excellent introduction to many of the topics of this section is the survey paper by Salvi et. al. [32], which can be consulted for further information.

Salvi and other categorize structured light techniques by pattern, identifying three main categories of patterns: *Direct Coding*, *Spatial Neighborhood*, and *Time-multiplexing*. Obviously, time-multiplexing can make use of temporal encoding, though there exist temporal versions of all of the techniques. All of the coding schemes are usually applied using a virtual camera.

Direct codes usually encode a semi-continuous property such as a point's u, v coordinate in the pattern as another easily projected semi-continuous property such as intensity from pure black to pure white [1]. While theoretically primitive, these encodings do possess some desirable properties. Foremost of these is that points with close properties have close encodings. This can mean that even erroneous recognition can still be useful [9].

Spatial Neighborhood uses complex graph theoretic properties to generate 2D patterns with uniquely identifiable “neighborhoods" [2,8,11,13,16,26,27,37,38]. These neighborhoods are made of smaller elements such as dots that are not themselves uniquely recognizable. To identify one of the sub-elements requires looking at its immediately surrounding elements. Often these patterns employ a finite number of colors.

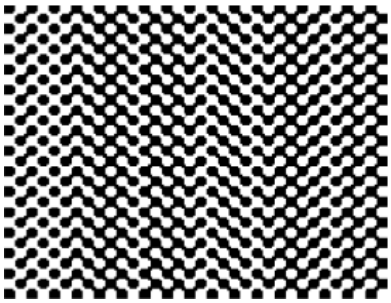

Fig. 3.1 A spatial neighborhood De Bruijin code. Each circular area has a unique combination of black and white neighbors. This allows the space around an area to uniquely identify it.

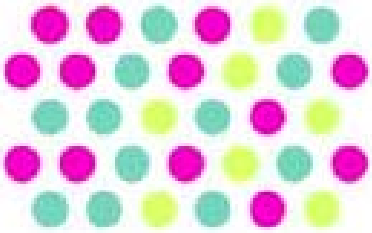

Fig. 3.2 A spatial neighborhood M array code. Here, M= 6. Any given dot is surrounded by a unique combination of colors. As with the De Bruijin codes, this allows the space around an area to uniquely identify it.

Time-multiplexing, as its name implies, makes sub-elements recognizable via a series of changes over time (i.e. between multiple images) [18, 24]. For example, sub-elements can be turned black or white over a period of time such that each sub-element represents a binary number. As the number of images can be extended somewhat arbitrarily, Time-multiplexing is the least constrained of the techniques in the amount of information it can encode. As

mentioned, the other encoding techniques can be extended to make use of time, but many forms of multiplexed patterns do not have a static analogue.

3.2 General Issues with Structured Light

The largest limitation of traditional structured light techniques is that they do not completely eliminate the need for some form of automatic recognition. While recognizing and decoding explicitly encoded patterns projected onto a scene is certainly a more closed recognition task than the actual correspondence problem, it is **still** a recognition task with all of the computational difficulty that implies. As a consequence, there is usually a limit to how small the elements of a structured light patterns can be made while still being reliably recognizable. Unfortunately, it is often this factor rather than camera or projector resolution that limits the resolution of structured light based reconstructions.

Related to this issue is the fact that most scenes/objects that one might want to reconstruct are usually made of surfaces that are far from ideal screens for the projections used by structured light. Most surfaces perform (at least) linear transformations on the light that reaches them. This further limits the fidelity structured light patterns can bring to bear. The work in this dissertation seems to indicate it is this “surface-distortion" that creates the observed [1] problems of direct coding.

3.3 Advanced Techniques

Due to their high quality reconstructions, I will here give a slightly more in-depth examination of three techniques from the literature: those of Scharstein, Zhang, and Fofi.

Scharstein's technique involves a calibrated camera and a projector with unknown geometry [36]. He uses a series of binary time multiplexed codes chosen so that close sub-elements share a close encoding (a so-called *gray code*). Using images of the gray codes obtained from two camera positions, he produces a 3D reconstruction of the scene. Comparing this reconstruction to his obtained images, he is able to reverse engineer the projector geometry. He is then able to apply a virtual camera technique to yield two more reconstructions (one for each camera and projector combination). He applies averaging and heuristics to combine his previous three reconstructions into a final 3D reconstruction.

To lessen problems with his patterns interacting with scene surfaces, Scharstein takes his images under four different conditions: the pattern series with exposure times of .5 and .1 seconds, and the inverted pattern series at exposure times .5 and .1 seconds. When it comes time to recognize individual pattern elements, Scharstein is thus able to choose the conditions that maximize the binary illumination differences at a per pixel level. This allows for the mitigation of almost any effect of surface $l(.)$'s. While this approach undoubtedly contributes to the high quality nature of his results, it also adds a great deal of procedural overhead to his technique. Besides needing a fully calibrated camera at up to 80 images are needed. Figure 3.3 shows a capture session.

Fig. 3.3 Scharstein's gray codes at work. This picture shows a typical structured light capture session. The scene is illuminated via a pattern from an LCD projector and imaged by a normal 2D camera from one or more viewpoints.

Zhang was among the first in the literature to realize the inherent limitations of the code and recognize approach. While his *Space-time Stereo* technique employs time-multiplexed color gray codes, it does not attempt to decode them into labels directly. Instead, he uses a multi-pass dynamic programming analysis to minimize the difference between the images predicted by a (iteratively refined) 3D reconstruction and the images actually observed by two cameras [42]. Lavoie [23] applies a similar analytic step.

In a giant step towards making traditional structured light techniques viable without calibration, research by Fofi and Furukawa [5, 6] uses post-capture analysis to allow for accurate 3D reconstructions from sparse 2D correspondence where the capture geometry is unknown. We consider this work highly complementary to our own as we facilitate the collection of **dense** correspondence data without calibration.

3.4 Stereo Vision

It is possible to consider CV-NICS a stereo vision technique. It is most strongly related to the early foundational stereo vision work as it utilizes basic similarity metrics (Pearson's ρ and cosine similarity) and a simple 2D search window approach. Stereo vision in general is concerned with features invariant to the viewpoint dependent imaging process $c(.)$. Structured light is usually distinguished from stereo vision by its dependence on projected patterns, and by its concentration on a priori known rather than natural scene features. CV-NICS features the former but not the latter. However, the use of projected patterns is an important distinction, and I believe CV-NICS is thus best classified as a structured light technique and not a stereo vision technique. In any case, that is the viewpoint I have chosen to take in this dissertation for the purposes of a self-contained analysis with familiar tools. I mention stereo vision here for the sake of completeness.

CHAPTER 4

CV-NICS

Attempting to reproduce many of the techniques I have discussed, I quickly became motivated to investigate techniques that would be simpler to implement and would be more forgiving of capturing conditions. As mentioned, most of the popular techniques in the literature required a pre-calibrated camera. This is fine from an academic perspective as camera calibration is its own research and engineering problem. However, from a laboratory perspective this meant that to actually use these techniques required a long, drawn-out, and tedious calibration process [43]. Those techniques which do exhibit some form of self-calibration often require taking so many additional images that for short capture sessions little or no labor is saved over having a calibration stage. As for actual programmatic implementation, most structured light techniques are not overly complicated but proper recognition, encoding, and decoding of the patterns involved is still far from a trivial programming task.

I began to examine the source of complexity in pre-existing techniques. The most obvious was camera calibration. Obviously, some kind of camera calibration is required to reconstruct geometry from correspondences but this is not why most pre-existing techniques require camera calibration. As I will show, subjectively good 3D reconstructions can be produced with almost no knowledge of camera geometry. Most structured light techniques require camera calibration not for accurate reconstruction but for accurate recognition. Recognition of (for example) spatial neighborhood codes requires that a certain degree of the

pattern distortion created by the act of projection be reversed. This requires camera geometry knowledge.

The complexity of the programs involved in structured light techniques is, like the need for camera calibration, tied to the fact that a recognition task is required. Pattern recognition is---of course---its own research problem and can be quite complicated.

Based on this, I decided to investigate tackling the correspondence problem directly instead of relying on an intermediate labeling. This would allow me to skip the implicit recognition task that labeling required. I was familiar with co-linearity and that such close correlation was one operational definition of identity. I decided to investigate how well a simple statistically driven structured light approach might perform. As a result of this investigation, I have developed CV-NICS.

It should be noted that I am not the first to attempt to build a structured light system based purely on summary statistics. In fact, many of the authors I have already cited (most notably Zhang) investigated the idea before dismissing it as unworkable. However, unlike previous systems CV-NICS seems to work well in practice [31], its performance seeming to derive from its exploitation of local illumination (see chapter 5).

CV-NICS is a hybrid process for solving the correspondence problem combining aspects of structured light and statistical classification. It utilizes a single camera and projector. The process can be roughly divided into three stages.

4.1 Capture

This stage is almost identical to a traditional structured light approach. A scene or object of interest is prepared. The projector is positioned as close as possible to the scene while retaining the ability to both focus and completely cover the desired details. Two camera positions (viewpoints) $\boldsymbol{a}$ and $\boldsymbol{b}$ are chosen, with the restriction that they are roughly coplanar with the projector. With the camera in position $\boldsymbol{a}$, the projector is set to display each of ten pattern frames in turn. The camera captures one image for each frame displayed. This process is repeated with the camera in position $\boldsymbol{b}$ and results in two image sets $\boldsymbol{A}$ and $\boldsymbol{B}$. Separating CV-NICS from other structured light techniques is that the pattern used during this stage is in some sense completely unstructured and consist of color static (see section 4.5). Figure 4.1 shows two image sets, one per row. Each row of pictures was taken from a slightly different viewpoint. Each column's images were captured under the same lighting conditions, illuminated by one of the 10 CV-NICS patterns.

