

研究室紹介：機械学習を応用した新しい映像・音響センシング

木下 裕磨*

Introductory Article: Novel Video and Audio Sensing Applying Machine Learning

by

Yuma Kinoshita*

(Received on Dec. 13, 2023 & accepted on Jan. 9, 2024)

Abstract

In the Kinoshita Laboratory, the primary research focus is on novel video and audio sensing technologies that leverage machine learning. When it comes to sensing images and sounds, there are often situations where it is challenging to obtain the desired information due to the physical limitations of sensors and environmental constraints. To overcome these challenges and improve the quality of video and audio sensing, the laboratory is engaged in research on sensing technology, media information processing, and machine learning techniques. This paper briefly introduces two research themes: "High Dynamic Range Imaging" in the context of video sensing, and "Acoustic Scene Analysis using Wireless Acoustic Sensor Networks" in the context of audio sensing.

Keywords: Sensing, signal processing, machine learning, deep learning, and multimedia

1. はじめに

木下研究室では、機械学習を応用した新しい映像・音響センシングを主な研究トピックとしている。映像や音のセンシングにおいては、センサの物理的な限界や周囲の環境の制約によって、所望の情報を得ることが難しい状況がしばしば存在する。これら困難を克服してより良い映像や音をセンシングするため、センシング技術とメディア情報処理技術、および機械学習技術の研究に取り組んでいる。本稿では、映像センシングに関する研究テーマとして「高ダイナミックレンジイメージング」を、音響センシングに関する研究テーマとして「無線音響センサネットワークによる音響シーン分析」を取り上げ、研究内容を簡単に紹介する。

2. 研究テーマ例

2.1 高ダイナミックレンジイメージング

撮像センサが扱うことができる輝度のダイナミックレンジは、現実シーンにおけるダイナミックレンジよりもはるかに狭い。そのため、現在のデジタルカメラで撮影された画像では、人間が知覚している輝度のダイナミックレンジを忠実に表現するには至っていない。本研究室では、一般の画像、すなわち、低ダイナミックレンジ (Low dynamic range, LDR) 画像の持つこの課題の解決を目的とし、より広い輝度のダイナミックレンジを記録した画像である高ダイナミックレンジ (High dynamic range, HDR) 画像をセンシングするための手法を研究している。

HDR 画像は、現実シーンにおける広い輝度のレンジの情報を



図 1 多重露出画像の例

持つ画像であり、写真撮影、コンピュータグラフィックス、監視カメラ、自動運転、医用画像など多くの分野への利用が期待されている。撮像センサにおけるダイナミックレンジの制限により、従来のカメラを用いた HDR 画像の直接的な撮影は困難な状況にある。図 1 に示すような同一シーンを異なる露出条件で撮影した複数枚の LDR 画像 (多重露出画像) の合成により HDR 画像を合成する方法も研究されているが、画像撮影時における時間的制約から十分な枚数の多重露出画像を撮影することが難しいという課題がある。

そこで、我々は、深層学習技術の利用により、単一の LDR 画像から高い精度で HDR 画像を推定する iTM-Net (図 2 参照) を提案した [1]。iTM-Net を用いることで、適切な多重露出画像の撮影が難しい状況や、そもそも多重露出画像が取得できない状況においても、より高い品質の HDR 画像を生成できる。

* 東海大学情報理工学部情報科学科

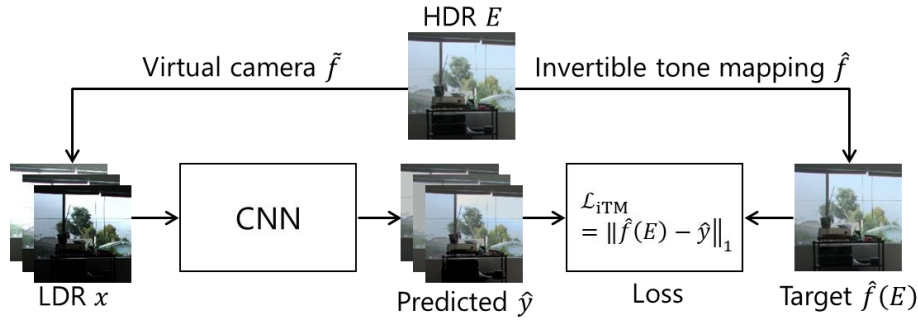


図2 iTM-Netの概要. 単一の LDR 画像から HDR 画像を生成するため, 畳み込みニューラルネットワーク (Convolutional neural network, CNN) を用いる. CNN の学習には, 教師 HDR 画像からランダム仮想カメラによる撮影をシミュレーションした LDR 画像と, 教師 HDR 画像に可逆なトーンマッピングを施した画像のペアデータを用いる.

