Depth Imaging
with Time-of-Flight Cameras and the Kinect

Loren Schwarz, Nassir Navab

3D Computer Vision II
Winter Term 2011 – 24.01.2012
Lecture Outline

1. Introduction and Motivation

2. Principles of ToF Imaging

3. Computer Vision with ToF Cameras

4. Case Studies

5. Other Range Imaging Techniques

6. Human Body Tracking with the Kinect
Introduction and Motivation

Time-of-Flight (ToF) Imaging refers to the process of measuring the depth of a scene by quantifying the changes that an emitted light signal encounters when it bounces back from objects in a scene.

Images from [2]

Regular Camera Image  ToF Camera Depth Image
Introduction and Motivation

Classification of Depth Measurement Techniques

Depth Measurement Techniques

- Microwaves
- Light Waves
- Ultrasonic Waves
  - Triangulation (Two Camera Views)
  - Time-of-Flight (Single ToF Camera)
Introduction and Motivation

Reminder: Depth Measurement Using Multiple Camera Views
Introduction and Motivation

Depth Measurement Using Multiple Camera Views

- **Disadvantages:**
  - At least two calibrated cameras required
  - Multiple computationally expensive steps
  - Dependence on scene illumination
  - Dependence on surface texturing
Introduction and Motivation

Depth Measurement Using a ToF Camera

Advantages:
- Only one (specific) camera required
- No manual depth computation required
- Acquisition of 3D scene geometry in real-time
- Reduced dependence on scene illumination
- Almost no dependence on surface texturing
Introduction and Motivation

3D Reconstruction
ToF Amplitude Image
ToF Depth Image
Introduction and Motivation

Applications for ToF Imaging

• Computer Vision
  – People and object tracking
  – 3D Scene reconstruction

• Interaction
  – Gesture-based user interfaces
  – Gaming/character animation

• Medical
  – Respiratory gating
  – Ambulatory motion analysis

Images from [8,12,14]
Lecture Outline

1. Introduction and Motivation

2. Principles of ToF Imaging

3. Computer Vision with ToF Cameras

4. Case Studies

5. Other Range Imaging Techniques

6. Human Body Tracking with the Kinect
Principles of ToF Imaging

Pulsed Modulation

- Measure distance to a 3D object by measuring the absolute time a light pulse needs to travel from a source into the 3D scene and back, after reflection
- Speed of light is constant and known, $c = 3 \cdot 10^8 \text{m/s}$
Principles of ToF Imaging

Pulsed Modulation

**Advantages:**
- Direct measurement of time-of-flight
- High-energy light pulses limit influence of background illumination
- Illumination and observation directions are collinear

**Disadvantages:**
- High-accuracy time measurement required
- Measurement of light pulse return is inexact, due to light scattering
- Difficulty to generate short light pulses with fast rise and fall times
- Usable light sources (e.g. lasers) suffer low repetition rates for pulses
Principles of ToF Imaging

Continuous Wave Modulation

- Continuous light waves instead of short light pulses
- Modulation in terms of frequency of sinusoidal waves
- Detected wave after reflection has shifted phase
- Phase shift proportional to distance from reflecting surface
Principles of ToF Imaging

Continuous Wave Modulation

• Retrieve phase shift by demodulation of received signal
• Demodulation by cross-correlation of received signal with emitted signal
• Emitted sinusoidal signal:

\[ g(t) = \cos(\omega t) \]

\( \omega \): modulation frequency

• Received signal after reflection from 3D surface:

\[ s(t) = b + a \cos(\omega t + \phi) \]

\( b \): constant bias
\( a \): amplitude
\( \phi \): phase shift

• Cross-correlation of both signals:

\[ c(\tau) = s \ast g = \int_{-\infty}^{\infty} s(t) \cdot g(t + \tau) dt \]

\( \tau \): offset
Principles of ToF Imaging

Continuous Wave Modulation

- Cross-correlation function simplifies to

\[ c(\tau) = \frac{a}{2} \cos(\omega \tau + \phi) + b \]