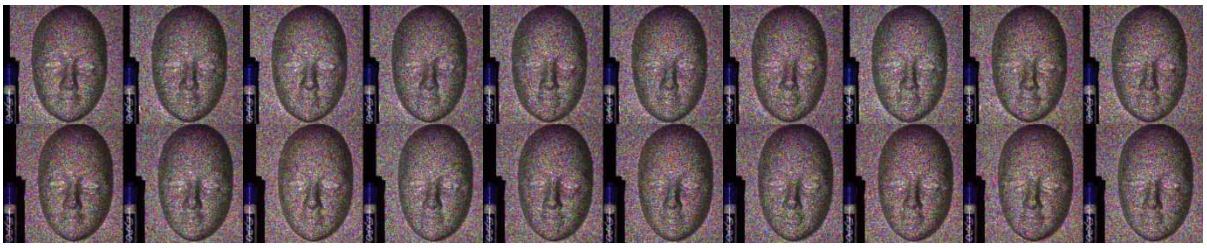

Fig. 4.1 A complete dataset for a CV-NICS capture session.

4.2 Search

We call each point of interest in $\boldsymbol{A}$ a *target*. For every u, v in both $\boldsymbol{A}$ and $\boldsymbol{B}$ there is an associated vector defined as the values of the three-channel color intensities that occur at that point. If $R_{u,v,1}$ indicates the red intensity in the first image of the set occurring at u, v, $G_{u,v,4}$ indicates the green intensity in the fourth image, and so on. Then the vector associated with u, v is:

$$[R_{u,v,1}, G_{u,v,1}, B_{u,v,1}, R_{u,v,2}, G_{u,v,2}, B_{u,v,2}, \dots, R_{u,v,10}, G_{u,v,10}, B_{u,v,10}]$$

For each target in $\boldsymbol{A}$, a search is conducted comparing its associated vector with every vector associated with the points in $\boldsymbol{B}$. The comparison is made by way of cosine similarity and the most similar vector is declared the target's *match*. The point associated with this matching vector, u', v', is considered to be in correspondence with the target point. This process is repeated for every u, v of interest and a $u, v, \rightarrow u', v'$ mapping is thus created, solving the correspondence problem.

4.3 Assessment

Common to all structured light techniques are the problems of occlusion and poorly reflective surfaces. Viewpoints being distinct, some surface points will not actually be visible in both images. Other surface points, such as those in shadows, may not image the projected patterns well and will have little temporal variance. Attempting to match such points will lead to erroneous matches. Rather than detect these situations beforehand, a threshold for

similarity is established and mappings generated in the last stage form vector pairs with low similarity are removed from the data set. See section 4.5 for details.

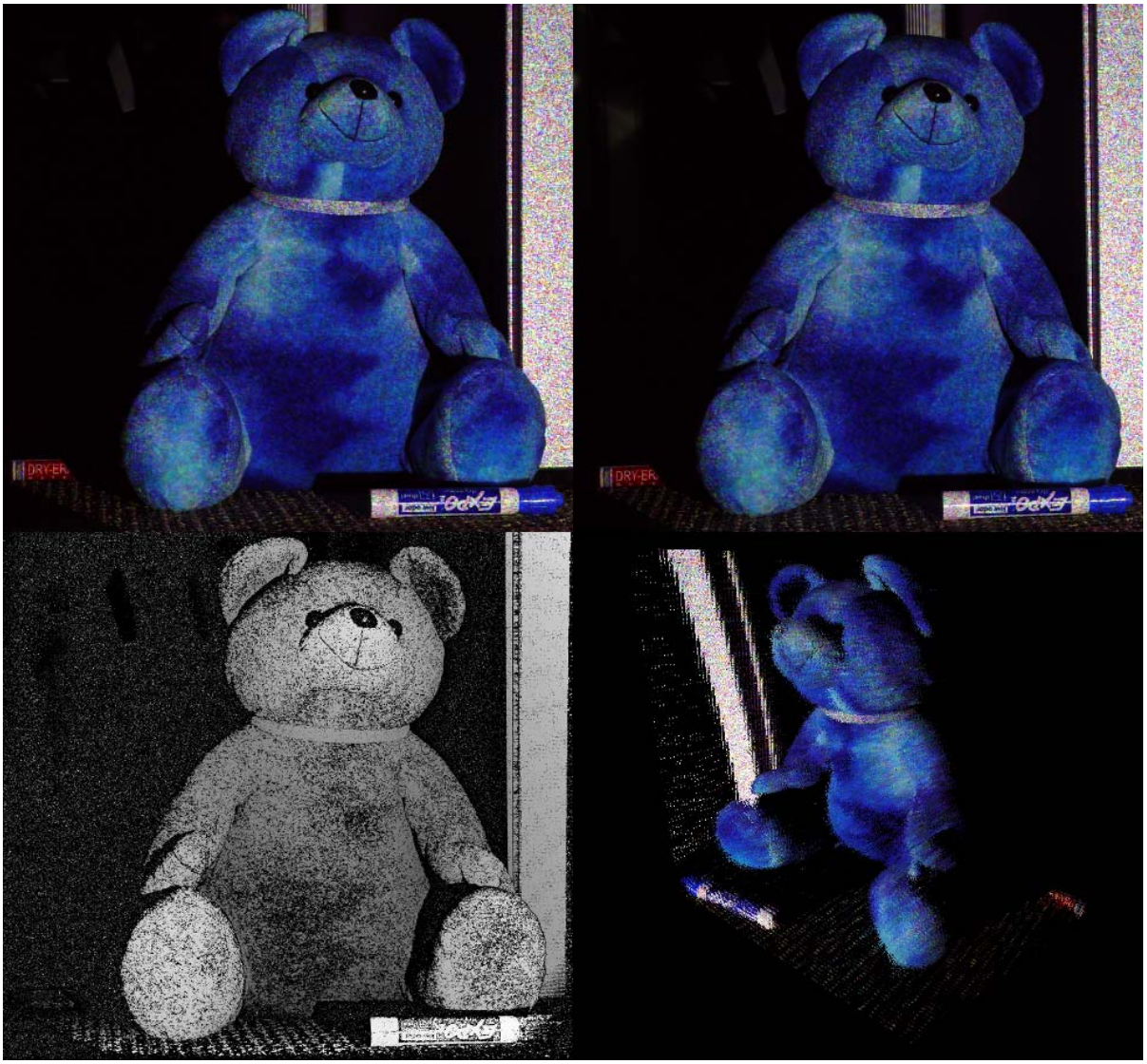

Fig. 4.2 Images representing different points in the CV-NICS process.

Images from several points in the CV-NICS process are shown in figure 4.2. The top two images were each taken from a slightly different viewpoint but while the same CV-NICS pattern was projected onto the subject, a plush toy. The bottom left image is a disparity map representation of the correspondence data generated. The bottom right image is a simulation

of the scene as viewed from an angle from which the camera was never actually placed (see Chapter 6).

The actual implementation of CV-NICS (see section 4.5) makes several optimizations, but the ideas presented above remain unchanged. Such a straight-forward procedure may seem unlikely to produce *any* usable correspondence data at all, but extensive experiments have shown just the opposite. CV-NICS is indeed capable of producing dense high-quality correspondence data under robust capturing conditions. Our probabilistic analysis (see section 5.1), suggest some probable causes for this surprising effectiveness and we discuss others in Chapter 7.

4.4 User Interaction and Optimizations

As will be discussed in section 5.1.3, the search window size plays a crucial role in CV-NICS accuracy. More important practically, it is the dominate factor of runtime, as it dictates the number of comparison operations needed per point of interest. The current implementation allows for the search window to be constrained by allowing the user to manually provide the correspondence pairs for the closest and furthest points of interest. The user simply identifies a common point in both images and provides the (image) coordinates of each. This constrains the size of the 2D search window, and in practical use has decreased CV-NICS runtime exponentially.

The current implementation provides one additional user guided optimization. The user can use normal image editing software to annotate an image from either viewpoint set. This is done by simply coloring over the existing image. Only the annotated points will be

search for correspondence pairs. This mechanism can be used to speed up the CV-NIS process when only a portion of the captured image is of interest.

4.5 Tuning Parameters and Implementation Specifics

The color static pattern used in the experiments was created by first assigning each projector pixel a random number. To generate the pixel value for the i^{th} image, the $3i^{th}$ through the $(3i+2)^{th}$ bits were multiplied by 255 and assigned to the R,G, and B color channels. The random numbers were generated modulo 2^{30} from a pseudo-Gaussian population with both standard deviation and mean of $\frac{2^{30}}{2}$.

Chapter 5 will discusses how an idealized version of CV-NICS would match on continuous wavelengths instead of the intensities of discrete sub-bands (i.e. RGB color). As an attempt to better approximate this ideal and to avoid having to use arbitrarily large number representations within the cosine similarity search, the current CV-NICS implementation does not perform assessment (see section 4.3) using cosine similarity, but instead Pearson's ρ and standard variance on grayscaled versions of the images. The actual search is still performed with cosine similarity (with Java double precision), only the assessment stage is carried out using Pearson's. Any match whose points have a temporal variance lower than 10 (in pixel intensity) or a correlation with its match of less than .8 is removed from the data set. Another reason Pearson's is used at this point is that it gives the filtering parameters a more intuitive meaning than the spatial cosine metric.

