情報処理学会論文誌　コンシューマ・デバイス & システム　Vol.9　No.2　11–21 (May 2019)

コンシューマ・デバイス論文

深層学習による美的評価エンジンの開発と構図推薦カメラへの実装

井上 義隆[1,a)]　松村 沢磨[2,b)]　深澤 佑介[1,c)]　山田 和宏[1,d)]

受付日 2018年10月9日, 採録日 2019年1月31日

概要：本研究では写真を美的評価するエンジンを深層学習で構築し，デジタルカメラ型デバイスへ実装した．デバイスでは深層学習の推論処理をリアルタイムに動作させ，撮影すべき構図を撮影者に推薦する機能を実装した．構図推薦カメラは初心者の撮影技術をサポートするだけでなく，熟練者に対してもシャッターチャンスの気づきや構図を追求する機会を与えることができる．

キーワード：深層学習，画像認識，デジタルカメラ，構図推薦

Development of Aesthetics Evaluation Engine based on Deep Learning and Implementation into Composition Recommendation Camera

Yoshitaka Inoue[1,a)]　Takuma Matsumura[2,b)]　Yusuke Fukazawa[1,c)]　Kazuhiro Yamada[1,d)]

Received: October 9, 2018, Accepted: January 31, 2019

Abstract: In this paper, we developed an aesthetics evaluation engine based on deep learning, and implemented it into a digital camera which is able to recommend photo compositions. The device inferences aesthetics and recommends compositions to the photographer in real time. The proposed engine and device can not only support photography techniques for amateur photographers, but also give opportunities of perfect shot and composition adjustment for expert photographers.

Keywords: deep learning, image recognition, digital camera, composition recommendation

1. はじめに

近年，デジタルカメラに内蔵されている受光センサや手ぶれ補正技術の性能向上により，照明が暗い環境下であっても，ノイズを少なく抑えて，手ぶれによる画像の品質劣化を低減できるようになった．また，顔や瞳の検出とオートフォーカスによって，顔領域のピントボケを軽減できるようになった．さらに 1 秒あたりの撮影可能枚数の増加は動物やスポーツ等の動体を被写体とした場合にシャッターチャンスを抑えやすくなった．

デジタルカメラの性能向上により，きれいな写真を撮影することが可能になったものの，必ずしもそれだけで本質的に「美しい」と感じる写真を撮影できるとは限らない．人間が写真を「美しい」と判断する際には，正確な露光やピントという光学的観点だけではなく，写真の構図や内容に基づく情緒的観点も影響する．後者は人物の魅力的な表情，大自然の中の非日常的な絶景，構図の幾何学的な精密さ，二度と訪れない決定的瞬間等があげられる．色彩を考慮しても，街中の色鮮やかな看板広告よりは，花畑や夕焼けの方が一般的に好まれるかもしれない．モノクロ写真では，コントラストが強く幾何学的な写真も，ノスタルジックな淡い写真も等しく好まれるかもしれない．

従来の画像処理技術で計算可能な低次の画像特徴量と，人間の高次の情緒的観点との間にはギャップがあることが指摘されている [1]．一方，近年では，大量の写真に対して

[1] 株式会社 NTT ドコモ
NTT DOCOMO, INC., Chiyoda, Tokyo 100–6150, Japan
[2] ドコモ・テクノロジ株式会社
DOCOMO R&D Center, Yokosuka, Kanagawa 239–8536, Japan
[a)] yoshitaka.inoue.ye@nttdocomo.com
[b)] matsumuratak@nttdocomo.com
[c)] fukazawayu@nttdocomo.com
[d)] yamadakazu@nttdocomo.com

図 1　構図推薦カメラ本体．（上）正面，（下）背面，（右）内部
Fig. 1　Composition recommendation camera. (Top) Front, (Bottom) Back, (Right) Inside.

人間による美的観点での主観品質評価結果が記録されており，かつ一般公開されている AVA データセット [2] を用いて，深層学習によって光学的観点と情緒的観点とを同時に抽出する手法が研究されている [3], [4]．これらの美的評価手法をデジタルカメラ型のデバイスに実装し，撮影者に対してリアルタイムに美的評価と構図推薦を行う製品は存在しない．

本研究では，AVA データセットを用いて深層学習による美的評価エンジンを開発し，写真の美しさを 3 段階で判別することを可能にした．撮影済み写真に対してだけではなく，撮影現場でリアルタイムに美的評価を実行できるように，エンジンを軽量な GPU マシン上に実装し，レンズとセンサと組み合わせて，手持ちで撮影できる構図推薦カメラを開発した（**図 1**）．構図推薦カメラはライブビュー画像を逐次評価するだけなく，より高い評価値が得られる構図を撮影者に推薦することができる．

本研究の貢献ポイントは以下のとおりである．撮影機能と美的評価の処理をデバイスローカルで完結させ，リアルタイムな動作を可能にした．さらに，撮影初心者が犯しがちな構図ミスを指摘したり，熟練者に対しても気づきや構図を追求する機会を与えるような構図推薦機能を実装した．

以下，2 章では関連研究について述べ，3 章で構図推薦カメラに求められる要求条件とアプローチについて述べる．4 章では美的評価エンジンの構築を，5 章では構図推薦カメラの実装について，ハードウェア構成，プロセス並列化，構図推薦処理の順に述べる．6 章で美的評価エンジンと構図推薦カメラの評価について述べる．最後に 7 章でまとめる．

2. 関連研究

2.1 特徴量設計によるアプローチ

写真を評価する際に構図は重要な要素である [5], [6], [7]．代表的な構図ルールとして日の丸構図，三分割構図，黄金分割構図，対角構図等があげられる．このような構図ルールに従った写真は美しさや安定感を与えるといわれている．構図や色を特徴量として抽出し，ルールベースで写真を美的評価する手法が研究されている．家田ら [8] は入力された写真を解析し，色による顔の検出，色彩的に際立つ領域，三角形，水平線，対角線，遠近法消失点を考慮して，構図ルールに近づくようにトリミング領域を推薦する手法を提案している．志津野ら [9] の手法は入力された写真から SURF 特徴量を抽出し，あらかじめ用意した構図ルールのテンプレートと比較して，従うべき適切な構図ルールを撮影者へ推薦する．撮影者が当該構図ルールに合わせるように撮影することで良い写真が得られる．Bhattacharya ら [10] は構図ルールに基づいた特徴量を抽出し，別途用意した 632 枚の美的評価済み写真を用いて，Support Vector Regression によって美的評価値を推定するモデルを構築している．しかしながら，特徴量設計のアプローチで計算可能な構図や色のような低次の画像特徴量と，人間の高次の情緒的観点との間にはギャップがあることが指摘されている [1]．