- Sample \( c(\tau) \) at four sequential instants with different phase offset \( \tau \):

\[ A_i = c(i \cdot \pi/2), \quad i = 0, \ldots, 3 \]

- Directly obtain sought parameters:

\[ \phi = \arctan2(A_3 - A_1, A_0 - A_2) \]
\[ a = \frac{1}{2} \sqrt{(A_3 - A_1)^2 + (A_0 - A_2)^2} \]

\[ \Rightarrow d = \frac{c}{4\pi \omega} \phi \]

\( b \): constant bias
\( a \): amplitude
\( \phi \): phase shift
\( \tau \): internal offset
Principles of ToF Imaging

Continuous Wave Modulation

**Advantages:**
- Variety of light sources available as no short/strong pulses required
- Applicable to different modulation techniques (other than frequency)
- Simultaneous range and amplitude images

**Disadvantages:**
- In practice, integration over time required to reduce noise
- Frame rates limited by integration time
- Motion blur caused by long integration time
Principles of ToF Imaging

Continuous Wave Modulation
• Simultaneous availability of (co-registered) range and amplitude images

![Depth Image](image1)
![Amplitude Image](image2)
Principles of ToF Imaging

Example Device: PMDVision CamCube

- Near-infrared light (700-1400 nm)
- Continuous wave modulation
- Sinusoidal signal

- Resolution: 204x204 pixels
- Standard lens, standard calibration
- Frame rate: 20 fps

- Multiple camera operation by using different modulation frequencies

Image from [3]
Lecture Outline

1. Introduction and Motivation

2. Principles of ToF Imaging

3. Computer Vision with ToF Cameras

4. Case Studies

5. Other Range Imaging Techniques

6. Human Body Tracking with the Kinect
Computer Vision with ToF Cameras

Measurement Errors and Noise

**Systematic distance error**
- Perfect sinusoidal signals hard to achieve in practice
- Depth reconstructed from imperfect signals is erroneous
- Solution 1: camera-specific calibration to know distance error
- Solution 2: alternative demodulation techniques not assuming perfect sinusoidal signals

Image from [1]
Computer Vision with ToF Cameras

Measurement Errors and Noise

**Intensity-related distance error**
- Computed distance depending on amount of incident light
- Inconsistencies at surfaces with low infrared-light reflectivity
- Correction by means of corresponding amplitude image

![Depth images of planar object with patches of different reflectivity](image)

Image from [1]
Computer Vision with ToF Cameras

Measurement Errors and Noise

Depth inhomogeneity

- Current ToF cameras have low pixel resolution
- Individual pixels get different depth measurements
- Inhomogeneous
- „Flying pixels“, especially at object boundaries
- Correction: discard pixels along rays parallel to viewing direction

Red circles: „flying pixels“
Computer Vision with ToF Cameras

Measurement Errors and Noise

**Light interference effects**

- Signal received on detector can be mixed with signals that were reflected in the scene multiple times (instead of direct reflection)
- Emitted light waves can be attenuated and scattered in the scene
- Interference by other sources of near-infrared light (e.g. sunlight, infrared marker-based tracking systems, other ToF cameras)
Computer Vision with ToF Cameras

Geometric Calibration of ToF Cameras

- Standard optics used in commercial ToF cameras
- Use ToF amplitude image for calibration
- Standard calibration procedure for camera intrinsics
  - $f_x = f m_x$: focal length in terms of pixel dimensions (x)
  - $f_y = f m_y$: focal length in terms of pixel dimensions (y)
  - $c_x$: principal point (x)
  - $c_y$: principal point (y)
  - Lens distortion parameters

- Typical approach:
  - checkerboard calibration pattern
  - World-to-image point correspondences
  - Linear estimation of intrinsic/extrinsic parameters
  - Non-linear optimization
Computer Vision with ToF Cameras