2.2 無線音響センサネットワークによる音響シーン分析

音響シーン分析とは, 音の情報のみを用いて人の行動や周囲の状況を認識する技術である. 例えば, 「包丁がまな板に当たる音」から「現在料理中である」と推測することが可能である. 音響シーン分析には, 広範囲の音情報が重要となる. しかし, 音源がマイクロフォンから離れた場合, 目的の音が他の雑音に紛れてしまい, 分析が困難となる.

この問題に対処するため, 本研究室では, 無線マイクを広範囲に分散配置し, これらのマイクで収集される音信号を統合する無線音響センサネットワークに着目し, センサネットワークから得られる音信号を分析する枠組みを開発している. 音信号の統合と分析に音深層学習を活用することで, より精度の高い音響シーン分析が可能になると考えている.

加えて, 本研究室では, 広範囲の音情報を簡単に取得するため, 音を光に変換する小型センサデバイス「プリンキー」を開発している. プリンキーの外観を図3に示す. プリンキーは, 内蔵マイクで録音した音を, LED を介して光信号として発信する. 分散配置したプリンキーをビデオカメラで観測することで, 様々な場所の音情報を取得できる. しかし, 空中の光信号伝搬では, 光の減衰やノイズによって信号に歪みが発生する. また, 通常 30[fps] というカメラの限られたフレームレートのため, 撮影される信号は強く帯域制限されてしまう. このことが, プリンキーの音響シーン分析への応用を困難にしている.

このような背景から, 本研究室では, 音響シーン分析のためのプリンキーにおける最適な音光変換プロセスを一貫して学習

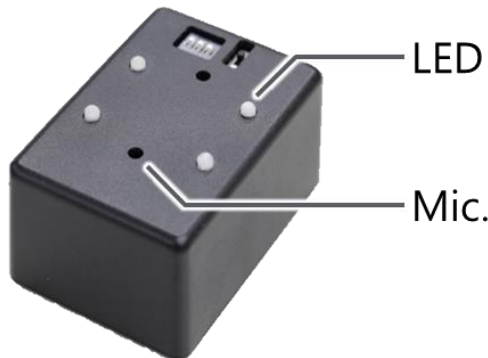


図3 音光変換デバイス「プリンキー」

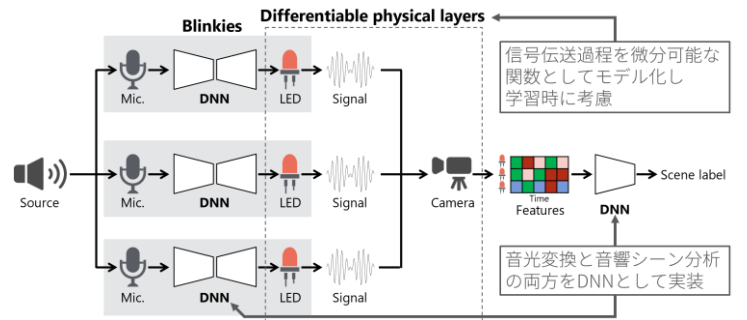


図4 プリンキーを用いた音響シーン分析の一貫学習

する, 音響シーン分析フレームワークの研究も行っている (図4参照) [2, 3]. このフレームワークでは, 物理的な制約や音響シーン分析の精度を考慮しつつ, データ駆動型のアプローチで適切な音光変換を得ることが可能となる.

3. おわりに

本稿では, 木下研究室で実施している研究テーマの一例を紹介した. 今後も, センサの物理的限界を信号処理と機械学習の活用によって克服するための研究を実施する.

参考文献

- [1] Yuma Kinoshita and Hitoshi Kiya, “iTM-Net: Deep Inverse Tone Mapping Using Novel Loss Function Considering Tone Mapping Operator,” IEEE Access, vol.7, no.1, pp.73555-73563, May 2019.
- [2] Yuma Kinoshita and Nobutaka Ono, “End-to-End Training for Acoustic Scene Analysis with Distributed Sound-to-Light Conversion Devices,” EURASIP European Signal Processing Conference, pp. 1010–1014, Dublin, Ireland, Aug. 2021.
- [3] Yuma Kinoshita and Nobutaka Ono, “Analysis on Roles of DNNs in End-to-End Acoustic Scene Analysis Framework with Distributed Sound-to-Light Conversion Devices,” APSIPA Annual Summit and Conference, pp. 1167–1172, Tokyo, Japan, Dec. 2021.