A five mega-pixel Sony DSC-F707 was used for all imaging. The projector used was a 1024 x 768 Toshiba TLP-XD2000U XGA.

The full CV-NICS source code is available on-line at:

http://code.google.com/p/cvnics .

CHAPTER 5

RESULTS

If CV-NICS were simply an effective structured light technique, it would not be particularly interesting from an academic point of view. However, being statistically based CV-NICS yields to several analysis methods difficult to apply to other structured light methods. This chapter presents one such analysis: a probabilistic bounds on CV-NICS performance. It also presents empirical evidence of CV-NICS effectiveness and some qualitative but still empirical observations

5.1 Analysis

We will now analyze CV-NICS from a machine-learning/statistical-classification perspective. We begin with a re-examination of some of the physical details.

5.1.1 Surface Properties and Illumination Constraints

We describe the illumination properties of a surface at point x, y, z as an arbitrarily complicated (non-linear) function $l_{x,y,z}(.)$ parameterized by a viewpoint and wavelength of interest (see chapter 2). Information about light sources (distance, angle, etc.) is encompassed in an $l(.)$ and is not independently modeled.

One way of framing the CV-NICS search process is that given a target point x, y, z and its associated $l_{x,y,z}(.)$, CV-NICS searches for the x', y', z' whose associated $l_{x',y',z'}(.)$ is

most similar. Given such a characterization, a loose upper bound on CV-NICS's spatial accuracy is the point at which it becomes impossible to distinguish one $l(.)$'s from another.

$l(.)$'s capture both a point's surface and illumination properties. Surface properties are not necessarily unique. Different physical surface points may be made of the same material, have the same surface normals, etc. However, illumination is a product not just of the relevant light sources but distance and angle to both those sources and the surfaces that might reflect light from those sources. Illumination is thus a spatial property. Points with non-identical locations cannot in principle have completely identical $l(.)$'s.

While the fact that no two physically separated points can have identical $l(.)$'s is physically true, such distinctions may well be beyond our process of characterizing an $l(.)$ by sampling its behavior according to cosine similarity from a particular viewpoint over a finite number of changes to the light source wavelengths. To our chosen analysis and equipment, it is very likely that very close surface points will seem to have identical $l(.)$'s. For the purposes of analysis, we thus use the more conservative assumption:

$l(.)$'s that do not share light sources cannot be identical.

Assuming we regard each pixel of the projector as a light source, a key question becomes: How much physical surface space does such a light source effect? This space is in some sense the maximum area within which $l(.)$'s cannot be distinguished from one another. The size of this area will be scene, projector, and camera specific, but is measurable. Figure 5.1 shows such a measurement. It is the result of combining two images of a paper screen captured with the camera and projector one meter distant. The superimposed grid represents the projector resolution (with roughly 100 camera pixels per projector pixel). The two patterns projected where 1) a completely black pattern and 2) a pattern with a single solid

green pixel. Gray marks those pixels which were changed by 5% or more by the introduction of the full green pixel.

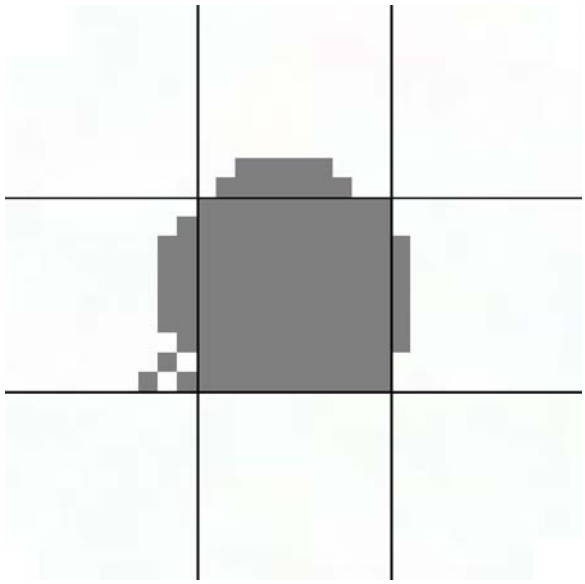

Fig. 5.1 The effects of a single projector pixel on its surrounding pixels. The superimposed grid represents the projector resolution, each projector pixel taking up approximately 100 camera pixels. Gray marks those areas whose color values change by 5% or more when the center projector pixel is fully illuminated.

The capturing conditions of figure 5.1 are almost ideal for light leakage. The surface involved is unusually reflective (white paper) and perfectly parallel with the projector, the projector and camera are unusually close, and the camera (like most commercial cameras) is more sensitive to the green color band than any other. Given how favorable these conditions are, it would be extremely unexpected for an $l(.)$ to have properties identical to an $l(.)$ separated by even one projector pixel. However, in the interest of a conservative analysis, this $+/-$ single projector pixel ambiguity is assumed.

In the next section we will examine the implications of this assumption to overall CV-NICS accuracy.

5.1.2 Conservative Accuracy Bounds

Due to the use of temporal patterns, in CV-NICS the wavelengths of the light sources change over time. Thus, over the entire capture process, an x, y, z points is associated not with a single $l(.)$ but many. We drop the concept of an $l(.)$, and replace it with an L: a random variable of an unknown distribution of wavelength combinations. As with the individual $l(.)$'s, each surface point x, y, z has an associated $L_{x,y,z}$ that is unique up to $+/-$ one projector pixel (see section 5.1.1).

Provided that surfaces lack prismatic properties (see section 7.1), observations of the wavelength *makeup* of a particular instance of an L at a specific time are viewpoint neutral. The individual intensities of the wavelengths observed will vary, but their presence or absence will hold except in cases of extreme detector distance and insensitivity. Conventional cameras do not record arbitrary wavelengths but intensities within defined bands (i.e. RGB), so in practice CV-NICS is an approximation of this ideal (again, see section 7.1).

We now characterize the probabilistic performance of CV-NICS as a statistical classifier. Given a target point u, v from an image set $\boldsymbol{A}$ and m candidate points from another image set $\boldsymbol{B}$, CV-NICS will hypothesize a particular candidate $\tilde{c}$ as being physically identical (representing the same $\mathbb{R}^3$ x, y, z).

CV-NICS picks a $\tilde{c}$ by examining each candidate point's associated L for cosine similarity with the L_t associated with the target point. The most similar candidate c is chosen as $\tilde{c}$.

We define $\theta(L_t, L_1)$ as the cosine similarity between vectors of samples from the L_t and L_1 random variables. We call $|E(\theta(L_t, L_1)) - E(\theta(L_t, L_2))|$ the θ-gap between L_1 and L_2. Of course, we do not have direct access to θ-gaps directly and must use estimates. These $\hat{\theta}$-gaps are based on a samples of finite vectors. $|\theta\text{-gap} - \hat{\theta}\text{-gap}|$ is the error on an estimate of a θ-gap, and we denote this quantity as ϵ. If the $\hat{\theta}$-gap between two L's is ϵ or less, we say they are ϵ-equivalent. We denote ϵ-equivalence with the $\approx$ operator.

We assume that the search window (the m points from $\boldsymbol{B}$) of candidate c's actually contains the corresponding point, which we denote $\overline{c}$. W can calculate the probability that CV-NICS will pick a $\tilde{c}$ that is ϵ-equivalent to this point.

Let us examine an $m = 2$ case with c's c_1 and c_2, associated L's L_1 and L_2, and an observed $\hat{\theta}$-gap between c_1 and c_2 that is greater than ϵ (otherwise both c_1 and c_2 would be ϵ-equivalent to each other).

In the case described, CV-NICS will choose c_1 as $\tilde{c}$ over c_2 because of the positive $\hat{\theta}$-gap between c_1 and c_2 suggest that $\theta(L_t, L_1) > \theta(L_t, L_2)$. The probability that c_1 is the wrong choice of $\tilde{c}$ is the same as the probability that the observed $\hat{\theta}$-gap is a complete overestimate, that the true θ-gap between c_1 and c_2 is zero or less.

To calculate the probability of completely overestimating the θ-gap, we first we observe that:

$$L_t \cdot L_1$$

is proportional to $\theta(L_t, L_1)$ when the L's involved have unit second order moments. We note that this equality could be enforced (by normalizing the data), but for analysis purposes it is here assumed. We further note:

$$L_t \cdot L_1 \equiv \sum_{i=1}^{3n} L_{t,i}\, L_{1,i}$$

by the definition of the dot product. Recall that each L vector contains a value for each color channel, and hence there are $3n$ components for the n sample images.

Hoeffding's inequality states that the likelihood of overestimating the sum of a set of n random variables by at least δ is less than or equal to:

(5.1)
$$2e\text{^} - \frac{2\delta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

where b_i and a_i are the bounds of the relevant random variable.

Substituting our derived summation into 5.1, we see that the probability of making an error of magnitude δ or larger when estimating the cosine similarity from a finite random sample is less than or equal to $2e^{-6\delta^2 n}$. However, we are not interested in arbitrary values of δ. We are interested in the case where δ is large enough to effect the CV-NICS choice of $\tilde{c}$.

