Learning to be a Depth Camera for Close-Range Human Capture and Interaction

Cem Keskin
Shahram Izadi
Pushmeet Kohli
David Kim
David Sweeney
Antonio Criminisi
Jamie Shotton
Sing Bing Kang
Tim Paek

Microsoft Research

Sean Ryan Fanello

iCub Facility - Istituto Italiano di Tecnologia

This is an *easy-read* summary of a **SIGGRAPH**2014

Technical Paper

As part of our annual NoirMaker social game, we aim to make some of the Conference more accessible to our international audience. For 2014, we have summarized and translated a few of the Conference's Technical Papers, to serve as catalyst for 'technical' conversations that will be part of our regional presentations, including:

'CG in Latin America'
Tuesday, August 12, 3 pm
in the ACM SIGGRAPH Theater | International Center
[room 210, West Building; VCC]

We hope that this will inspire you to access so many more interesting Technical Papers that will be presented during the Conference week in Vancouver:
http://bit.ly/s2014-technical-papers

And remember that Technical Papers can be found after the Conference, as part of the ACM Digital Library:
http://dl.acm.org

Brought to you by:

ACM **SIGGRAPH** IRC
International Resources Committee

Latin America

In summary, our paper makes the following contributions:

We demonstrate a new technique for turning a cheap color or monochrome camera into a depth sensor, for close-range human capture and interaction. Our hope is to allow practitioners to more rapidly prototype depth-based applications in a variety of new contexts.

We present two practical hardware designs for depth sensing:
a modified web camera for desktop depth sensing, and
a modified cellphone camera for mobile applications.
We demonstrate efficient and accurate hand and face tracking in both scenarios.

We propose specializations of existing multi-layered decision forests algorithms for the task of depth prediction which can achieve 100Hz performance on commodity hardware.

We present experimental and real-world results that illustrate depth estimation accuracies for our specific scenarios that are comparable to state of the art consumer depth cameras.

Hardware Setup

> … our hardware setup consists of a regular commodity camera with minor modifications. First, we remove the IR cut filter typically present, permitting sensitivity to the spectrum range of ~400-1100nm. Next, an IR bandpass filter operating at 850nm (±10nm) is used to limit all other wavelengths. This makes the camera sensitive only to this specific NIR range. Finally, we add diffuse LED illumination emitting at this spectral range. To ensure uniform lighting and limit shadowing, we build a ring of six NIR LEDs around the camera, with a minimal baseline. …
>
> The camera images are downsampled by a factor of three to 640x480 for our implementation (both devices support full HD capture). Downsampling can aid performance, and mitigate issues of defocus blur... We prefilter with a Gaussian filter, prior to subsampling, which substantially removes the effect of the different gains of the Bayer pattern of RGB filters in the IR spectrum...

… our depth prediction algorithm that learns to map a given a pixel **x** in the NIR image I to an absolute depth value. We model this continuous mapping **y(x|I)** with a multi-layered decision forest... The problem can be significantly simplified by restricting the depths of the objects to a certain range (primarily because we cannot use depth invariant features)... For such a constrained set, an expert forest can be trained to regress continuous and absolute depth values more efficiently. Thus, our first layer learns to infer a coarsely quantized depth range for each pixel, and optionally pools these predictions across all pixels to obtain a more reliable distribution over these depth ranges. The second layer then applies one or more expert regressors trained specifically on the inferred depth ranges. We aggregate these results to obtain a final estimation for the absolute depth **y** of the pixel **x**. …

Multi-Layered Forest Architecture

Coarse, Discrete Depth Classification: Given an input pixel x and the infrared image **I**, the classification forest at the first layer infers a probability distribution **p(c|x, I)** over coarsely quantized depth ranges indicated by **c**, where **c ∈ {1, . . . , C}**. The forest learns to map the pixel and its spatial context into one of the depth bins for each pixel. The experts at the next layer can be chosen based on this local estimate of **c** (denoted 'local expert network', or LEN), or alternatively the individual local posteriors can be aggregated (and averaged) over all the pixels to form the more robust estimate **p(c|I)** (denoted 'global expert network', or GEN)

Fine, Continuous Depth Regression: … each pixel or image is assigned a set of posterior probabilities over the depth bins in the first layer. In the second layer, we evaluate all the expert regression forests to form a set of absolute depth estimates. The final output depth **y** is a weighted sum over the estimates **yc** of the experts, where the weights wc are the posterior probabilities estimated by the first layer …

Multi-layered forests carry two main benefits over conventional single-layered forests. First, the multi-layered model can infer potentially useful intermediate variables that simplify the primary task, which in turn increases the accuracy of the model. Second, multi- layered forests have a reduced memory footprint than training deeper single-layered forests (as trees grow exponentially with depth). ...

Furthermore, the ability of the global weighting to aggregate depths across all image pixels has the potential to make the final prediction more robust than possible with a single layer.

Limitations

Whilst we have demonstrated the utility of our approach, there are clearly limitations. Firstly we train for uniform surface albedo (in this case skin) which limits our depth estimation to human faces or hands, or any other specific object. Our system fails to predict depth for surfaces with varying reflectance properties. …

Conclusion

In this paper, we proposed and demonstrated a low-cost technique to turn any 2D camera into a real-time depth sensor with only simple and cheap modifications. Diffuse NIR LEDs illuminate objects near the camera, and capture the reflected light with the help of an added band pass filter. The actual depth calculation is done by a machine learning algorithm, and can learn to map a pixel and its context to an absolute, metric depth value. As this is a data driven, discriminative machine learning method, it learns to capture any variation that exists in the dataset, such as changes in shape, geometry, skin color, ambient illumination, complex inter-object reflections and even vignetting effects, without the need to explicitly formulate them. To capture this much information via simple rules encoded in the decision forests, we employed a multi-layered forest that simplifies this problem in the first layer by predicting coarse quantized depth ranges for the object. … Whilst this method cannot replace commodity depth sensors for general use, our hope is that it will enable 3D face and hand sensing and interactive systems in novel contexts.

Get the complete paper:
http://bit.ly/s2014-paper-depth
Catch the SIGGRAPH 2014 presentation:
Tuesday, 12 August; 345-515 pm
[ballroom A, East Building; VCC]

ACM Transactions on Graphics, Vol. 33, No. 4, Article 86, Publication Date: July 2014