6K-04

実世界対象の構成要素の素材推定とその打撃音生成

久門 香穂† 長尾 確†
名古屋大学 大学院情報学研究科†

1 はじめに

近年，センサやユーザデバイスの発展によって現実空間の様々な要素が情報空間に取り込まれるようになった．多くの VR(Virtual Reality)・AR(Augmented Reality)コンテンツは，実世界の情報を視覚的に仮想環境へ取り込むことで，現実世界と同様の活動を可能にしている．そして実世界から取り込む情報は，視覚だけでなく，音や触覚などの次の段階へと発展している．その中でも（音声を除く）音情報は，仮想環境において，オブジェクトの物理特性，環境，衝撃の力などのユーザにとって環境認識のために重要な情報を提供する．

しかし，VR や AR での実世界を利用したコンテンツでは変化する環境に合わせて音を変える必要があり，オブジェクトやアクションに対応した音を設定するには多大な労力がかかるという問題がある．

そこで，本研究では仮想環境内での動的な音の生成を目的とし，実世界環境から認識したオブジェクトのマテリアル推定を行い，推定したマテリアルラベルを基に Generative Adversarial Network(GAN)を用いたスペクトログラム画像生成による打撃音の生成手法を提案する．

2 打撃音生成のシステム概要

本システムの構成を図 1 に示す．本研究では，ユーザのカメラで認識したオブジェクトに対して，色情報を用いてマテリアルごとのセグメンテーションを行う．マテリアルと音を関連付けるための中間情報として，取得したマテリアル画像から DEP-Net [1]を用いたマテリアル分類を行う．そのマテリアル情報に基づいて学習した GAN の Generator によってスペクトログラム画像を生成する．更に，それを Griffin-Lim アルゴリズム [2]によって音声データへ変換を行うことで，ユーザへのフィードバックを行う．

Material Estimation of Components of Real-World Objects and its Hit Sound Generation
†KUMON, Kaho (kumon@nagao.nuie.nagoya-u.ac.jp)
†NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)
†Graduate School of Informatics, Nagoya

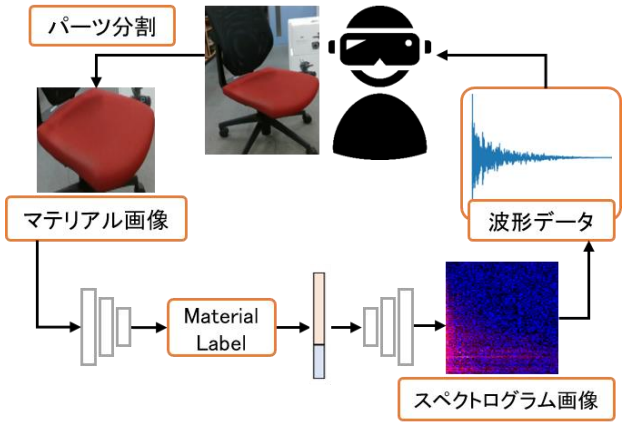


図 1 打撃音生成システム図

3 オブジェクト画像のパーツ分割による画像推定

画像からマテリアル分類を行うためにまず，画像された複数のマテリアルからなるオブジェクトをマテリアルごとにパーツ分割を行う必要がある．そのための手法として，オブジェクト画像を距離・色ごとに近い画素を super pixel [3]としてまとめることで色情報を削減し，取得した HSV 値によって K-Means 法を用いてクラスタリングを行う．

さらに，上記の手法によってマテリアルごとに分割された画像に対して，打撃音生成のために必要なマテリアルラベルの推定を行う必要がある．マテリアル推定には，Xue らの DEP-Net [1]を使用した．学習には The Greatest Hits dataset [4] に含まれるクラスのから布(cloth)，ガラス(glass)，金属(metal)，紙(paper)，プラスティック(plastic)，木(wood)の 6 クラスを教師ラベルとして選択して学習を行った後，実環境で取得可能なデータを用いて転移学習(Fine tuning)を行った．学習を行った結果，全クラスにおいて 9 割以上の精度が得られた．