Recall again the $m = 2$ case of c_1 and c_2 with a positive $\hat{\theta}$-gap. Let us say that the $\hat{\theta}$-gap is γ and favors c_1. Obviously, to shift the choice of $\tilde{c}$ from c_1 to c_2 the cumulative error must be greater than the $\hat{\theta}$-gap of γ. In the spirit of a conservative analysis, we approximate the chances of such a cumulative error with the probability that any single δ exceeds $\gamma/2$. By substitution into the Hoeffding inequality the probability that a *single* cosine estimate will be inaccurate by $\gamma/2$ or more is calculated to be less than or equal to:

$$2e^{-1.5\gamma^2 n}$$

Applying the union bound, we see that the probability that *neither* of the estimates will exceed $\gamma/2$ is greater than:

(5.2) $$1 - 4e^{-1.5\gamma^2 n}$$

Thus for the $m = 2$ case, the $\Pr(\tilde{c} \approx \bar{c}) > 1 - 4e^{-1.5\gamma^2 n}$. Notice the use of the $\approx$ rather than the $=$ operator. $\tilde{c}$ is not guaranteed to be equal to the true match just ϵ-equivalent for $\epsilon \leq \hat{\theta}$-gap (i.e. γ).

If we let all the observed $\hat{\theta}$-gaps be characterized by their minimum, we can generalize 5.2 for $m > 2$, yielding:

(5.3) $$\Pr(\tilde{c} \approx \bar{c}) > 1 - 2me^{-\min(\hat{\theta}\text{-gap}/2)^2 n 1.5}$$

5.1.3 Characterization

We now characterize equation 5.3 by its big-O form:

(5.4) $$O(-me^{-\min(\hat{\theta}\text{-gap})^2 n})$$

The most interesting observation regarding equation 5.4 is the balance between the $\min(\hat{\theta}\text{-gap})^2$ term and the m due to the presence of e. They are dominant to be sure, but geometric constraints (in the form of search window size m) do not dominate accuracy as heavily as would first appear.

Another interesting observation is that given that $\min(\hat{\theta}\text{-gap})$ is not degenerate (i.e. 0), CV-NICS cannot fail but succeed if given enough input (n).

CV-NICS performance seems to be dependent on 1) available geometric constraints, 2) scene challenge, and 3) input size in that order.

5.2 Empirical Performance

We now present empirical evidence of CV-NICS performance by comparing its correspondence data to human generated results. To show the flexibility of the CV-NICS approach, this is done for two subjects at the extreme ends of the structured light capturing scale: a macro scene and a scene with an extremely large object.

5.2.1 Macro Object

We begin by capturing the mask shown in figure 5.2. The two viewpoints were horizontally separated by approximately 30 centimeters with the camera located approximately 60 centimeters from the subject. The mask is shown in figure 5.2 with the projected grid pattern used for manually collecting data. The mask makes an interesting object choice because it is a completely featureless uniform shade of white (the grid shown in figure 5.2 is being projected). This lack of features make it particularly challenging for most heuristic or AI based stereo vision techniques and ideally suited to a structured light approach.

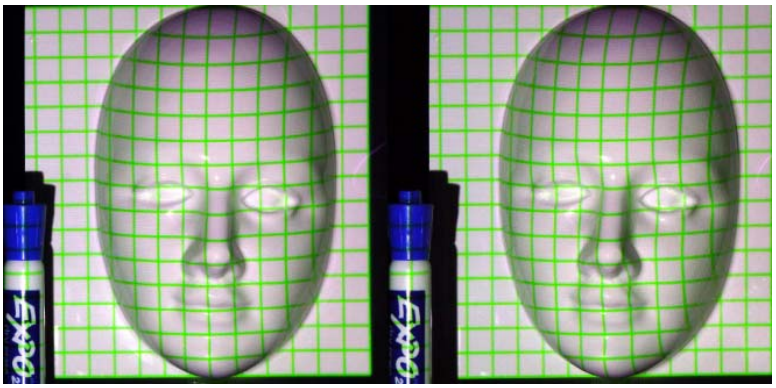

Fig. 5.2 The two views used to generate the macro dataset.

For a sense of scale, the mask is shown in figure 5.2 with a standard Expo white board marker. Both views were captured with the camera approximately 90 centimeters away with a horizontal separation between views of approximately 30 centimeters.

5.2.2 Room Scale Object

CV-NICS differentiates itself from the majority of structured light techniques from the literature by not being limited to small objects or diorama scenes. The second test case (shown in figure 5.3) is of a *room-scale* object over 8 feet tall and 7 feet wide. The scene was imaged from a distance of approximately ten meters with a vertical separation between views of approximately 40 centimeters (the camera was moved up and down on a vertically adjustable tripod).

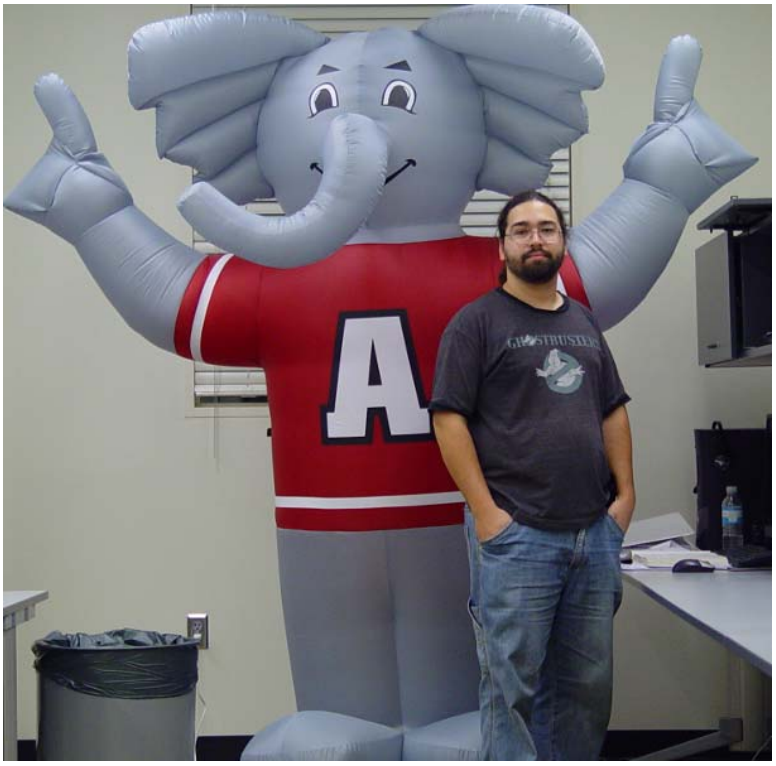

Fig. 5.3 The authors and his 8’ by 7’ pachyderm companion Al. This photo was taken from one of the two viewpoints used to generate the room-scale dataset.

5.2.3 Manual Matching

We here compare the automated results of our process to a rigorous procedure for manually finding a limited number of correspondences. In addition to random noise, we also projected onto each view a grid pattern (see figure 5.2) and took one additional image with this pattern visible. These images were then loaded into a photo manipulation program where the top-left and bottom-right points of grid intersections were manually identified. This information was used to construct roughly 12 by 12 rectangular collections of pixels we call *patches*. The patches within each view were then manually mapped to each other, providing rough correspondence data. To ensure accuracy, this process was carried out twice. Conflicts

between attempts (.e.g. different mappings, different top-left and bottom-right points) were manually resolved. If a patch covered a truly ambiguous point and a manual decision could not be reached, that patch was thrown out. Patches that covered occluded or non-reflective points were also removed. Following this procedure, 100 patches of known correspondence were generated for the macro data set, and 72 for the room-scale data set. The 100 macro patches can be seen in figure 5.3. The results of comparing the automatic computed correspondences to the manually matched data can be seen in table 5.1.

Scene		Average	Std. Dev.	Coverage
macro:				
	pre-filtering:	1.01	0.87	
	post-filtering:	0.97	0.89	85%
room-scale:				
	pre-filtering	0.87	1.03	
	post-filtering	0.73	0.75	54%

Table 5.1 Disagreement between manual and automatic matching in absolute pixel distance before and after error-correcting filtering. The statistics are for a subset of pixels within the images that were matched manually. Coverage is percentage of this subset kept by filtering *not* percentage of total image pixels kept.

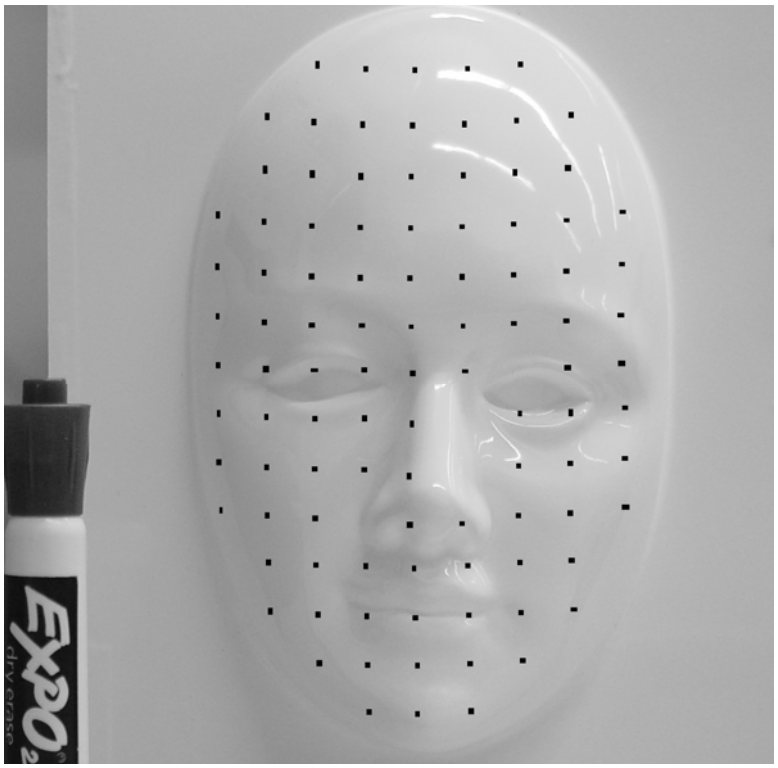

Fig. 5.4 The mask under normal lighting conditions and with manually matched patches super-imposed. These small collections of pixels were carefully placed within two separate views to provide an approximation of ground truth correspondence.