Extraction of Metric 3D Geometry from ToF Data

- ToF data: depth $d$ in meters for every pixel location $x = (x, y)^T$
- Desired data: 3D coordinates $X = (X, Y, Z)^T$ for every pixel

- Write image coordinates in homogeneous notation $(x, y, 1)$
- Apply inverse of intrinsic parameters matrix $K$ to points

$$x = PX = K[R|t]X = K[I|0]X$$

Camera projection 3D to 2D

$$X = \frac{(x - c_x)Z}{f_x}$$
$$Y = \frac{(y - c_y)Z}{f_y}$$

Inverse relation for $X$ and $Y$
Computer Vision with ToF Cameras

Extraction of Metric 3D Geometry from ToF Data
- Simply taking measured depth $d$ as $Z$ coordinate is not sufficient
- Depth is measured along rays from camera center through image plane
Computer Vision with ToF Cameras

Extraction of Metric 3D Geometry from ToF Data

- Ray from camera center into 3D scene:

\[
\begin{pmatrix}
\frac{(x - c_x)Z}{f_x} \\
\frac{(y - c_y)Z}{f_y} \\
Z
\end{pmatrix} \rightarrow \begin{pmatrix}
\frac{(x - c_x)}{f_x} \\
\frac{(y - c_y)}{f_y} \\
1
\end{pmatrix} =: \tilde{X}
\]

- Normalize to unit length (keep only direction), multiply with depth: \( X = \|\tilde{X}\| \cdot d \)
Computer Vision with ToF Cameras

Combining ToF with Other Cameras
- Additional, complementary information (e.g. color)
- Higher-resolution information (e.g. for superresolution)
- Example: combination with a high-resolution RGB camera
- Approach: Stereo calibration techniques, giving $R$, $t$ and $K_{\text{tof}}$, $K_{\text{rgb}}$

![Diagram](Attach an image of the diagram here)

Reconstruction using $K_{\text{tof}}$
Projection using $K_{\text{rgb}}$
3D Points
Lecture Outline

1. Introduction and Motivation

2. Principles of ToF Imaging

3. Computer Vision with ToF Cameras

4. Case Studies

5. Other Range Imaging Techniques

6. Human Body Tracking with the Kinect
Case Studies

Semantic Scene Analysis [4]

- Extract geometric representations from 3D point cloud data for object recognition
- Application: scene understanding for mobile robot
- RANSAC for fitting geometric models (e.g. plane, cylinders) to point data
- Points belonging to a detected model (e.g. table) are subsequently removed
- Final step: classification of remaining point clouds to object types

Image from [4]
Case Studies

Mixed/Augmented Reality [5]

- Real-time 3D scene augmentation with virtual objects
- Substitution for traditional chroma-keying (blue or green background) used in TV studios
- Combined ToF-RGB camera system
- Segmentation of moving objects
- Occlusions and shadows between real and virtual objects
- Tracking of camera location by co-registration of 3D depth data
Case Studies

Acquisition of 3D Scene Geometry [6]

- Combined ToF and RGB cameras
- Real-time acquisition of 3D scene geometry
- Each new frame is aligned to already previously aligned frames such that:
  - 3D geometry is matched
  - color information is matched
- Point cloud matching algorithm similar to Iterative Closest Points (ICP)
- Color information compensates for low depth image resolution
- Depth image compensates for hardly textured image regions
Case Studies

Medical Respiratory Motion Detection [8]

- Patient motion during examinations such as PET, CT causes artifacts
- Several breathing cycles during image acquisition
- Reduce artifacts when breathing motion pattern is known
- Measure breathing motion using ToF camera above patient
- Plane fitting to 3D data in specific regions of interest
- Continuous breathing signal
Case Studies

Gesture Recognition [9]

- Recognition of upper-body gestures
- Invariance to view-point changes (limited invariance)
- Representation of human point cloud using 3D shape context descriptors
- Rotational invariance by means of spherical harmonics functions
Case Studies