2.2 Deep Learning によるアプローチ

25 万枚の写真を含み，1 枚あたり 78 人から 549 人の評価者によって 10 段階の主観評価が記録されている AVA データセット [2] を用いた深層学習による美的評価手法が研究されている．Lu ら [3] の手法は，Convolutional Neural Network（以下，CNN）の入力層が規定する解像度になるように入力画像を変換し，画像全体と細部をそれぞれ分割して同時にネットワークへ入力する．従来の CNN は入力層が規定するサイズに入力画像をリサイズしなければならないため，画像が歪むという問題がある．しかしながら，入力画像の解像度やアスペクト比は，撮影条件やカメラの設定，撮影後の編集に依存するためあらかじめ想定することができない．Lu らは次の手法でこの問題を解決している．画像の全体の成分として，入力画像をクロップ（入力画像の短辺の長さで中央を正方形をトリミング），ワープ（長辺のみを縮小），パディング（長辺を合わせて，短辺方向に生じる隙間を黒で埋める）し，さらにそれぞれを正方形の入力層に合わせてリサイズする．これは構図やグラデーションを情報として含む．また，画像の細部の成分として，画像の部分領域をランダムに選択し，入力層に合わせてトリミングする．これは画像のテクスチャ情報を残している．それぞれを個別のネットワークに入力し，最終レイヤで統合する．AVA データセットを用いて，評価スコ

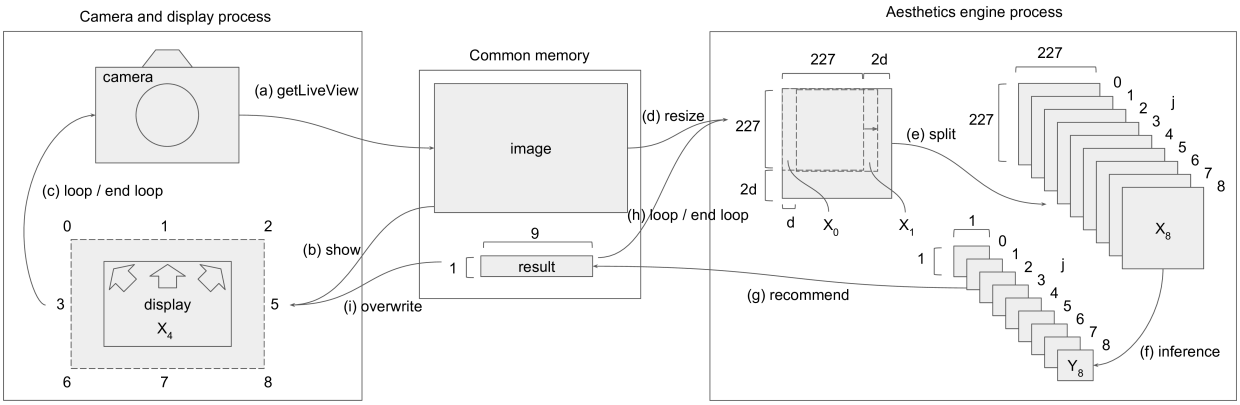

図 2　構図推薦カメラと美的評価エンジンのアルゴリズム

Fig. 2　Algorithm of composition recommendation camera and aesthetics evaluation engine.

アの 2 値分類（高スコア，低スコア）で精度 74.46%を達成している．Kao ら [4] は画像のセマンティクスを考慮して CNN で美的評価を行っている．セマンティクスとは画像の内容が人物なのか，建築物なのか，食べ物なのか，といった情報である．セマンティクスを考慮することで 2 値分類精度を 79.08%まで向上させている．

しかしながら，AVA データセットでは画像の評価スコアが平均値付近に集中する正規分布に近い分布をとるため，Lu らおよび Kao らの手法のように評価スコアの平均値を閾値とする 2 値分類モデルとして設計するのは望ましくない．平均値付近の写真は枚数が多いにもかかわらず，美的評価の判断は微小な差異の影響を受けやすく，2 値のいずれかに一方に極端に振れてしまう恐れがあるためである．

2.3　アプリケーション

写真の品質を自動的に評価する，あるいは品質の高い写真を撮影できるようにサポートする機能を持つアプリケーションがスマートフォンや PC 向けに提供されている．Picscore [*1] はユーザが入力した撮影済み写真 2 枚のうち美的評価が高い方を推薦する．写真の解析に約 30 秒を必要とするため，撮影時リアルタイムでの動作は実現されていない．Adobe Lightroom CC [*2] は撮影後の写真群の中から品質の高い写真を自動的に選択する．しかしながら，デバイス上での撮影時リアルタイム動作は実現されていない．構図カメラ [*3] および PoseCam [*4] はいずれも撮影時に構図ルールをライブビュー上にオーバレイ表示させることで，構図ルールに従った撮影をサポートする．しかしながら，構図ルール以外の要素は考慮されておらず，自動的な美的評価も実現されていない．ガイドカメラ [*5] は撮影時にお手本となる写真を透過オーバレイ表示させて，似たような構図で撮影することをサポートする．しかしながら，お手本写真はユーザが用意する必要があり，自動的な美的評価も実現されていない．

3.　問題設定

3.1　要求条件

本研究は美的評価の精度をある程度に保ちつつ，撮影作業をリアルタイムにかつインタラクティブにサポートするデバイスの開発を目的とする．これにより初心者向けには撮影技術の向上を，熟練者向けにはシャッターチャンスの気づきや良い構図を追求する機会を与える．具体的に本研究で実現する美的評価エンジンおよび構図推薦カメラの要求条件は以下のとおりとする．ただし，以下ではリアルタイム性の要求条件を定義する際に**図 2** を参照するが，図 2 の詳細は 5.3 節および 5.4 節で述べる．