Each scene was twice compared to a manual matching: before and after filtering out pixels that did not meet the statistical requirements presented in section 4.5. In each case, disagreement was measured in pixel distance. If the automatic process proposed a correspondence point of u, v and the manual process found it to be $u - 2, v$, this was marked as a disagreement of 2.

The filtering percentages track which pixels inside the manually marked patches were removed by the filtering step. This is an important distinction to note. In the case of the room scene, 46% of the pixels within manually marked patches were removed *not* 46% of the pixels in the image.

Before filtering, the macro process disagreed by 1.01 pixels on average with a standard deviation of 0.87 pixels. Filtering improved this somewhat resulting in an average disagreement of 0.97. This was at the expense of a somewhat wider standard deviation of 0.89. Only 15% of the patch pixels were removed by the filtering process for this scene.

The room-scale scene was more improved by filtering than the macro scene. Originally disagreeing by 0.87 pixels on average, this was improved to 0.73. Its standard deviation was improved as well: from 1.03 to 0.75. This was at the expense of coverage. The filtering process removed 46% of the room scene's patch pixels.

While both datasets are relatively small, their equally small standard deviations suggest that we can conclude the manual and automatic processes are in (statistical) agreement.

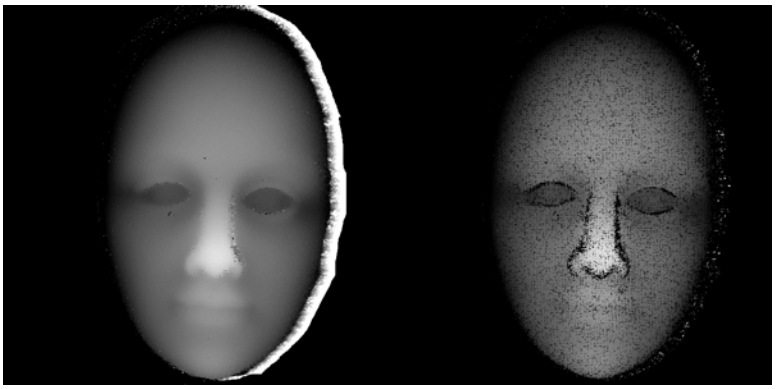

Fig. 5.5 A (left) unfiltered and (right) filtered disparity map of the macro data set. The correspondences of each pixel was used to calculate a horizontal shift. This was then normalized to a gray value between 255 (no shift) and 1 (maximum shift within the data set). Absolute black represents missing data.

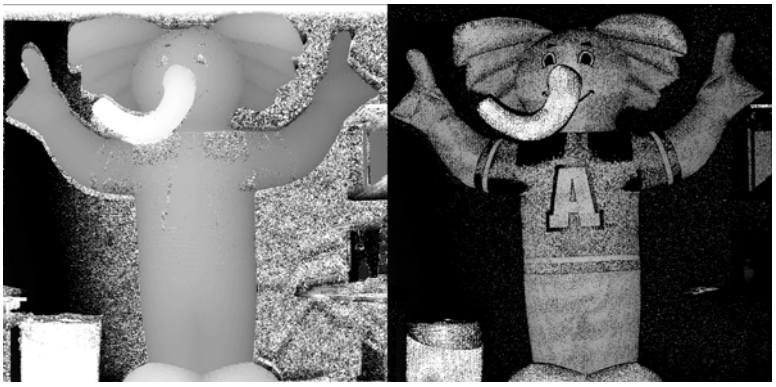

Fig. 5.6 A (left) unfiltered and (right) filtered disparity map of the room-scale data set.

Figures 5.5 and 5.6 contain the filtered and un-filtered disparity maps generated from the macro and room-scale correspondences, respectively. Each pixel is gray shaded by its normalized correspondence offset (i.e. how far away horizontally it moves between images). The possible shades are between 255 (white, no movement between images) and 1 (dark, the maximum horizontal shift in the data set). Actual black represents missing or ignored pixels.

We here investigate the convergence of the automatic correspondence process towards the manually generated data. For the case of the macro data set, Figure 5.7 displays such convergence in terms of both the average absolute error and its standard deviation.

The average disagreement with the manual data set (in terms of absolute pixels) improves exponentially. The standard deviation improves linearly. Figure 5.8 shows linear and exponential curves that have been fit to the data set. The fitness of these models can be evaluated in figure 5.9 where they have been composited with the original data.

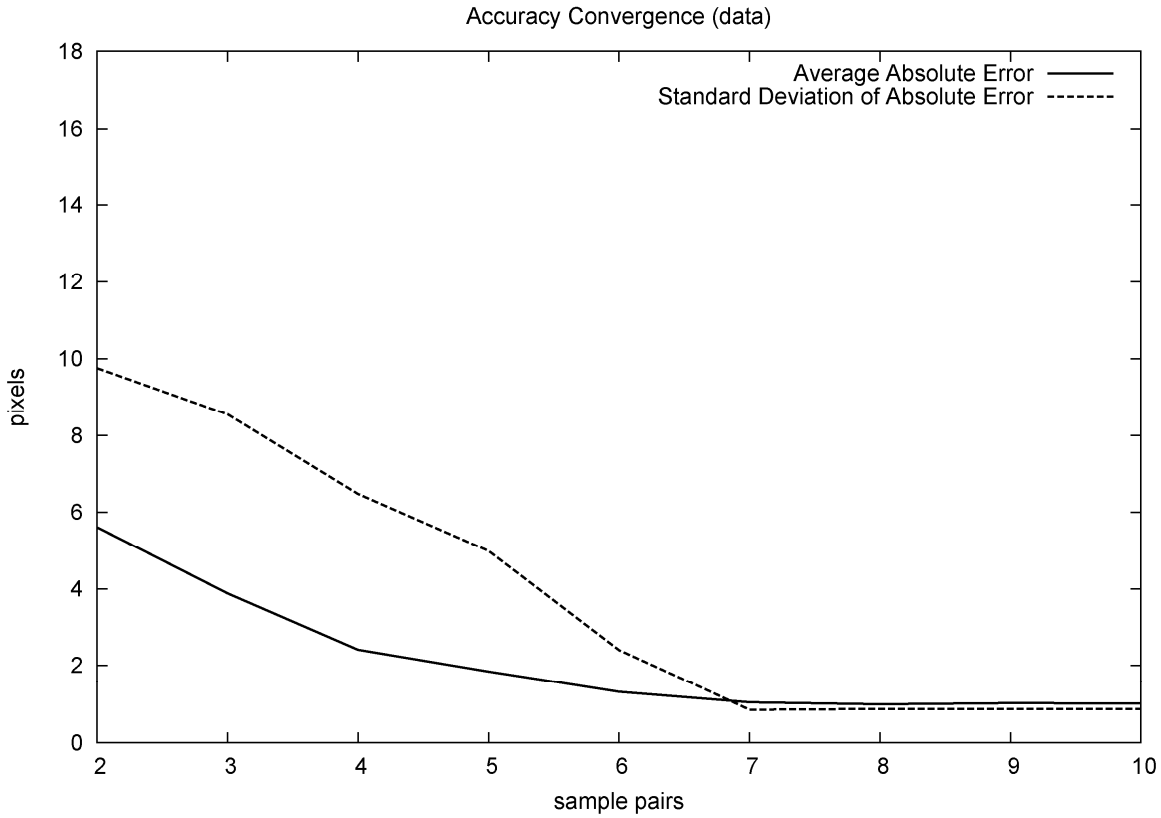

Fig. 5.7 Raw data trends for the macro dataset.

