5C-05

クロスモーダル学習による時間情報を考慮した楽器音からの演奏動画生成

中川 智愛† 井上 勝文‡ 吉岡 理文‡

大阪府立大学 大学院工学研究科† 大阪公立大学 大学院情報学研究科‡

1. はじめに

音楽情報から，その音楽情報に適した画像を生成する研究の中で，楽器音から楽器の奏者画像を生成するタスクがある．従来研究では時間情報を含む音楽情報より 1 枚の画像を生成している[1][2][3]．しかし，1 枚の画像には音楽情報の持つ時間情報までは考慮されていないという問題がある．そこで本研究では，生成画像にも時間情報を持たせるために，音楽情報から演奏動画を生成するタスクに拡張する．具体的には，単純に短時間音楽情報から動画フレームを順次生成するのではなく，入力する音楽情報に時間情報ラベルを埋め込みつつ 1 度に複数枚の画像を生成する．この生成された画像の前後関係を考慮することで，奏者の動作が滑らかに変化する動画フレームを生成する手法を提案する．

2. 従来手法

従来研究で提案されている CAR-GAN[1]を図 1 に示す．CAR-GAN は 2 段階の学習によって音楽情報から演奏画像を生成する．ここで，A, I, I', I''はそれぞれT秒から$T+0.5$秒の GT（Ground Truth）のメルスペクトログラム画像とT秒目の GT のフレーム画像，1 段階目の生成画像，2 段階目の生成画像である．また，G_1, G_2の構造は同一で，重みは共有されている．さらに，L_a, L_i, L_rはそれぞれ楽器の種類を表すクラスラベル，生成画像の予測クラスラベル，そしてこれらの差分から得られる残差クラスラベルである．まず，Generator G_1にAとL_aを入力し，I'を得る．次にこのI'を Classifier Cに入力することで$L_{i'}$を求める．そして，L_aとの差分であるL_rとI'をG_2に入力し，最終的な出力となるI''を得る．学習時には Discriminator Dを用いて敵対的学習を行う．

3. 提案手法

本研究ではT秒から$T+0.5$秒のメルスペクトログラム画像から$T+0.1$秒から$T+0.4$秒目までの 4 枚の画像を同時に生成する．図 2 は CAR-GAN をベースとした提案手法のネットワーク図である．CAR-GAN と異なる点は 2 つあり，1 つ目は入力データであるAに時間情報ラベルの埋め込み，2 つ目は Discriminator D_vの追加である．ただし，D_iは図 1 におけるDと同一のネットワークである．

1 つ目の相違点のAに対する処理については 2 種類の方法があり，図 3 に1秒から1.5秒のAから1.1秒目のフレーム画像を生成する場合の例を示す．図 3(a)の TSE（Time Segment Emphasis）では，$T+0.1$秒目のフレーム画像を生成する場合に，Aの RGB の 3 チャネルに対して$T+0.05$秒から$T+0.15$秒の区間の値をそれぞれ 2 倍にしている．一方，図 3(b)の TSL（Time Segment Label）では TSE と同区間を 1，それ以外の区間を 0 で埋めた層を追加し，4 チャネルのデータとする．これらの処理を行なった$A_{1\sim4}$をそれぞれG_1に入力し，$I'_{1\sim4}$を得る．

次に，2 つ目の相違点である生成画像の時系列判定を行うD_vについて説明する．4 枚の生成画像に対して，時系列として正しい順に入力されたものに対して Real，シャッフルした並びのものに対して Fake と識別を行うことで，奏者の動作における前後関係の学習を行う．

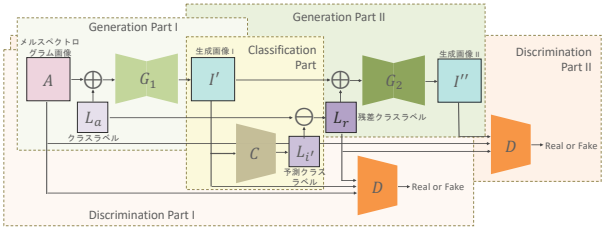

図 1 CAR-GAN[1]のネットワーク図

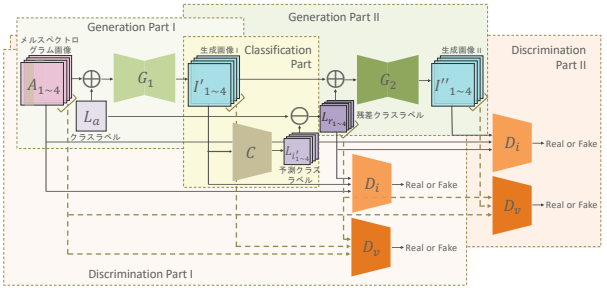

図 2 提案手法のネットワーク図

"Cross-modal Learning-based Musical Performance Movie Generation from Instrumental Sound by Considering Time Consistency"