4 GAN を用いたスペクトログラム画像の生成

4.1 Generator と Discriminator の構成

本研究では動的な音の生成のために，音声をスペクトログラム画像に変換することで GAN での生成を可能にしている．GAN の入力には 3 節で学習

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

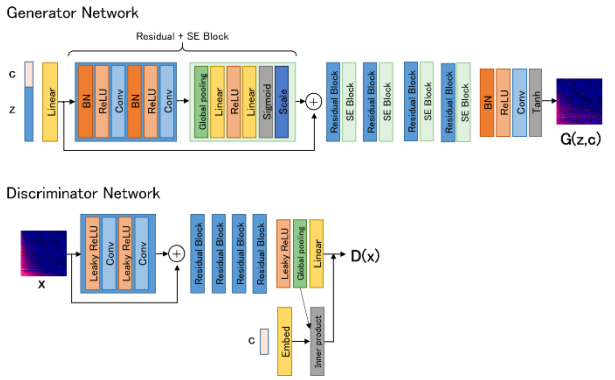


図 2 GAN のアーキテクチャ

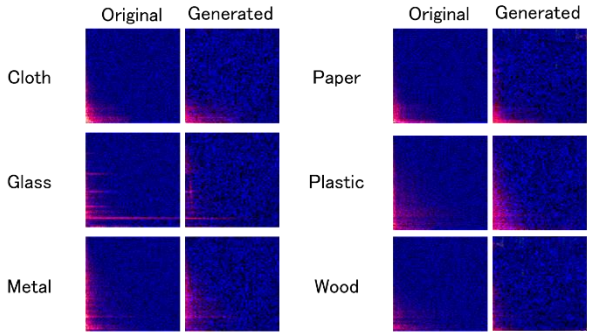


図 3 各クラスの教師データと Generator で生成したスペクトログラム画像

したマテリアル分類モデルを Encoder に使用，その出力として得られるマテリアルラベルの one-hot ベクトル c と乱数ベクトル z を結合したベクトルを用いた．学習に使用した Generator と Disciminator のアーキテクチャを図 2 に示す．

Generator は x=G(z,c)の生成を目的とし，[5]で用いられているような Residual Block を主体としたアーキテクチャをベースとした．更に，各 Residual Block 構造の後に [6]によって提案された SE Block を挟むことによって，より精度の高い画像の生成を行うことが可能になった．Disciminator は入力画像がデータセット内の画像もしくは Generator で生成した画像かを判断した確率の D(x)の出力を目的とし，Generator と同様の Residual Blcok を主体とした構造に，Projection を導入した構造となっている．両者とも損失関数には AC-GAN および DRGAN を参考にして構築した．

データセットには，3 節でも使用した The Greatest Hits dataset[4]を用いてクラスごとに打撃音を切り出した．教師データは，クラスごとに打撃開始時間から 0.4 秒間の音声データが提供され，サンプリング周波数は 44.1kHz である．この音声データを短時間フーリエ変換(サンプル点 512,ハミングサイズ 512,ホップサイズ 128)によりスペクトログラム画像を生成した．Generator の学習にはこの画像を 128×128 の画像として生成を行った．学習の optimizer は Adam を用い，ミニバッチサイズは 64 である．

4.2 Generator の学習結果

Generator の学習として，6 クラスの教師データのスペクトログラム画像と Generator で生成したスペクトログラム画像を図 3 に示す．GAN の定量評価として Fréchet Inception Distance(FID)を用いた．比較モデルとして，TactGAN と同様のモデル,DCGAN と WaveGAN，更に SE Block を用いない Residual Block のみの Generator 構造での学習を FID 値によって比較を行った．その結果を表 1 に示す．

表 1 FID 値による生成モデル比較

Method	FID
提案手法 (ResBlock + SEBlock)	17.69
ResBlcok	20.18
TactGAN	63.21
DCGAN	37.86
WaveGAN	22.32

5 まとめ

本研究では，実世界から取得したオブジェクト画像からオブジェクトの打撃音を生成することで，仮想環境内での動的な音を生成する手法について提案した．打撃音をスペクトログラム画像として扱うことで GAN による画像生成を可能にし，データセットに限定されない打撃音の生成が可能になった．今後の課題としては，GAN の入力をマテリアルラベルではなく，Encoder によって学習された中間表現を用いることで，より画像に適合した打撃音の生成ができると考えられる．

参考文献

[1] J. Xue, H. Zhang, K. Dana, Deep Texture Manifold for Ground Terrain Recognition, In CVPR, 2018, pp. pp. 558-567.

[2] Y. Masuyama, K. Yatabe and Y. Oikawa, “Griffin-Lim like phase recovery via alternating direction method of multipliers,” *IEEE Signal Process. Lett.*, 第 巻 26, 第 1, p. pp.184-188, 2019.

[3] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, “SLIC Superpixels,” *EPFL Technical Report*, 第 149300, 2010.

[4] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, W. T. Freeman, Visually indicated sounds, 2016, p. 2405–2413.

[5] Miyato, T., Koyama, M., “cGANs with Projection Discriminator,” In ICLR, 2018.

[6] J. Hu, L. Shen, and G. Sun., “Squeeze-and-Excitation Networks,” CoRR abs/1709.01507, 2017.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.