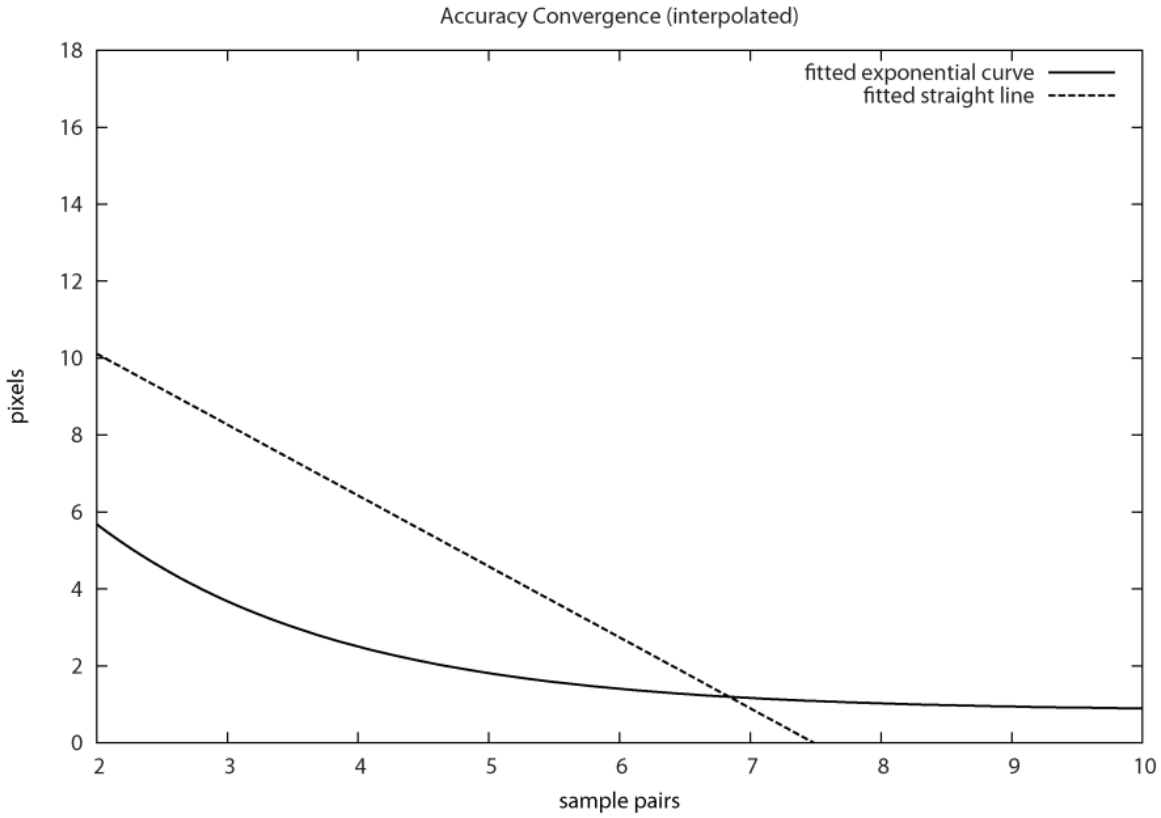

Fig. 5.8 Interpolated models for the macro dataset.

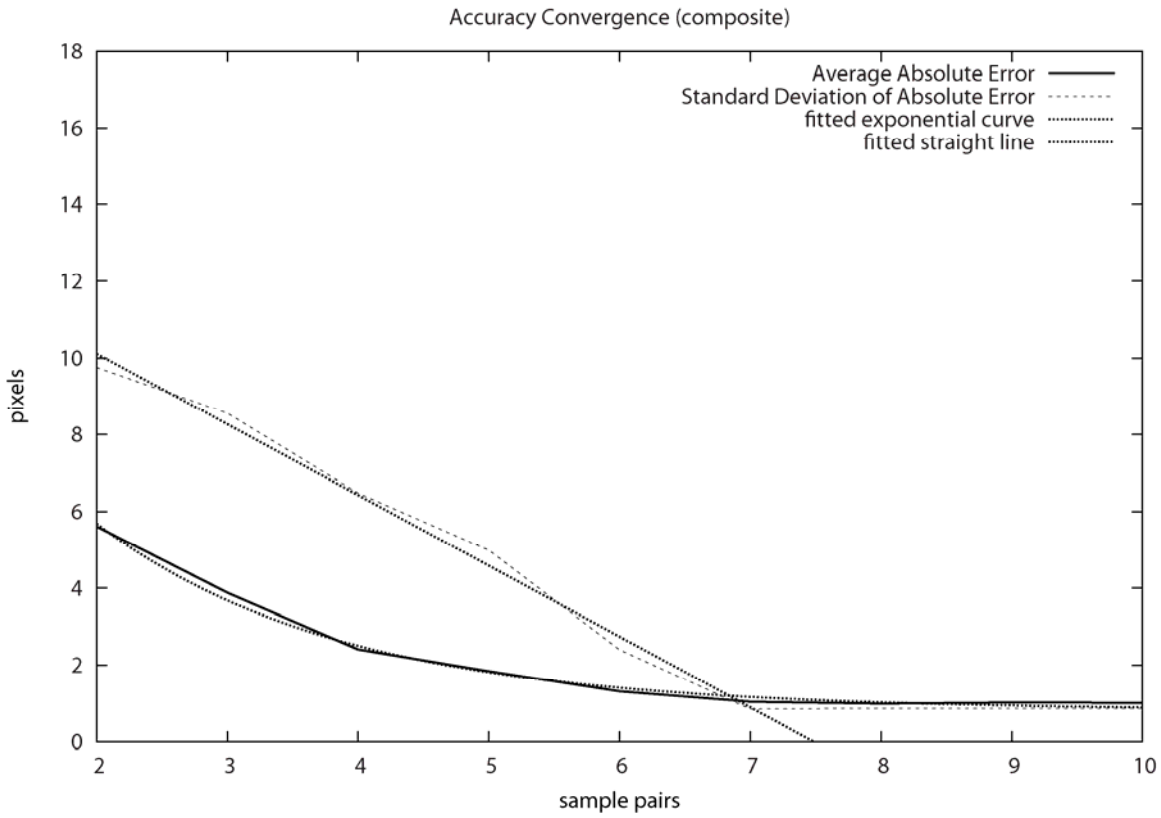

Fig. 5.9 Data and models composited together.

5.3 Empirical Characterization

This section focuses on empirical results that directly relate to the qualitative features of CV-NICS as well as the characterization made in section 5.1.3. I will begin with an example of CV-NICS's robustness to poor capturing conditions.

Examine figure 5.10, a single image from a set captured under less than ideal conditions. The three objects involved, a clock, a doll, and a telephone were captured from approximately one meter away by a set of CCD cameras meant for robotic and embedded system use. The image is low resolution, poorly focused, and improperly color balanced. The

patterns are being projected by a low luminance pocket projector. However, despite these restrictions CV-NICS is still able to generate a large number of correspondence pairs as shown in figure 5.11.

Fig. 5.10 A sub-optimal input image.

Fig. 5.11 Disparity map based on the data from figure 5.10

The depth features of the telephone are almost fully discernable as is a large portion of the doll. Interesting to note is the effect of poor focus in both the image and projector. They reduce the effective resolution of the disparity information. Note how the disparity map has large “clumps” of pixels sharing the same disparity.

While figure 5.10 shows a particularly pathological case, none of the images presented in this dissertation were captured under conditions that would be considered optimal by the literature. Color balance was handled completely in camera and except for the room scale dataset where a tripod was employed no rigging of any kind was utilized to restrain the camera's movement to simple horizontal translation. *Every* dataset shown had some degree of extraneous vertical motion, showing that CV-NICS is robust to such errors.

The analysis of section 5.1 suggest that CV-NICS derives its power from the exploitation of local illumination properties. This notion makes intuitive sense because on their own the patterns used by CV-NICS are very low bandwidth with each pixel encoding at most 2^{30} bits of information (three channels on or off over ten images). In the literature, those that attempted to use corollary information and found it lacking [41] used a virtual camera approach which cannot exploit this local information. Figure 5.12 and 5.13 supports the idea that CV-NICS performance is due mostly to its making of a feature to feature comparison. Figures 5.12 and 5.13 show disparity maps of the same action figure. Figure 5.12 was created using a virtual camera approach that compared one image set to a properly distorted reference pattern. Figure 5.13 was created by comparing image sets captured from different positions. The failure for the virtual camera version to converge (seemingly at all) supports the assertion that CV-NICS performance derives from its sensitivity to local illumination.

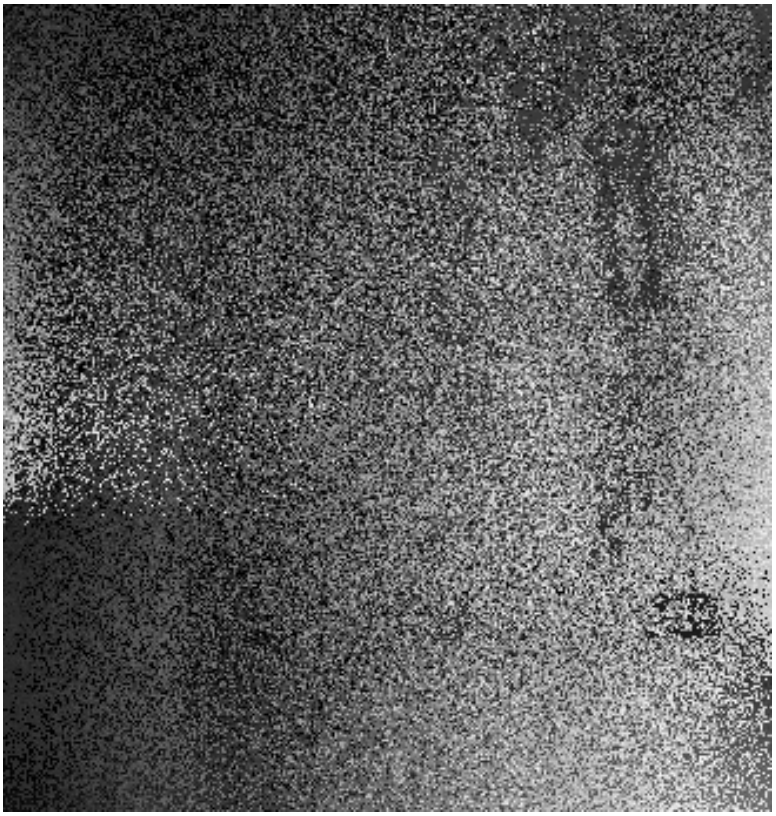

Fig. 5.12 Disparity Map of an action figure generated using a virtual camera approach