多段階評価　評価スコアが平均値付近の画像による影響を避けるために，3 段階以上の多段階評価あるいは評価値の回帰が可能な美的評価モデルを作成する．ただし，実環境での動作速度や実運用の利便性を検討した結果，3 値分類で十分とする．

リアルタイム性　画像の取得と表示の処理は 25.0 fps 以上とする（図 2 (a) → (b) → (c) のループ）．撮影者がスムーズに構図を探索し，シャッターチャンスを逃さないよう，実風景と表示内容にラグが存在してはならない．また，フレームレートは標準的な Blu-ray Disc 映像と同等であることを基準とする．推論処理は 2.5 fps 以上を目標とする（図 2 (e) → (f) → (g) → (h) のループ）．実環境が 1 秒の間に大きくは変化しないと仮定している．

[*1] https://itunes.apple.com/jp/app/id1082932364
[*2] https://blogs.adobe.com/jkost/2018/06/new-features-and-updates-in-lightroom-cc-web.html
[*3] https://play.google.com/store/apps/details?id=com.makaroni.composicamera
[*4] https://play.google.com/store/apps/details?id=com.wagachat.composecamera
[*5] https://play.google.com/store/apps/details?id=jp.co.rugle.guidecamera

構図推薦　実風景のうち，撮影画角の周辺の画像を含めて美的評価の高い写真が得られる構図を探索して撮影者に推薦する．撮影技術向上のためには撮影者が気づかない構図を撮影者に提示する必要がある．

3.2　アプローチ

前節の要求条件を次のアプローチによって解決する．まず，AVA データセットの各画像のスコアをもとに 3 つのラベルに分類し，3 値分類問題としてモデルを構築する．次に，異なるデータセットで学習済みのモデルをもとに美的評価エンジンをファインチューニングすることで精度を改善する．また，リアルタイム性については以下の 2 つの要素から実現する．(1) 通信遅延を発生させないためにデバイスローカルで処理を完結させる．(2) 表示と推論をプロセス並列化することでスムーズな操作性を実現する．さらに，構図推薦については，ライブビュー画像を重畳するように 9 枚の画像に分割して，9 枚の画像に対して同時に美的評価を行い，評価結果をもとに最適構図で撮影するための移動方向を撮影者に提示することで解決する．

上記それぞれのアプローチについては，3 値分類の定義を 4.1 節で述べ，モデル構築を 4.2 節で述べ，リアルタイム性のための (1) デバイスを 5.1 節，(2) プロセス並列化を 5.3 節でそれぞれ述べ，構図推薦アルゴリズムを 5.4 節で述べる．

4.　美的評価エンジンの構築

4.1　3 値分類

多段階評価の要求条件を満たすために下記のとおりデータセットを整理して 3 値分類問題として定義する．AVA データセット [2] は，画像のインデックスを $i \in I$，画像を X_i として，画像 X_i それぞれについてスコア $\{s \in \mathbb{N} \mid 1 \leq s \leq 10\}$ を付与した評価者の人数分布が $n_{i,s}$ として記録されている．画像 X_i ごとに評価者人数 $N_i = \Sigma_{s=1}^{10} n_{i,s}$ は異なり，AVA データセットにおいては最小で 78 人，最大で 549 人であった．画像 X_i ごとに評価者で平均したスコア $S_i = \frac{1}{N_i}\Sigma_{s=1}^{10} n_{i,s}s$ を算出する．画像で平均したスコア $\mu = \frac{1}{|I|}\Sigma_i S_i$ と，標準偏差 $\sigma = \sqrt{\frac{1}{|I|}\Sigma_i (S_i - \mu)^2}$ とを基準として，ラベル $Y_i \in \{\mathrm{High}, \mathrm{Middle}, \mathrm{Low}\}$ を付与する．具体的にはラベル High は $\mu + \sigma \leq S_i$，Middle は $\mu - \sigma \leq S_i < \mu + \sigma$，Low は $S_i < \mu - \sigma$ を満たす画像 X_i に付与する．

分類された画像の枚数はラベルごとに異なる．いずれのラベルも画像の枚数が同数程度になるようにランダムにサンプリングし，サンプリングされた画像を 90%と 10%に分割して，それぞれをトレーニングデータ，テストデータとする．ただし，汎化性能を向上させるために，トレーニングデータについては画像 X_i を水平方向に反転した画像も追加する．被写体が左右非対称であることを事前知識として持っている場合（看板の文字，衣服のボタン等），見た目に違和感を受けるかもしれないが，そのような写真の数はごく一部であって，動植物，人物，建築物，都市や自然の風景については左右反転しても気づかず，美的品質に影響しないと考えられる．なお，テストデータには水平方向に反転した画像は含まれない．ラベルごとの画像枚数を**表 1** に示す．左列から，元の画像枚数，トレーニングデータの画像枚数，テストデータの画像枚数である．

表 1　データセットの 3 値分類定義
Table 1　Definition of dataset classified by three labels.

Label	#original images	#training images	#test images
High	39,580	68,959	3,888
Middle	177,643	68,951	3,784
Low	38,307	68,979	3,839

表 2　代表的なネットワーク構造と層数
Table 2　Networks and their layers.

Network	Layers
AlexNet [11]	8
VGG [13]	16, 19
GoogLeNet [14]	22
ResNet [15]	152

4.2　モデル構築

本研究では AlexNet [11] のネットワーク構造を採用する．**表 2** に示すように，その他の代表的なネットワーク構造の中で層が浅い [12], [13], [14], [15] ためリアルタイム動作を実現可能なこと，かつファインチューニングに利用するリファレンスモデルが公開されていることの両面から AlexNet を選択した．リファレンスモデルとは，1,400 万枚の画像を含む ILSVRC2012 データセットを用いて，セマンティクス観点での分類問題を，AlexNet のネットワーク構造で学習した Caffe モデルファイル（caffe_reference_imagenet_model）である[*6]．ILSVRC2012 は画像の枚数が AVA データセットよりはるかに多いこと，かつアマチュア写真家による投稿画像以外の幅広い内容の画像をカバーしていることから，リファレンスモデルの各層における重みパラメータを，提案モデルにおける初期値として与えてから美的評価を再学習すること，すなわちファインチューニング（以下，FT）を行うことで，美的評価の精度向上が期待できる．