Markerless Human Motion Tracking [12]

- Person segmentation by background subtraction
- Graph-based representation of 3D points
- Geodesic distance measurements (almost) invariant to pose changes
- Detection of anatomical landmarks as points with maximal geodesic distance from body center of mass
- Self-occlusion handling by means of motion information between frames
- Fitting skeleton to landmarks using inverse kinematics
Case Studies

Markerless Human Motion Tracking [12]
Case Studies

Markerless Human Motion Tracking [10,11]

• Background segmentation
• Extraction of many interest points at local geodesic extrema with respect to the body centroid
• Classification as anatomical landmarks (e.g. head, hands, feet) using classifier trained on depth image patches
Case Studies

Human Body Tracking and Activity Recognition [13]

- ToF-based feature descriptor for human poses
- Sampling of extremal points of 3D surface corresponding to person
- Features: distances of extremal points to centroid of point cloud
- Descriptor varies smoothly with motion
Manifold Learning for ToF-based Human Body Tracking and Activity Recognition

BMVC 2010

Loren Arthur Schwarz, Diana Mateus, Victor Castaneda, Nassir Navab

Chair for Computer Aided Medical Procedures and Augmented Reality
Technische Universität München, Germany
1. Introduction and Motivation

2. Principles of ToF Imaging

3. Computer Vision with ToF Cameras

4. Case Studies

5. Other Range Imaging Techniques

6. Human Body Tracking with the Kinect
Other Range Imaging Techniques

Structured Light Imaging

- Project a known light pattern into the 3D scene, viewed by camera(s)
- Distortion of light pattern allows computing the 3D structure

Picture from Wikipedia
Other Range Imaging Techniques

Structured Light Imaging
• Example: **Microsoft Kinect**
  – Depth resolution: 640x480 px
  – RGB resolution: 640x480 px

![Projected Structured Light Pattern](Picture from [15])
Other Range Imaging Techniques

Structured Light Imaging
  • Example: Microsoft Kinect

RGB Image
Depth Image
3D Reconstruction
Lecture Outline

1. Introduction and Motivation
2. Principles of ToF Imaging
3. Computer Vision with ToF Cameras
4. Case Studies
5. Other Range Imaging Techniques
6. Human Body Tracking with the Kinect
Kinect Body Tracking

• **Given:** Metric 3D point cloud data (background-subtracted)
• **Wanted:** Location of body joints in 3D
Kinect Body Tracking

Original algorithm by Jamie Shotton et al. presented this year at International Conference on Computer Vision and Pattern Recognition (CVPR) 2011:
Previous Solutions

How to determine human body pose using visual systems, e.g. cameras?

• Marker-based optical motion capture
  – Infrared cameras track locations of reflective markers
  – Very high precision and frame rates
  – Cumbersome, limited freedom of movement
  – Expensive professional systems

• Marker-less human pose estimation (no learning)
  – Multiple cameras observing a scene
  – Optimization to fit a human body model into each view
  – No markers, but still constrained environment
  – Sensitive to illumination and occlusions
  – Computationally very expensive
Previous Solutions

How to determine human body pose using visual systems, e.g. cameras?

- Learning-based human pose estimation
  - Training data for creating models of typical human movements for individual activities (e.g. walking)
  - Machine learning methods
  - Simpler observations, e.g. single camera view
  - Generalization problems
    - Actions that have not been learned?
    - Movement styles that differ from training style?
    - Clothing that looks different from training setting?
    - Persons that look different from training person?
  - Training computationally expensive, tracking easier
  - Mainly research applications

Images: Jaeggli et al. in IJCV 2009
Method Overview

Training:
1. Marker-based motion capture
2. Generation of synthetic training depth data
3. Training random forest classifier

Testing:
1. Predicting body part for every depth pixel
2. Extracting joint locations from distribution of predicted body parts
Training