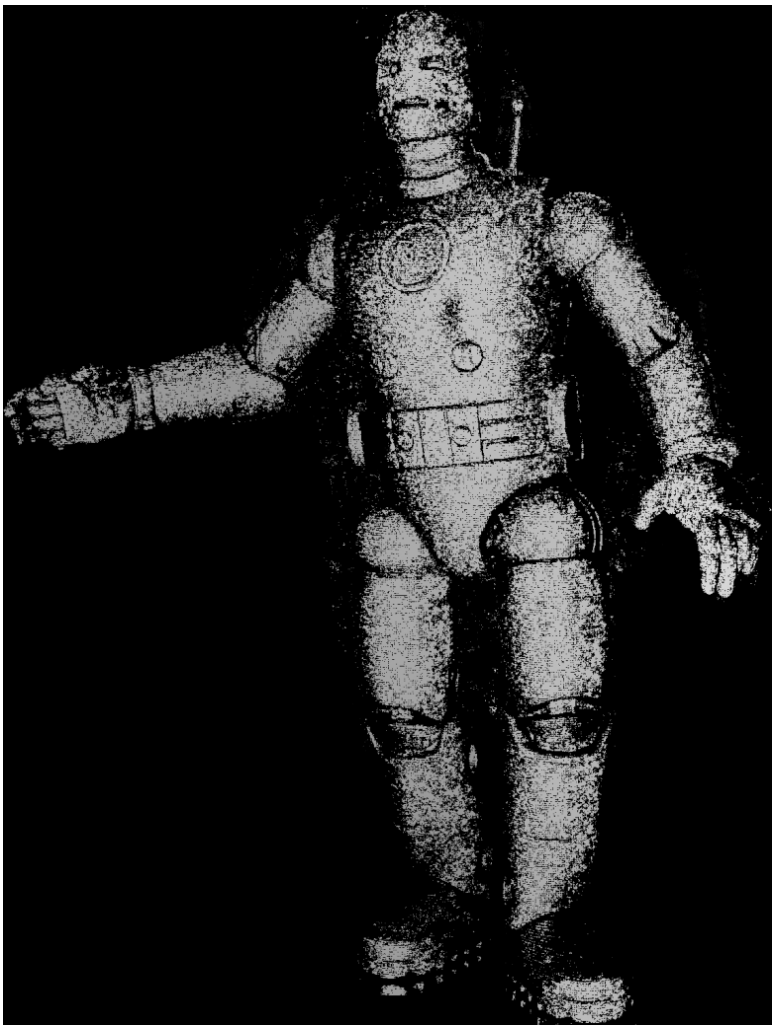

Fig. 5.13 Disparity map of the same action figure as in figure 5.12 generated using a mult-position approach

A natural area of investigation is the role of the similarity metric in CV-NICS performance. However, perhaps because of the temporal nature of the comparisons performed initial investigation into this topic has revealed little differentiation between metrics. Figure 5.14 is a good example. It shows a cosine similarity based disparity map on the left and a Pearson's rho based map on the right. In section 5.1 unit normality of the incoming vectors was assumed for ease of analysis. Figure 5.14 suggest that such an

assumption is not unjustified. By their definition Pearson's rho and cosine similarity are equivalent when the data is normalized. The equivalent performance of the two metrics suggest that the data is not so far from such a state.

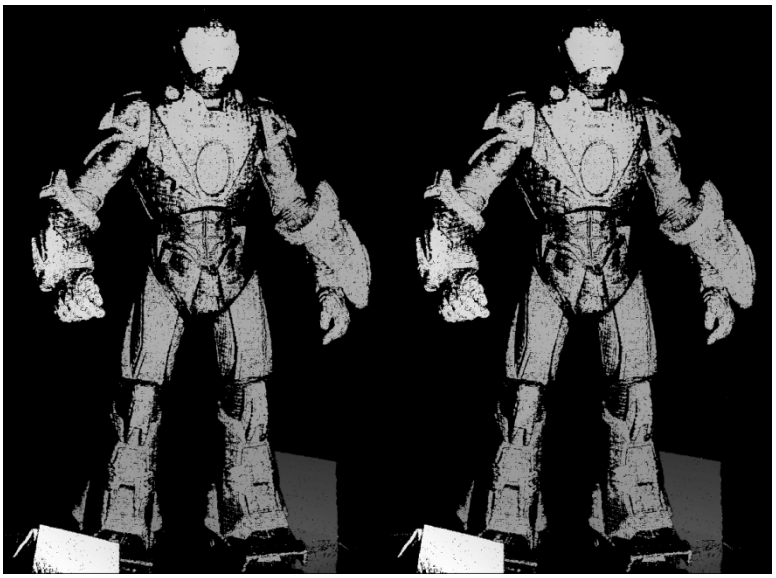

Fig. 5.14 Disparity maps of an action figure generated using two different similarity metrics. Cosine similarity was used on the left, while Pearson's rho was used on the right.

CHAPTER 6

APPROXIMATE 3D RECONSTRUCTION

Throughout, this dissertation demonstrates the density of CV-NICS correspondence data by constructing ad-hoc 3D reconstructions. This chapter briefly describe the methodology for such reconstructions.

Begin with the correspondence data of interest C, in the form: u, v, hd, vd where u and v are the coordinates in image space and hd and vd are the intra-image disparities (in horizontal and vertical pixels). Estimate a volume b that would bound the points of interest in $\mathbb{R}^3$.

We now define linear transformations between the space of C and the real world:

$$x = (\max(u) - u)\frac{b_w^2}{\max(u) - \min(u)}$$

$$y = (\max(v) - v)\frac{b_h^2}{\max(v) - \min(v)}$$

$$z = (\max(hd) - hd)\frac{b_d^2}{\max(hd) - \min(hd)}$$

where: b_w, b_h, and b_d are the bounding volume's width, height, and depth. Please note that these equations actually generate the *mirror image* of a coordinate in C in real world coordinates.

With the above equations, every point in C can be converted to a equivalent depth cloud D. Figure 6.1 shows such a cloud, generated from the room scale data set.

As with the image in figure 4.2, figure 6.1's depth cloud is viewed from an angle from which images were not actually captured and represents a rough view reconstruction.

Important to note is that neither figure 4.2 or figure 6.1 contain *any* interpolated data. For any non-black pixel in either image, there is a calculated correspondence pair. Since both reconstructions were performed post-filtering, they are good representations of how dense the correspondence data from CV-NICS tends to be.

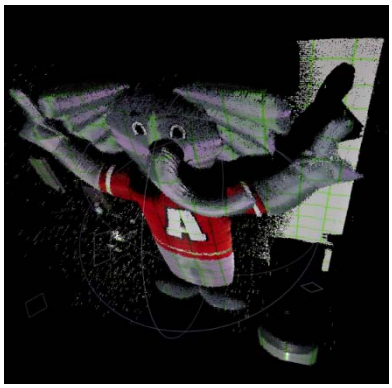

Fig. 6.1 Point cloud of the room scale dataset generated by applying Equations X, Y, and Z.

As with the image in figure 4.2, figure 6.1's depth cloud is viewed from an angle from which images were not actually captured and represents a rough view reconstruction.

Important to note is that neither figure 4.2 or figure 6.1 contain *any* interpolated data. For any non-black pixel in either image, there is a calculated correspondence pair. Since both reconstructions were performed post-filtering, they are good representations of how dense the correspondence data from CV-NICS tends to be.

CHAPTER 7

DISCUSSION

This chapter presents a high-level closing discussion, beginning with the threats to the validity of the analysis presented.

7.1 Threats to Validity

The largest threat to the validity of the presented analysis is that commercial cameras do not allow direct access to light wavelengths and instead provide intensity values for the red, green, and blue sub-bands. With access to only this information, a viewpoint neutral comparison of color makeup is not strictly possible. However, the error introduced by this approximation is very likely smaller than the uncertainty introduced by the limits on projector and camera resolution. The effects of this approximation are further mitigated by the use of extreme intensities within the random patterns (see section 4.5).

The current analysis completely ignores the possibility of prismatic effects or purely secular surfaces, where color depends on viewing angle. CV-NICS simply can't be applied to surfaces such as mirrors, holograms, or transparent glass. However, even when such surfaces are present they will not disturb the processing of the remaining scene surfaces.

The most significant threat to internal validity of the presented empirical work is a lack of ground truth data for performance evaluation. The manually provided dataset is small and its relative accuracy is unknown. However, the extremely small statistical variances

provide confidence that the manual and automatically generated datasets do not significantly differ. This alone has practical value for a number of applications where manual correspondence matching is currently used (e.g. special effects).

7.2 Discussion

Given that a point in the CV-NICS pattern set contains at most 2^{30} bits of information (see section 4.5), the cosine based discrimination of CV-NICS performs far better than expected. This is because the information encoded in the projected pattern is not the primary source of data for CV-NICS discrimination. As discussed in chapter 2 and section5.1, surface illumination properties are inherently local. CV-NICS performs a feature to feature search, and can thus utilizes this local information in its point comparisons.

There has been a trend in the structured light literature towards temporal pattern sets because they are invariant to many of the transformations of projection and image capture. On its own such invariance is desirable, but when these are combined with direct recognition the implementing technique becomes invariant to the potentially useful information of local illumination properties as well. The high performance of CV-NICS relative to its complexity indicates that such information is very rich. This suggest that if existing temporal pattern structured light techniques were modified to use a multi-viewpoint feature to feature comparison instead of direct decoding, they might be able to significantly reduce the number of images require for the same level of performance.

Another possible advantage to incorporating a feature to feature comparison instead of direct encoding is analyzability. It is difficult, for example, to image how an analysis like

that of section 5.1 would be constructed for a technique that directly decoded a gradient or used an m-array. The probabilistic nature of a feature comparison allows for natural quality assessment.

