H-020

番組ホームページ制作支援のための代表画像選定技術の検討

On the selection of thumbnail images for TV program website production support

前澤 桃子† 遠藤 伶† 望月 貴裕†
Momoko Maezawa Rei Endo Takahiro Mochizuki

1. はじめに

放送局では，視聴者の番組接触率向上を目的として，番組ホームページ（HP）の充実化が進んでいる．番組 HP は，閲覧者が大まかな番組内容を把握できるよう，番組映像から選んだ代表画像を掲載することが多い．番組 HP を制作する際，まず，番組構成にしたがって重要な映像区間を選ぶ．次に，その映像区間毎に番組 HP に掲載する代表画像を選定する．この代表画像選定作業は映像区間全体の確認が必要であり，大きな労力を要する．

本研究の目的は，番組編集担当者に代表画像の候補を自動提示する技術の実現である．それにより，番組編集担当者の代表画像選定の作業時間削減が期待できる．

本稿では，ニューラルネットワーク（NN）を用いた画像のスコア付けにより，番組映像から代表画像を選定する手法を提案する．NN の学習には，番組の代表画像としての適性の有無が付与された大規模な画像データセット（以下，番組代表画像データセット）が必要となる．しかし，番組映像の各フレームに対して，番組編集担当者が番組代表画像としての適性を評価するため，多くの作業時間がかかり大規模な画像データセットを構築することは難しい．本手法では，代表画像選定に類似するタスクのためのデータセットを併用することで少量の番組代表画像データセットで学習可能となる．データセットの併用学習に対応した 2 種類の NN を検討し，評価実験により本手法の有効性を検証する．

2. 関連研究

AVA database [1]は，視覚的な芸術性に基づき画像が high / low の 2 クラスに分けられたデータセットである．Jin らは，AVA database を学習データとして，画像を high / low に分類する NN を考案した[2]．まず，特徴抽出ネットワークが画像特徴を計算し，次にクラス分類ネットワークが画像特徴から 2 クラスの確率分布を計算する（**図 1**）．

AVA database は，画像の視覚的な芸術性に関するデータセットであり，代表画像選定のためのデータセットではない．しかし，番組編集担当者による代表画像選定作業では，色彩や構図などの芸術性に関連した要素が考慮されている．そのため，代表画像選定タスクとの類似度は高い．

3. 提案手法

本稿では，NN を用いた画像のスコア付けにより，番組映像から代表画像を選定する手法を提案する．本手法では，類似タスクのデータセットの併用により少量の番組代表画像データセットで学習することができる．

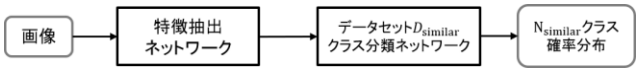

図 1　既存手法[2]の概要（$\mathrm{N}_{\mathrm{similar}} = 2$）

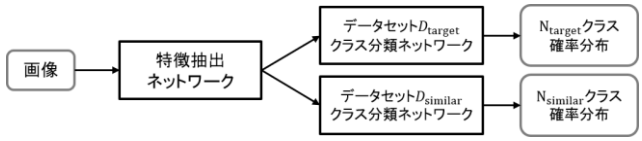

図 2　並列クラス分類モデルの概要

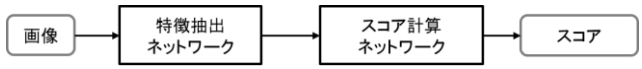

図 3　単一スコア回帰モデルの概要

本章では，複数のデータセットを学習するための，並列クラス分類モデルと単一スコア回帰モデルについて述べる．

どちらのモデルも 2 つのネットワークで構成される．第 1 のネットワークは，既存手法[2]と同様の構造の，画像特徴を計算するための特徴抽出ネットワークである．第 2 のネットワークはモデルによって異なる．並列クラス分類モデルはクラス分類ネットワーク，単一スコア回帰モデルはスコア計算ネットワークを用いる．

次節で各モデルの詳細を述べる．メインタスクのためのデータセットD_{target}のクラス数を$\mathrm{N}_{\mathrm{target}}$，類似タスクのためのデータセット$D_{\mathrm{similar}}$のクラス数を$\mathrm{N}_{\mathrm{similar}}$とする．

3.1 並列クラス分類モデル

既存手法[2]をもとに，特徴抽出ネットワークのあとのクラス分類ネットワークをデータセット毎に用意して，確率分布を出力する NN を，並列クラス分類モデルとする．**図 2** にこのモデルの概要を示す．入力は画像であり，出力は各クラスの確率分布である．

特徴抽出ネットワークは，画像の視覚的な芸術性を考慮する既存手法のネットワーク構造を利用する．GoogLeNet [3]の先頭から 1/3 の層と Batch Normalization [4]を用い，1024 次元の特徴ベクトルを生成する．クラス分類ネットワークは，既存手法と同様に全結合層 2 層からなり，データセットのクラス数に応じて出力部を改変する．D_{target}のクラス分類ネットワークは$\mathrm{N}_{\mathrm{target}}$クラスの確率分布，$D_{\mathrm{similar}}$のクラス分類ネットワークは$\mathrm{N}_{\mathrm{similar}}$クラスの確率分布を出力する．

