LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR submission

Paper ID ****

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

contributions

- introduce idea of RGB + single SPAD depth imaging
- ... gesture recognition, face authentication, ... (e.g. "learning to be a depth camera" microsoft paper)
- build prototype, demonstrate results

2. Related Work

Depth Imaging

- stereo and multiview
- structured illumination and random patterns (kinect, etc.), active stereo
- time of flight (continuous wave and pulsed)
- what we do: like pulsed but much simpler setup; no scanning, no spad array, ...

Monocular Depth Estimation

- summary of architectures and cost functions: u-net type architecture with reverse huber loss
- what we do: same thing, but augment with global hints (inspired by these approaches, we do ...)

Deep Sensor Fusion global hints for super-resolution, colorization, depth estimation

- colorization
- david's 2018 paper for depth estimation and denoising (see david's 2019 sig paper for related work)
- what we do: slightly different application

3. Method

In this section, we describe the histogram formation model for a single-pixel time-of-flight lidar sensor under diffuse, pulsed laser illumination. We also use this model to simulate the formation of a histogram given a depth map.

3.1. Histogram Formation Model

Consider a laser which emits a pulse at time $t = 0$ with shape $g(t)$. The waveform passes through a diffuser which spreads the light evenly over some 3D scene $z(x, y)$, which we describe as depth as a function of (x, y) position. To assess the distribution of returning photons from the whole scene, we first consider photons returning from a single location (x, y). The expected number of photon events detected from this location in the time interval $(t, t + \Delta t)$ is given as

$$\lambda_{x,y}[n] = \int_{n\Delta t}^{(n+1)\Delta t} (f * g)(t - 2z(x, y)/c)dt. \quad (1)$$

where c is the speed of light, and f models the temporal uncertainty in the detector. Since the detection of each photon can be described as a Bernoulli random variable, the total number of accumulated photons in this time interval follows a Poisson distribution according to

$$h[n] \sim \mathcal{P}\left(\sum_{x,y} \alpha_{x,y} \eta \lambda_{x,y}[n] + d\right) \quad (2)$$

where $\alpha_{x,y} = r_{x,y}/z(x, y)^2$ captures the attenuation of the photon counts due to the reflectance $r(x, y)$ of the scene and due to the inverse square falloff $1/z(x, y)^2$. In addition,

η is the detection probability of a photon triggering a SPAD event, and $d = \eta a + b$ is the number of ambient photons a and the number of "dark count" events b, which is a property of the SPAD.

Pick an example scene, show side-by-side with depth map and histogram of gt depth and histogram from SPAD (to make intuitive)

3.2. Monocular depth estimation with global depth hints

Given a single RGB image $I(x, y)$ and a histogram of photon arrivals $h[n]$ collected as in equation 2, we seek to reconstruct the ground truth depth map $z(x, y)$. Our method has two parts. First, we initialize our estimate of the depth map from the single RGB image via a monocular depth estimator. Second, we refine this depth map using the captured histogram $h[n]$ via a process we call Differentiable Histogram Matching. Differentiable histogram matching as a tool for post-processing the image to match the depth map to the statistics we capture from the SPAD.

Initialization via CNN Convoluational Neural Networks have become increasingly capable of leveraging monocular depth cues from to produce accurate estimates of depth from only a single image. We therefore choose to initialize our depth map estimate $z^{(0)}(x, y)$ using a CNN. However, any depth estimator reliant on only a single view will be unable resolve the inherent scale ambiguity in the scene resulting from the tradeoff between size of and distance to an object. Our next step, differentiable histogram matching, will resolve this ambiguity using the depth information present in the SPAD histogram.