The current implementation of CV-NICS is a simple threaded search. The points of interest in image set $\boldsymbol{A}$ are split into eight groups and a single thread processes each group, searching for the point best matching in set $\boldsymbol{B}$. The only state kept per point is the current best match, so there is no need for threads to communicate. In principle, this approach could be extended to a thread-per-point SIMD approach, but we have found current runtime performance acceptable. The current implementation can process a five mega-pixel dataset in approximately 20 minutes on a modern 2.4Ghz Intel Core 2 Quad processor. The original goal for CV-NICS was to be a light-weight alternative to traditional structured light approaches, with lower implementation, capture, and data requirements. Our experiments indicate that CV-NICS can indeed fit this role. Figure 7.1 was produced using only 10 images per viewpoint and using no camera rigging at all but by instead repositioning the camera by hand. For this kind of performance, CV-NICS is the most casual correspondence data collection method the authors are aware of.

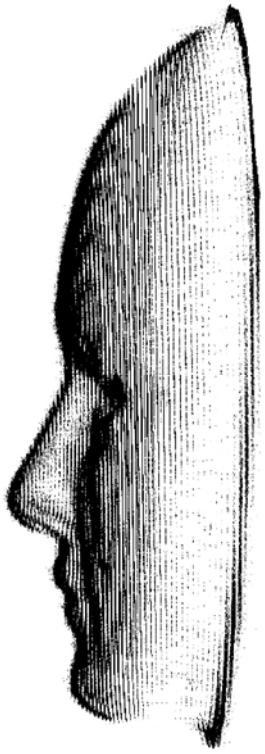

Fig. 7.1 A reconstructed profile view of the masked based solely on the images in figure 4.1.

7.3 Future Work

Besides tightening the bounds of chapter 5, the most promising area for investigation seems to be the θ-gaps (see chapter 5). Are there surface properties other than poor reflectance that yield smaller θ-gaps? What can be done to the patterns themselves to improve these gaps? The patterns projected are semi-uniform random, but this property is not guaranteed to be preserved after transformation by the $p(.)$ and $l(.)$ functions. It would be interesting to create an "adaptive" version of CV-NICS which used the observed images to attempt to mitigate these effects by making real-time changes to the projected patterns. Would such a system perform better or worse than the current system? Such a system could

also be used to determine the ideal role of the $l(.)$ function. Such investigations of specific θ-gap properties are the next logical step for CV-NICS research.

7.4 Conclusion

We have presented a novel structured light approach, CV-NICS, that introduces elements of statistical classification. We have characterized its performance both through mathematical analysis and empirical experimentation.

For applications requiring a more casual correspondence data collection process or a straight-forward implementation, CV-NICS presents many advantages. CV-NICS requires relatively few input images to produce dense correspondence data and it does not require camera or projector geometry knowledge. Since the pairing process of CV-NICS is based on similarity metrics, each pairing has a natural quality assessment statistic and correspondence pairs can be thresholded on this quality.

We have shown that CV-NICS is able to perform as well as more complex structured light techniques by exploiting local surface illumination properties in a feature to feature similarity search process rather than recognizing predetermined pattern features. This is our most important contribution. It suggest possible extensions to other temporal structured light techniques.

REFERENCES

[1] R. Hummel B. Carrihill. Experiments with the intensity ratio depth sensor. Computer Vision, Graphics and Image Processing, 32(301):337--358, 1985.

[2] Joseph Shamir Dalit Caspi, Nahum Kiryati. Range imaging with adaptive color structured light. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(5):470--480, 1998.

[3] James Davis, Diego Nehab, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27(2):296--302, February 2005.

[4] Olivier D. Faugeras, Quang-Tuan Luong, and Stephen J. Maybank. Camera self calibration: Theory and experiments. In ECCV '92: Proceedings of the Second European Conference on Computer Vision, pages 321--334, London, UK, 1992. Springer-Verlag.

[5] David Fofi, Joaquim Salvi, El Mustapha Mouaddib, Universit Picardie, and Jules Verne. Uncalibrated vision based on structured light. In IEEE Conference on Robotics and Automation, 2001.

[6] Ryo Furukawa and Hiroshi Kawasaki. Dense 3d reconstruction with an uncalibrated stereo system using coded structured light. In CVPR'05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, page 107, Washington, DC, USA, 2005. IEEE Computer Society.

[7] Andrew S. Glassner. An Introduction to Ray Tracing. Morgan Kaufmann, 1989.

[8] Paul M. Griffin, Lakshmi S. Narasimhan, and Soung R. Yee. Generation of uniquely encoded light patterns for range data acquisition. Pattern Recognition, 25(6):609—616, 1992.

[9] J. Guehring. Dense 3d surface acquisition by structured light using off-the-shelf components. In Proceedings of Videometrics and Optical Methods for 3D Shape Measuring, pages 220--231, 2001.

[10] G. Maitre H. Hugli. Generation and use of color pseudo random sequences for coding structured light in active ranging. In Proceedings of Industrial Inspection, volume 1010, pages 75--82, 1989.

[11] Olaf Hall-Holt and Szymon Rusinkiewicz. Stripe boundary codes for real-time structured-light range scanning of moving objects. In Eighth International Conference on Computer Vision (ICCV), July 2001.

[12] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301):13--30, 1963.

[13] E. Horn and N. Kiryati. Toward optimal structured light patterns. 3D Digital Imaging and Modeling, International Conference on, 0:28, 1997.

[14] Yi-Chih Hsieh. Decoding structured light patterns for three-dimensional imaging systems. Pattern Recognition, 34(2):343--349, 2001.

[15] D. C. Douglas Hung. 3d scene modelling by sinusoid encoded illumination. Image Vision Comput., 11(5):251--256, 1993.

[16] Minoru Ito and Akira Ishii. A three-level checkerboard pattern (tcp) projection method for curved surface measurement. Pattern Recognition, 28(1):27--40, 1995.

[17] R.A. Jarvis. Range sensing for computer vision. In 3DORS93, pages 17--56, 1993.

[18] M.D. Altschuler J.L. Posdamer. Surface measurement by space-encoded projected beam systems. Computer Graphics and Image Processing, 18:1--17, 1982.

[19] Bela Julesz. Foundations of Cyclopean Perception. University of Chicago Press, 1971.

[20] A. C. Kak K. L. Boyer. Color-encoded structured light for rapid active ranging. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 9:14--28, 1987.

[21] Robert R. Korfhage. Information Storage and Retrieval. Wiley, 1997.

[22] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. Technical Report TR692, 1998.

[23] Philippe Lavoie, Dan Ionescu, and Emil Petriu. A high precision 3d object reconstruction method using a color coded grid and nurbs. In ICIAP '99: Proceedings of the 10th International Conference on Image Analysis and Processing, page 370, Washington, DC, USA, 1999. IEEE Computer Society.

[24] T. Sakai M. Minou, Takeo Kanade. A Method of Time-Coded Parallel Planes of Light for Depth Measurement, E64(1):521{528, August 1981.

[25] Jessie Macwilliams and Neil Sloane. Pseudo-random sequences and arrays. In Proceedings of the IEEE, 1976.

[26] T. P. Monks and J. N. Carter. Improved stripe matching for colour encoded structured light. In CAIP '93: Proceedings of the 5th International Conference on Computer Analysis of Images and Patterns, pages 476--485, London, UK, 1993. Springer-Verlag.

[27] Raymond A. Morano, Cengizhan Ozturk, Robert Conn, Stephen Dubin, Stanley Zietz, and Jonathan Nissanov. Structured light using pseudo-random codes. IEEE Trans. Pattern Anal. Mach. Intell., 20(3):322--327, 1998.

[28] Faugeras Olivier. Three-dimensional computer vision: a geometric viewpoint. MIT Press, 1996.

[29] Bui T. Phong. Illumination for computer generated pictures. Commun. ACM, 18(6):311--317, June 1975.

[30] C. Rocchini, Paulo Cignoni, C. Montani, P. Pingi, and Roberto Scopigno. A low cost 3d scanner based on structured light. Computer Graphics Forum, 20(3):299--308, 2001.

[31] F. Matsuda S. Inokuchi, K. Sato. Range imaging system for 3-d object recogntion. In Proceedings of the International Conference on Pattern Recognition, pages 806--808, 1984.

[32] J. Salvi. A robust-coded pattern projection for dynamic 3d scene measurement. Pattern Recognition Letters, 19(11):1055--1065, September 1998.

[33] Joaquim Salvi, Jordi Pages, and Joan Batlle. Pattern codification strategies in structured light systems. Pattern Recognition, 37(4):827--849, April 2004.

[34] K. Sato. Range imaging based on moving pattern light and spatio-temporal matched lter. pages I: 33--36, 1996.

[35] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47:7--42, April 2002.

[36] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using

structured light. volume 1, pages 195--202, 2003.

[37] Jiqiang Song and Michael R. Lyu. A hough transform based line recognition method utilizing both parameter space and image space. Pattern Recognition, 38(4):539--552, April 2005.

[38] H. J. W. Spoelder, F. M. Vos, Emil M. Petriu, Senior Member, and F. C. A. Groen. Some aspects of pseudo random binary array-based surface characterization. IEEE Transactions on instrumentation and measurement, 49:1331--1336, 2000.

[39] R. J. Valkenburg and A. M. M C Ivor. Accurate 3d measurement using a structured light system. Image and Vision Computing, 16:99--110, 1998.

[40] P. Vuylsteke and A. Oosterlinck. Range image acquisition with a single binary-encoded light. pattern. PAMI, 12, 1990.

[41] Li Zhang, Brian Curless, and Steven M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In The 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission, pages 24--36, 2002.

[42] Li Zhang, Brian Curless, and Steven M. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 367--374, June 2003.

[43] Zhengyou Zhang and Zhengyou Zhang. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22:1330--1334, 2000.