前処理にかかる計算量を削減するために，画像のワープのみを実行してサイズを入力層に合わせる．すなわち長辺方向を縮小して，縦 227，横 227，3 チャンネルのサイズに統一する．また，本研究では AlexNet のネットワーク構造の出力層の次元を 3 に変更する．4.1 節で定義したラベルの 3 値分類を学習する．出力される値は画像 X_i ごとの，

[*6] http://caffe.berkeleyvision.org/model_zoo.html

3 種類のラベルそれぞれへの所属確率 $P(Y_i \mid X_i)$ となる．また，1 枚の画像について 3 カテゴリへの所属確率の和は $\Sigma_{Y_i} P(Y_i \mid X_i) = 1$ となる．

モデル構築環境は，Amazon Web Service の EC2 インスタンスを用いる．インスタンスタイプは g2.2xlarge [*7] とし，Ubuntu にインストールした Caffe を用いる．

5. 構図推薦カメラの実装

5.1 ハードウェア構成

リアルタイム性の要求条件を満たす要素 (1) として，美的評価エンジンを NVIDIA Jetson TX1 [*8]（以下，TX1）へ実装し，ローカルで推論処理を行う．TX1 は NVIDIA GPU Tegra X1 を搭載した小型軽量のコンピュータである．CPU，GPU，メモリ，ストレージを含む TX1 モジュールと，Wi-Fi，HDMI 端子，GPIO，電源等を含むキャリアボードとで構成されている．TX1 モジュールを別売りの小型のキャリアボード Orbitty Carrier for NVIDIA Jetson TX1 [*9]（以下，Orbitty Carrier）に載せ替えた．Orbitty Carrier は縦 5 cm，横 8.5 cm である．

Orbitty Carrier と LiPo バッテリ，小型の液晶ディスプレイを接続した．アクリル板を加工してカメラの筐体を作成し，TX1，LIPO バッテリを筐体内に収納し，ディスプレイは筐体背面側のに固定した．図 1 右図は液晶ディスプレイを取り外した状態の筐体内部である．筐体内部の左側が TX1 モジュールと Orbitty Carrier である．

レンズユニットは SONY ICLE-QX1 [*10]（以下，QX1）を用いた．QX1 はセンサとシャッターボタンのみ備えており，ライブビューを確認するためのディスプレイを搭載しておらず，設定変更操作に必要なダイヤルやボタンも備えていない．本研究では QX1 をレンズユニットとして採用し，QX1 と TX1 を接続し，TX1 と液晶ディスプレイを接続した．なお，レンズは SONY SEL35F18（以下，レンズ）を用いた．

QX1 は筐体外側に固定した．操作性を考慮して，QX1 に備わっているシャッターボタンとは別に新たなシャッターボタンを作成して筐体外側に配置した．またボタン操作検知用マイコンボードを筐体内に配置し，GPIO で Orbitty Carrier と接続した．図 1 上下図のように，QX1 およびレンズが筐体に固定されている．また，図 1 下図の右下に見えるシルバーのボタンがシャッターボタンである．

以上のハードウェア構成で，筐体の幅 19.0 cm，高さ 9.5 cm，厚さ 6.3 cm，ディスプレイからレンズの先端までの長さ 21.0 cm，全体の重量は 1,082 g となった．デジタルカメラのハイエンド機のと比較しても重量差はあまりない．SONY α 7 II のズームレンズキットはバッテリとレンズ込みで 894 g，Canon EOS 5D Mark IV のレンズキットはバッテリとレンズ込みで 1,490 g である．

5.2 ソフトウェア環境

TX1 の Ubuntu 上に，QX1 との通信，CNN による推論，表示等の処理を Python で実装した．QX1 は SONY Camera Remote API [*11] に対応している．QX1 内部に Wi-Fi アクセスポイントを立ち上げ，TX1 から Wi-Fi 接続することで，QX1 で取得したライブビューを取得したり，QX1 に対してシャッター命令を行うことができる．このような処理は HTTP 通信で行われる．TX1 上の Python から QX1 と通信できるようにライブラリを構築した．また，画像処理と表示は OpenCV を，推論は Caffe を用いた．

5.3 プロセス並列化

リアルタイム性の要求条件を満たす要素 (2) として，プロセス並列化と共有メモリによるアプローチをとる．撮影現場での活用のためには，カメラの指示に従って撮影者がカメラを動かし，カメラに新しい画像が入力されてカメラが撮影者への指示を変更する，というような撮影者とカメラとの間のインタラクティブなやりとりを遅延なくスムーズに実現しなければならない．

そのため本研究では，図 2 のように画像の取得および表示をコントロールする表示系プロセス（Camera and display process）と，美的評価を逐次行う推論系プロセス（Aesthetics engine process）とを分割し，共有メモリ空間（Common memory）を設置する．もし，2 つのプロセスを同一のプロセスで実行すると，推論の際に負荷がかかり，表示画像が固まるようなラグが生じてしまう．たとえば，構図の誘導指示に従ってカメラをスライドさせる際に，ラグが生じるたびにディスプレイ上のライブビューが一時静止してしまい，滑らかに動作しない．プロセスを分割することでスムーズな動作を可能にする．

表示系プロセスは，図 2 (a) のように取得したライブビュー画像を共有メモリに格納し，図 2 (b) のようにディスプレイにライブビュー画像を表示する．このとき，図 2 (i) のように共有メモリに推論結果が格納されていればそれをライブビュー画像に重畳表示する．図 2 (c) のように表示後には再度ライブビュー取得へ戻る．一方，推論系プロセスは，図 2 (d) のように共有メモリに画像が格納されていれば，その画像に対して美的評価を行う．また，図 2 (g) のように推論結果を共有メモリに格納する．格納後は再度共有メモリ上の画像確認に戻る．

[*7] https://aws.amazon.com/jp/ec2/previous-generation/
[*8] https://www.nvidia.com/ja-jp/autonomous-machines/embedded-systems-dev-kits-modules/
[*9] http://connecttech.com/product/orbitty-carrier-for-nvidia-jetson-tx2-tx1/
[*10] https://www.sony.jp/ichigan/products/ILCE-QX1/
[*11] https://developer.sony.com/ja/develop/cameras/