†Graduate School of Engineering, Osaka Prefecture University

‡Graduate School of Informatics, Osaka Metropolitan University

表 1　定量評価

		Sax	Flute	Bassoon	Horn	Clarinet	Cello	Trumpet	Oboe	Trombone	Violin	Double Bass	Tuba	Viola
Ours (TSE)	FID	**271.004**	**193.625**	185.607	**217.054**	**113.797**	48.364	164.341	**324.280**	130.690	171.442	**76.528**	92.570	**149.816**
	LPIPS	**0.227**	0.298	0.351	**0.197**	0.443	0.170	0.139	0.161	0.223	0.273	0.362	0.207	0.172
Ours (TSL)	FID	271.721	204.117	**176.342**	268.475	264.480	**42.450**	**110.192**	364.821	**123.833**	**147.829**	97.253	**92.539**	163.219
	LPIPS	0.236	0.307	0.354	0.216	0.446	0.159	**0.136**	**0.155**	0.225	0.220	**0.188**	**0.188**	0.175
CAR-GAN	FID	315.263	227.307	186.226	223.038	286.463	45.717	203.585	409.218	134.553	236.821	130.351	94.713	242.119
	LPIPS	0.272	**0.215**	**0.225**	0.222	**0.326**	**0.141**	0.138	0.199	**0.185**	**0.164**	0.242	0.215	**0.143**

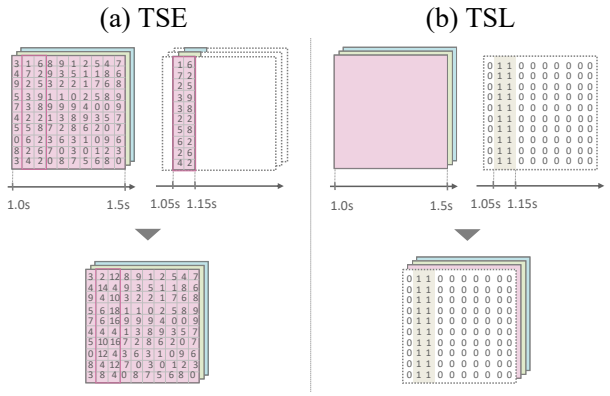

図 3　入力データの処理例

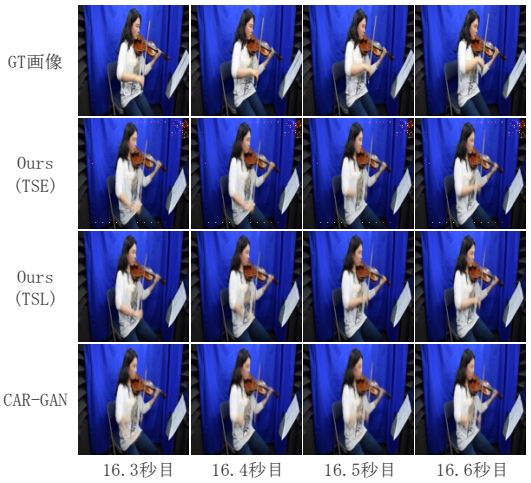

図 4　Violin の定性評価

4. 実験

4.1 実験条件

実験には学習とテストで奏者が一致しているデータを使用し，13 種類の楽器クラスを持つ Sub-URMP データセットのうち 52749 枚のメルスペクトログラム画像とフレーム画像を用いた．メルスペクトログラムはサンプリング周波数が 44.1kHz，長さ 2048，移動幅 512 のハニング窓で求めた．提案手法として TSE と TSL の 2 種類を入力した結果と従来手法である CAR-GAN による結果を比較した．バッチサイズは 4 とし，200 エポックずつ学習を行なった．

4.2 実験結果

表 1 に生成画像と GT 画像において，画像群間の距離を表す FID と，画像間の距離を表す LPIPS の平均値による定量評価を示す．ただし，太字は最も精度が高い値を表している．FID については全ての楽器において，提案手法が従来手法の結果を上回った．また，提案手法の中でも TSE の方が良い結果となった．LPIPS については従来手法よりも高精度，もしくは同等の結果となった．これらから，メルスペクトログラム画像への時間情報ラベルの埋め込みと画像の複数枚同時生成によって，生成精度が上がったと考えられる．

次に，図 4 に Violin の定性評価を示す．提案手法において，TSE と TSL はどちらも GT 画像と同様に腕が下から上に動いていることがわかる．一方で，従来手法においては腕の部分がぼやけている画像が生成されており，時間経過に対して画像内の奏者の動きの変化が少ない．これらからも，提案手法の有効性を確認できる．

5. おわりに

本研究では楽器音のメルスペクトログラム画像から楽器演奏動画の生成するネットワークを提案した．実験結果より，メルスペクトログラム画像に時間情報ラベルを埋め込み，同時に複数枚の画像を生成することで，高精度な奏者の動きを生成することができた．今後は楽器以外のデータへの適用により，提案手法の妥当性のさらなる検証が課題として挙げられる．

参考文献

[1] B. Duan et al. “Cascade attention guided residue learning gan for cross-modal translation” In Proc. of ICPR, 2020
[2] L. Chen et al. “Deep cross-modal audio-visual generation.” In Proc. of Thematic Workshops’17
[3] W. Hao et al: “Cmcgan: A uniform framework for cross-modal visual-audio mutual generation”, In Proc. of AAAI, 2018