1. Marker-based Motion Capture
   • Record motion capture data (joint angles over time)
   • Capture as many activities as possible:
     • driving, dancing, kicking, running, navigating menus, etc.
   • Discard poses that are very similar
   • Final database size: 100,000 frames
Training

2. Generation of synthetic training depth data
   • Use motion capture data to animate synthetic human body models
   • Systematic variation of size, weight, body proportions, etc.
   • Automatic labeling of 31 body parts according to predefined scheme
   • Rendering to obtain synthetic depth images
   • Systematic variation of camera perspective

Question: Would such an automatic generation of synthetic training data also work for an approach using regular RGB cameras?
Training

3. Training random forest classifier

Excursion: Random Trees and Forests

- Classification approach
- Simple decision rules
- Applied hierarchically

taken from wikipedia
Training

3. Training random forest classifier

Excursion: Random Trees and Forests

- Classification approach
- Simple decision rules
- Applied hierarchically
Training

3. Training random forest classifier

Excursion: Random Trees and Forests

- Classification approach
- Simple decision rules
- Applied hierarchically
Training

3. Training random forest classifier

Excursion: Random Trees and Forests

**GOAL:** Partition the feature space
- Using simple rules
- In a hierarchical fashion

**Tree:**
- Ensemble of nodes
- Each node consists of a decision function and a threshold
Training

3. Training random forest classifier
Training

3. Training random forest classifier

Excursion: Random Trees and Forests
Training

3. Training random forest classifier

Excursion: Random Trees and Forests
Training

3. Training random forest classifier

Excursion: Random Trees and Forests
## Training

### 3. Training random forest classifier

**Excursion: Random Trees and Forests**

**Training a Random Tree**
- Input: observations and labels
- Iteratively generate random split candidates
- Keep split candidates that best separate feature space

**Generating Split Candidates**
- Randomly chose a dimension of feature space
- Randomly choose a threshold along that dimension
- Or other approaches…

**Prediction using a Random Tree**
- Drop a feature vector down the tree until it reaches a leaf

Expensive

Cheap!
3. Training random forest classifier

Excursion: Random Trees and Forests

Prediction using Random Tree
• Unknown test point \( X \)
• Pushed down the tree until a leaf is reached
• Leaves store class posteriors learned in training step
• Probability of a point arriving in this leaf to belong to any class
Training

3. Training random forest classifier

Random Forest
- Ensemble of random trees
- Each random tree is a weak classifier, their ensemble is superior
- Each tree partitions feature space in a different way
- Final decision by aggregating votes over all trees
Training

3. Training random forest classifier

**Excursion: Random Trees and Forests**

**Random Forest**
- Final decision by aggregating votes over all trees
Training

3. Training random forest classifier
   • What is classified: individual depth image pixels
   • What are the classes: $k=31$ body part labels

Depth image pixel to be classified

Depth image pixel to be classified

Probability distributions of the pixel belonging to one of $k$ classes
Training

3. Training random forest classifier

• Features to be examined in each node:
• Depth difference between two pixels at static offsets from the pixel
• Each tree node has fixed offsets
• Offsets learned during training phase

Images: Shotton et al. in CVPR 2011

Pixel to be classified
 Pixels whose depth is compared
1. Predicting body part for every depth pixel

Aggregation of votes
Testing

2. Extracting joint locations from distribution of predicted body parts

- Simplest solution: joint location is average of all 3D points assigned a corresponding body part
- Better solution: Mode finding in body part distribution using Mean Shift algorithm
- Smaller influence of outliers
Summary

• Pose estimation formulated as a classification problem
• Every depth pixel is independently classified to a body part
• No temporal smoothing over subsequent frames
• Synthetic training dataset for coping with variability of human bodies
• Random forest classifier with 3 trees, each of depth 20
• Training time: 1 day on 1000 core cluster
• Runtime: 200 fps on Xbox
Evaluation

Images: Shotton et al. in CVPR 2011
References


References