5.4　構図推薦処理

構図推薦の要求条件を満たすために，下記の構図推薦処理を実装する．図 2 (d) に示すように，QX1 から取得した画像を $227+2d$ 四方に縮小させた後に，図 2 (e) に示すように，幅 d でスライドしながら重畳するように画像を 9 分割する．227 とは 4.2 節で定義した CNN の入力画像サイズである．9 分割した画像のインデックスを $\{j \in \mathbb{N} \mid 0 \leq j \leq 8\}$，各画像を X_j とする．ただし，左上から水平方向優先で走査順にインデックスを割り当てるものとし，たとえば左上端を $j=0$，右上端を $j=2$，中央を $j=4$ とする．

次に，図 2 (f) に示すように，美的評価エンジンに 9 枚同時に入力し，ラベル 3 値への所属確率 $P(Y_j \mid X_j)$ を出力し，式 (1) で総合スコアを算出する．

$$\mathrm{TotalScore}_j = \Sigma_{Y_j} a(Y_j) P(Y_j \mid X_j) \tag{1}$$

ただし，$(a(\mathrm{High}), a(\mathrm{Middle}), a(\mathrm{Low})) = (1.0, 0.5, 0.0)$ とする．

図 2 (g) に示すように，評価結果に基づく推薦方向を result に格納する．具体的には，中央の画像 X_4 の総合スコア $\mathrm{TotalScore}_4$ があらかじめ設定した閾値 th_{OK} 以上である場合，もしくは中央の画像 X_4 の総合スコア $\mathrm{TotalScore}_4$ が，9 枚の中で最大である場合は，推薦方向 result $= 4$ とする．この場合には図 2 (i) で現在の構図のまま撮影を指示する．より高い総合スコアを得られる方向が X_4 以外に 1 個以上存在する場合は，それらの方向をすべて配列として推薦方向 result に与え，図 2 (i) でライブビュー画像に矢印で重畳表示して最適構図への移動方向を推薦する．撮影者はこの矢印の方向へカメラの向きを移動させることでより美しい写真が得られる領域を探索することができる．ただし，本来は QX1 の受光センサで広範囲の画像を取得しているが，撮影者にとっては中央部分 X_4 のみが撮影対象となる．撮影者には X_4 だけを表示して，X_4 以外の領域を非表示にしてもよい．

5.5　構図推薦カメラのアルゴリズム

5.3 節と 5.4 節の処理の全体をアルゴリズム 1, 2 に示す．表示系プロセス（CameraAndDisplayProcess）内部でライブビュー取得画像 image と推薦結果 result を，推論系プロセス（AestheticsEngineProcess）との共有メモリ上の変数として定義する．また表示系プロセスは推論系プロセスを起動（start）する．表示系プロセスのループでは，image の取得（getLiveView），表示（show）を行い，result が存在する場合には推薦結果を重畳表示する（overwrite）．また，シャッターボタンが押された場合は，撮影して記録保存する（takePictureAndSave）．推論系プロセスは美的評価を行う．推論系プロセスのループでは，新しい image が存在する場合は画像を重畳分割（split）後に推論（inference）を行い，推論結果に基づいて計算（recommend）した推薦結果を result に格納する．以上の 2 つのプロセスが image と result を逐次更新する．

Algorithm 1 CameraAndDisplayProcess

```
global image ← null
global result ← null
camera.initialize()
aestheticsEngineProcess.start()
loop
    image ← camera.getLiveView()
    display.show(image)
    if result ≠ null then
        display.overwrite(result)
    end if
    if camera.shutterButton.onPressed() then
        camera.takePictureAndSave()
    end if
end loop
```

Algorithm 2 AestheticsEngineProcess

```
cnn ← loadModel()
loop
    if image ≠ null then
        image ← resize(image)
        X[0, ..., 8] ← split(image)
        Y[0, ..., 8] ← cnn.inference(X)
        result ← recommend(Y)
    end if
end loop
```

6.　評価

6.1　多段階評価の検証

本節では多段階評価の精度と FT による改善を確認する．AlexNet のパラメータの初期値をランダムに設定して美的評価を学習したモデル（CNN），初期値としてリファレンスモデルのパラメータを与えて美的評価を再学習したモデル（CNN+FT）の 2 通りのモデルを学習した．学習時の Caffe パラメータ設定を**表 3** に示す．base_lr は複数通りの設定値（0.0001, 0.001, 0.01, 0.1）を試して，両モデルともに最終的な Accuracy が最も高かったパラメータ 0.001 を選択した．その他のパラメータはリファレンスモデルでの設定値を採用した．

ラベル $Y_i \in \{\mathrm{High}, \mathrm{Middle}, \mathrm{Low}\}$ への所属確率 $P(Y_i \mid$

表 3　Caffe パラメータ設定
Table 3　Caffe parameters setting.

Parameter	Description	Value
base_lr	Base learning rate	0.001
momentum	Weight of the previous update	0.9
weight_decay	Regularization term	0.0005
lr_policy	Learning rate policy	“step”
gamma	Drop the learning rate	0.1
max_iter	Iterations total	100000
batch_size	Number of images at each iteration	256

X_i) のうち，最大所属確率 $\hat{Y}_i = \arg\max_{Y_i} P(Y_i \mid X_i)$ をとるラベル $\hat{Y}_i$ を推定ラベルとする．テストデータ画像 X_i の正解ラベル Y_i と，推定ラベルの $\hat{Y}_i$ の組合せ $(Y_i, \hat{Y}_i)$ ごとに枚数を集計した結果を**図 3** に示す．左が CNN のみ，右が CNN+FT の混同行列である．対角成分の値が大きいほど分類精度が高い．CNN+FT は，美的評価をより正確に推論できていることが確認できた．また，CNN に比較して CNN+FT は特に (High, High) および (Middle, Middle) を大きく改善した．

図 3 をもとに分類精度を算出した結果を**表 4** に示す．Accuracy は CNN の場合 57.7%だったが，CNN+FT では 70.0%まで改善した．Recall の Low を除き，CNN+FT によって各指標は大幅に改善している．本節では FT によって精度が改善し，Accuracy 70.0%で 3 値分類可能なことを確認した．