3.2 単一スコア回帰モデル

並列クラス分類モデルのようにデータセット毎の各クラスの確率分布を出力するのではなく，すべてのデータセットの画像に対して共通のスコアを出力する NN を，単一ス

† NHK 放送技術研究所
NHK Science & Technology Research Laboratories

コア回帰モデルとする．**図 3** にこのモデルの概要を示す．入力は画像であり，出力はその画像に対するスコアである．出力されるスコアから閾値に基づいて分類される．

特徴抽出ネットワークは並列クラス分類モデルと同様である．スコア計算ネットワークは全結合層 2 層からなり，Sigmoid 関数によってスコアは 0.0 以上 1.0 以下の値となる．

単一スコア回帰モデルの学習においては，各クラスに 0.0 以上 1.0 以下の正解スコアを付ける．各クラスの正解スコアは，等間隔にする必要はなく，データセットの性質や利用目的に応じて適切に設定すべきである．例えば，データセットD_{target}において$\mathrm{N}_{\mathrm{target}} = 3$のとき，クラス 1 の正解スコアを 1.0，クラス 2 を 0.7，クラス 3 を 0.0 とすることができる．代表画像としての適性等，クラスに順序関係がある場合は単一スコア回帰モデルが有効であると考える．

4. 評価実験

4.1 番組代表画像データセット

NHK 番組映像の各フレームに対して，番組編集担当者が番組代表画像としての適性を great / good / bad の 3 段階で評価したデータセットを作成した．これを番組代表画像データセットとして利用した．

great は顔のアップ等，実際に番組代表画像として使えるフレームである．bad はカメラの動きの途中で被写体が切れていたり，ぶれていたり，あるいは明るすぎたりする等，明らかに番組代表画像として使えないフレームである．good はそのどちらにも当てはまらないフレームがすべて含まれている．中間クラスである good が曖昧なため，クラス分類が難しいと考えられる．なお，番組代表画像として使える画像かどうかの判断には，被写体が番組の中で重要な登場人物かどうか等，視覚的な情報だけではわからない基準も存在する．

1 枚 1 枚人手で評価をつけているため 6500 枚程度しかなく，学習データとしては十分でない．

4.2 実験概要

既存手法[2]，並列クラス分類モデル，および単一スコア回帰モデルの比較実験について述べる．

代表画像の選定が目標であり，great / good / bad の 3 クラスのうち great を見分けることが重要なタスクとなるため，great ラベルの判別精度を評価指標とした．なお，単一スコア回帰モデルにおいてはスコア 0.75 以上を great とした．

評価データは番組代表画像データセット中の約 1 割（676 枚）とし，9 割を学習に用いた．類似タスクのデータセットとして AVA database を利用した．AVA database 中の 25553 枚を学習データとした．なお，学習する際，既存手法は番組代表画像データセットのみを用い，並列クラス分類モデル，単一スコア回帰モデルは番組代表画像データセットと AVA database を併用した．

単一スコア回帰モデルの正解スコアについては，AVA database では high を 1.0，low を 0.0，番組代表画像データセットでは great を 1.0，good を 0.5，bad を 0.0 として学習した．

表 1　評価結果

	適合率	再現率	F 値
既存手法[2]	0.49	0.48	0.48
並列クラス分類モデル	0.59	0.21	0.31
単一スコア回帰モデル	0.78	0.34	0.47

4.3 評価結果

great ラベルの判別精度について，適合率，再現率，F 値を**表 1** に示す．適合率が低い場合，番組編集担当者は，代表画像として不適切な画像を多く確認することになる．作業時間の削減という観点では，great のとりこぼしが少ない再現率が高いモデルより，great でない画像が選ばれにくい適合率が高いモデルが望ましい．したがって，適合率が 78%となり，F 値が既存手法と同等である単一スコア回帰モデルが実運用には適していると確認できる．

単一スコア回帰モデルでは，特徴抽出ネットワークのあと，どちらのデータセットでも共通のスコア計算ネットワークを用いて画像に対するスコアを出力している．当然，great と high，bad と low がまったく同じ意味を持つわけではないが，それぞれ似た意味合いを持つため，適切に学習できたと考えられる．異なるタスクのデータセットであっても，類似のタスクであれば併用可能なことがわかった．

5. おわりに

本稿では，NN を用いた画像のスコア付けにより番組映像から代表画像を選定する手法を検討した．画像の視覚的な芸術性に関するデータセットの併用により少量の番組代表画像データセットで学習可能な仕組みを実現した．評価実験では，番組代表画像データセットを利用して手法の有効性を示した．今後は，代表画像選定の精度向上を目指すとともに，実用化を進めていく．

謝辞

番組代表画像データセットの作成にご協力いただいた株式会社 Preferred Networks，NHK 放送総局デジタルセンターに感謝の意を表する．

参考文献

[1] N. Murray, *et al.* “AVA: A large-scale database for aesthetic visual analysis.” In Proc. of CVPR, 2012.

[2] X. Jin, *et al.* “ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation.” IET Computer Vision 13.2 (2018): 206-212.

[3] C. Szegedy, *et al.* “Going deeper with convolutions,” In Proc. of CVPR, 2015.

[4] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In Proc. of ICML, 2015.