Differentiable Histogram Matching Given our initial depth map estimate, our goal is now to refine that depth map so that it agrees with the information provided by the SPAD histogram. Figure showing just the differentiable histogram matching part of the model. At a high level, our goal is to formulate the SPAD histogram formation model as a differentiable function of the depth map and RGB image so that the output $\hat{h}[n]$ of the SPAD simulation can be compared to the observed SPAD histogram $h[n]$ and the depth map can be optimized via gradient descent or other first-order method. Given per-pixel depth, we can simulate SPAD histogram formation using equation 2, where we approximate $\lambda_{x,y}[n]$ using Kernel Density Estimation as:

$$\hat{\lambda}^{(i)}_{x,y}[n] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(n - \hat{z}^{(i)}(x, y))\right) \tag{3}$$

where $\hat{z}^{(i)}(x, y)$ is our estimate of the depth at pixel (x, y) at iteration i.

	$\delta^1 \uparrow$	$\delta^2 \uparrow$	$\delta^3 \uparrow$	RMSE ↓	rel	log_{10}
DORN	1	2	3	4	5	6

Once we have $\lambda^{(i)}$, we compute $\hat{h}[n]$ according to the rest of the SPAD histogram formation model, neglecting the Poisson sampling, as follows

$$\hat{h}[n] = \sum_{x,y} \alpha_{x,y} \eta \lambda_{x,y}[n] + d \tag{4}$$

model. (Explain where d and η come from.) Finally, given two histograms, we can compute the Sinkhorn Distance between them. As described in **??**, the Sinkhorn Distance is a measure of distance between two probability distributions. It is particularly desirable to us because unlike other distances between probability distributions, such as the KL divergence or the simple RMSE, the Sinkhorn Distance scales with separation as well as magnitude. Furthermore, unlike the Wasserstein Distance from which it is derived, the computation of the Sinkhorn Distance can be computed via Sinkhorn Iterations, which are fast and differentiable. We refer the reader to **??** for the details. "Minimize blah s.t. blah blah"

3.3. "How you actually solve it"

3.4. Implementation Details

For the Monocular Depth Estimator, we use a pre-trained version of the the Deep Ordinal Regression Network (DORN) []. We use the implementation of the Sinkhorn Iteration (for calculating the entropy-regularized Wasserstein Loss) from **??**uturi et. al. We update the depth map using gradient descent. Everything is implemented in PyTorch.

Table with hyperparameters (defer to supplementary info?) - Kernel Density Estimation (sigma) - Sinkhorn Iterations (maximum number, epsilon tolerance) - Gradient Descent (learning rate, epsilon tolerance, max iterations)

4. Hands and Faces

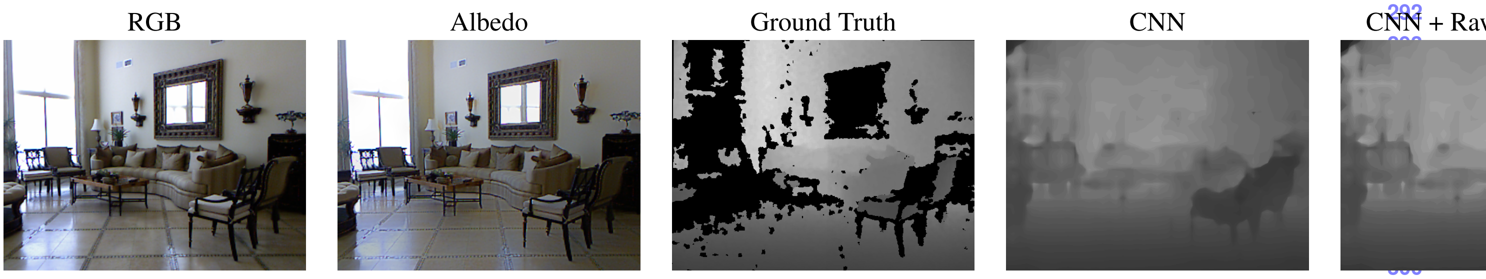
5. Assessment

5.1. Implementation

hardware, calibration, etc.

5.2. Results

6. Discussion


RGB
Albedo
Ground Truth
CNN
CNN + Rav