CNN+FT で分類された画像の例を**図 4** 示す．推定ラベルが High である写真は，色合いが鮮やかで，コントラストが強く，被写体の内容がはっきりしているもの，構図が練られている写真が多い印象を受けた．推定ラベルが Low である写真はブレやノイズを含み，被写体の内容が不鮮明である写真が多い印象を受けた．図 4 左下に位置する，正解ラベルが Low であって，推定ラベルが High である写真は，リンゴ，猿，花，猫，3 人のポートレートのように，比較的，被写体が明確であって，背景がシンプルである印象を受けた．逆に図 4 右上に位置する，正解ラベルが High であって，推定ラベルが Low である写真は，鍵と鎖，氷の上のアイスキャンディー，植物と鳥，割れたクルミと工具，イルミネーションのように，主となる被写体とその他の背景との区別が明確でなく，背景への物体の写り込みが多い印象を受けた．

学習によって作成された第 1 畳み込み層のフィルタ構造を**図 5** に示す．左が CNN のみの場合，右が CNN+FT の場合である．CNN のみの場合は最上段の左から 2 番目および 3 番目のようにノイズのような構造的機能を解釈しにくいフィルタが多く構成されている．一方，CNN+FT の場合は上段のように輝度成分のエッジを検出するフィルタ

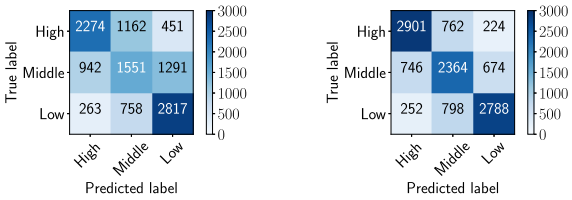

図 3　混同行列．（左）CNN，（右）CNN+FT

Fig. 3　Confusion matrix. (Left) CNN, (Right) CNN+FT.

表 4　精度評価

Table 4　Evaluation by accuracy/precision/recall.

		CNN	CNN+FT
Accuracy		57.7%	**70.0%**
Precision	High	65.4%	**74.4%**
	Middle	44.7%	**60.2%**
	Low	61.8%	**75.6%**
Recall	High	58.5%	**74.6%**
	Middle	41.0%	**62.5%**
	Low	**73.4%**	72.6%

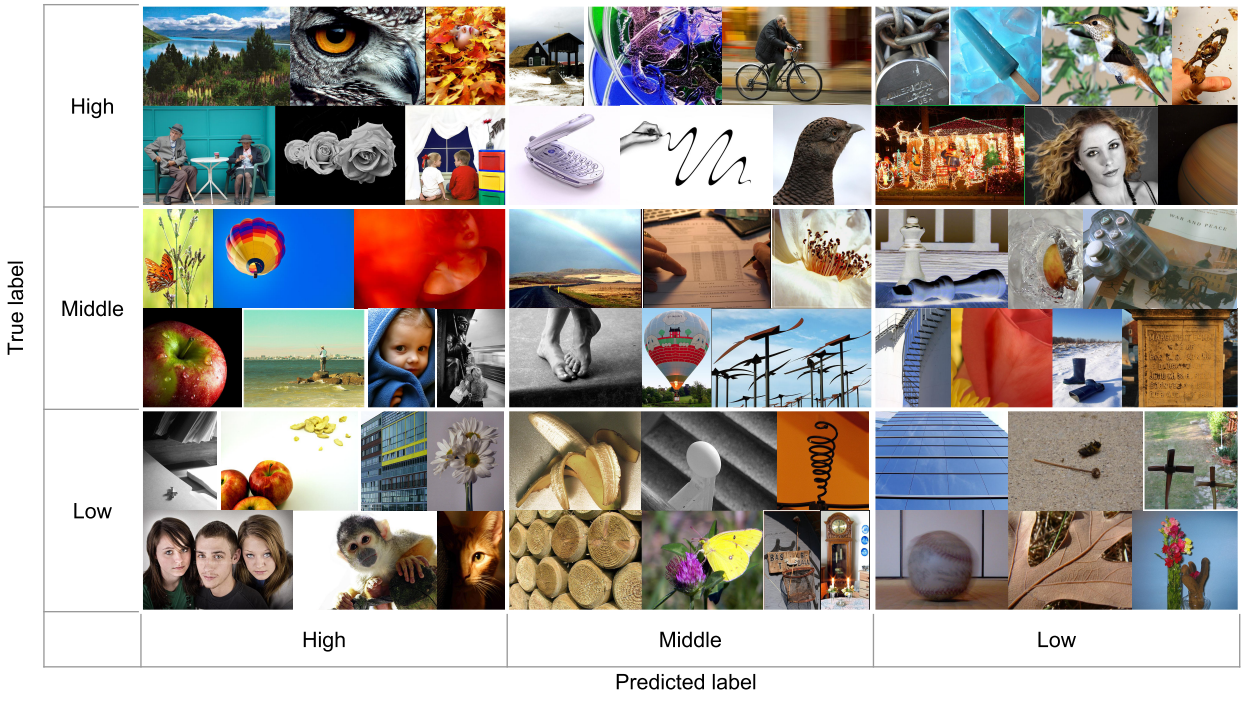

図 4　分類された画像の一例

Fig. 4　Examples of classified images.

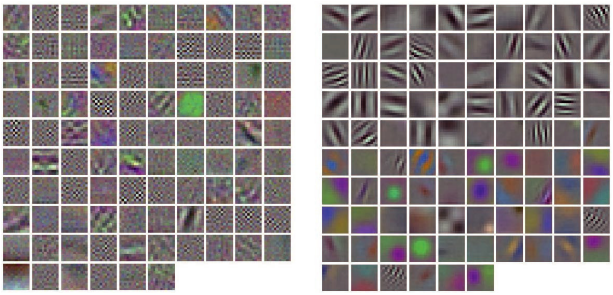

図 5　第 1 畳み込み層のフィルタ構造．（左）CNN，（右）CNN+FT
Fig. 5　Filter structures in 1st covolution layer. (Left) CNN, (Right) CNN+FT.

や，下段のように色相成分のグラデーションを検出フィルタが構成されている．以上の可視化結果から，CNN+FT は各フィルタが有効に機能していると考えられる．

6.2　リアルタイム性の評価

本節では構図推薦処理のリアルタイム性を確認する．5.4 節で述べたように 9 分割された画像どうしが重畳する幅を $d = 90$ として設定した．QX1 が取得するライブビューの解像度は高さ 424，幅 640 であるが，$d = 90$ と設定することで，まず 407 四方に縮小される．次に画像どうしが $d = 90$ で重畳するように 9 分割して，227 四方の画像 X_j が 9 枚得られる．また，中央画像 X_4 での撮影を指示する閾値を $th_{OK} = 0.9$ とした．

まず屋内環境でカメラを手に持ち，1 分間歩いてフレームレートの変動を測定した．このとき時々刻々と入力される画像は変化する．計測結果を**図 6** の凡例 indoor（high power）として示す．図 6 上図のようにプロセス並列化により，表示系プロセスはつねに約 25.0 fps で処理され，推論系プロセスの負荷によって 10.0 fps を下回るような表示ラグが発生することはなかった．図 6 下図のように，推論系プロセスは約 4.9 fps（ライブビュー画像約 5 フレームごとに推論）で処理可能であったが，被写体の状況が 1 秒間以内に大きく変化することはないと想定して，実装上は約 2.5 fps（10 フレームごと）に抑えて動作させた．以降の実験は約 2.5 fps に抑えた実装条件で実施する．

次に，屋内環境，屋外環境で同様の実験を行い，環境影響を比較した．計測結果を図 6 の凡例 indoor および outdoor として示す．図 6 上下図に示すとおり，屋内および屋外の環境条件にかかわらず，また時間の経過（入力される画像の変化）に依存せず，表示系プロセスは約 25.0 fps で，推論系プロセスは約 2.5 fps でフレームレートは安定することが確認された．図 6 各グラフの平均と標準偏差を**表 5** に示す．以上の処理速度からリアルタイム性の要求条件を満たすことを確認した．

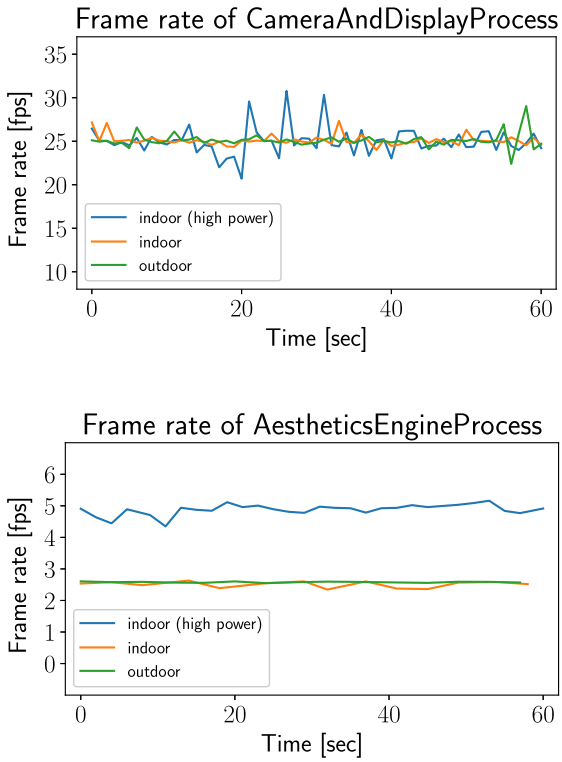

図 6　フレームレート．（上）表示系プロセス，（下）推論系プロセス
Fig. 6　Frame rate. (Upper) CameraAndDisplayProcess, (Lower) AestheticsEngineProcess.

表 5　フレームレートの平均と標準偏差
Table 5　Mean and standard deviation of frame rates.

Process	Environment	Mean	SD
CameraAndDisplay	Indoor (high power)	25.01	1.62
	Indoor	25.10	0.59
	Outdoor	25.09	0.77
AestheticsEngine	Indoor (high power)	4.88	0.17
	Indoor	2.51	0.10
	Outdoor	2.58	0.02

6.3　構図推薦の動作例

本節では構図推薦の挙動を確認する．推薦表示の例として，特に期待された効果を示した例を**図 7** に，期待以上の気づきが得られた例を**図 8** に示す．いずれの図も構図推薦カメラのライブビュー画像である．5 章で述べたように，中央の緑色の矩形内部が画像 X_4，すなわち撮影者が意識する撮影画角を示している．X_4 に重畳するように赤色の三角形で構図推薦方向を示し，「OK」という赤色のテキストで撮影タイミングを指示している．また，$\mathrm{TotalScore}_4$ が 0.75 以下の場合は「GOOD」，0.25 未満の場合は「BAD」と緑色で表示している．

図 7 は期待された効果を示したポートレート撮影の例である．撮影者から見て左に男性が，右に女性が立っている．図 7 の左図の状況では，男性の顔の上半分と女性の頭部が画像 X_4 の外側に見切れていた．すると，右上への移動を指示する構図推薦結果が表示された．この推薦結果に

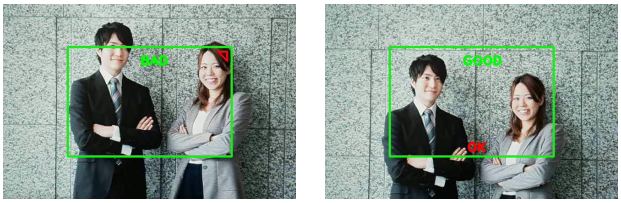

図 7　ポートレート撮影での動作例．（左）構図推薦，（右）撮影指示
Fig. 7　Demonstration for portrait. (Left) Recommend composition, (Right) Photo opportunity.

図 8　風景撮影での動作例．（左）構図推薦，（右）撮影指示
Fig. 8　Demonstration for landscape. (Left) Recommend composition, (Right) Photo opportunity.

従い，撮影者が右上方向へ構図を変更したのが図 7 の右図の状況である．男性と女性の顔全体が X_4 に収まった瞬間に撮影指示が表示された．

図 7 のように，被写体が見切れている状態というのは，初心者に見られる単純な構図のミスであるが，構図推薦カメラはこれを見逃さずに指摘していることが確認できた．教師あり学習では本来，良い構図と悪い構図が含まれていれば，それぞれを分類することができるが，本件では必ずしも悪い構図が含まれてはいない．一方，提案手法では画像 X_4 との相対評価による推薦を行っていることで，結果的に良い構図を推薦できている．

図 8 は期待以上の気づきが得られた風景撮影の例である．奥に森林，手前に湖面がある．湖面に木々が反射している．図 8 の左図の状況では，X_4 の下端近くに湖面と森林の境界線が位置していた．構図推薦は左下，下，右下への移動を示した．撮影者が下方向へ構図を調整したのが図 8 の右図の状況である．湖面と森林の境界線が X_4 の下端から 1/3 に位置した瞬間に撮影指示が表示された．

図 8 の右図のように撮影画角（緑枠）内を縦横 3 等分した線に風景の水平線や地平線を合わせたり，あるいは縦横の線の交点に被写体を配置する構図は 3 分割構図 [5], [6], [7] と呼ばれる．AVA データセットにもともと含まれており，かつ提案手法が High と推定した結果の中から，撮影者自らが 3 分割構図に従って撮影したと推察される写真を筆者が手動で抽出して**図 9** に例示する．3 分割構図は熟練者が意識するような複雑な構図であるが，構図推薦カメラの推薦結果はこの構図に従う傾向があった．本研究は明示的に 3 分割構図を学習したわけではないが，ラベル High に含まれる画像の多くに 3 分割構図が出現していたため，3 分割構図を推薦したと考えられる．

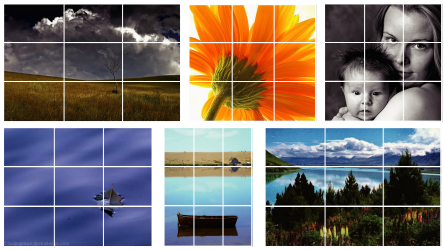

図 9　3 分割構図の例
Fig. 9　Rule of thirds.

表 6　満足度評価結果
Table 6　Evaluation results of satisfaction.

Answer	#answers
Strongly agree	1
Agree	6
Neither agree nor disagree	0
Disagree	1
Strongly disagree	0

6.4　構図推薦機能の満足度評価

構図推薦機能に対するユーザの満足度を評価する．被験者 A 群 8 人に 2 種類の被写体を撮影してもらう．2 種類の被写体は自然風景写真（図 8）および人物ポートレート写真（図 7）とした．いずれの被写体も 50 インチのディスプレイに 1 枚ずつ表示し，被験者はその中の部分領域を撮影する．被験者は構図推薦機能が ON および OFF を切り替えて撮影する．撮影の順番については次のとおりとする．記号の意味は構図推薦機能の ON または OFF，および被写体の種類を示す．

- 被験者 1：ON 風景，ON 人物，OFF 風景，OFF 人物．
- 被験者 2：OFF 風景，OFF 人物，ON 風景，ON 人物．
- …
- 被験者 8：OFF 風景，OFF 人物，ON 風景，ON 人物．

たとえば，被験者 1 は，構図推薦機能の ON に設定して，両被写体を順番に撮影する．その後，構図推薦機能を OFF に設定して，再び両被写体を順番に撮影する．被験者番号が奇数の場合は，被験者 1 と同じ順序で撮影する．被験者番号が偶数の場合は，構図推薦機能を OFF に設定してから実験を開始し，その後 ON に設定する．手法の試行順が結果に影響する順序効果が発生する恐れがあるため，その対策として，被験者ごとに ON および OFF の撮影順序を入れ替えた．

被験者 A 群には実験終了後に「美的評価が高い推薦ができていましたか？」という質問に対して，5 段階の主観満足度（非常にそう思う，そう思う，どちらでもない，そう思わない，非常にそう思わない）で回答してもらった．**表 6** のように 8 人中 1 人が「非常にそう思う」，6 人が「そう思う」，1 人が「そう思わない」と回答した．構図推薦機能の利用に対する高い満足度が確認された．

表 7　美的品質評価結果
Table 7　Evaluation results of aesthetics.

Score	Answer	#answers ON	#answers OFF
5	Excellent	6	1
4	Good	12	5
3	Average	10	18
2	Below average	3	7
1	Poor	1	1
Mean score		3.59	2.94

表 8　金額的評価結果
Table 8　Evaluation results of monetary value.

Monthly fee (yen)	0	100	300	1,000	3,000
#answers	3	6	18	2	0

6.5　構図推薦による撮影結果の美的品質評価

被験者 A 群の撮影した写真について第三者が主観に基づいて美的評価を行う．A 群とは異なる被験者 B 群 16 人を用意し，A 群が撮影した写真に対して，5 段階の主観美的評価スコア（5：非常に良い，4：良い，3：どちらでもない，2：悪い，1：非常に悪い）で回答してもらった．A 群の被験者 1 名が撮影した 4 枚の写真は B 群の被験者 2 名に割り当てられ，B 群の被験者 1 名は 4 枚の写真を評価する．B 群の被験者は A 群の実験条件（提示された写真と構図推薦機能 ON および OFF の対応関係，および順序）について一切知らされていない．

評価結果を**表 7** に示す．構図推薦 ON は主観美的評価スコアの平均値 3.59，OFF は 2.94 と，構図推薦 ON の場合に高い評価が得られた．またカイ二乗検定による p 値は 0.035（< 0.050）となり，構図機能の ON と OFF には有意な差があることが確認された．

6.6　構図推薦機能の金額的評価

構図推薦機能に対する金額的評価を行う．A 群および B 群とも異なる被験者 C 群 29 人に構図推薦カメラを操作してもらい，アンケートを実施した．従来のデジタルカメラに構図推薦機能を付加した場合に，追加で支払いうる金額を質問とした．なお，デジタルカメラの価格帯は数万円から数十万円と幅広く，被験者の相場感覚が一定ではないことから，構図推薦機能の利用に対して月額料金を要すると仮定し，選択肢（無料，100 円，300 円，1,000 円，3,000 円）の中から回答してもらった．評価結果を**表 8** に示す．選択肢の中で月額 300 円が最も多くの回答者を得た．全 29 人中，無料と回答した 3 人を除く 26 人により，構図推薦機能に一定の価値があることが認められた．構図推薦機能を備えたデジタルカメラであれば価格帯が上昇しても購入検討対象となりうることが確認できた．

また，2.3 節で説明したアプリケーションのうち，撮影時にデバイス上で構図をサポートするアプリケーションは構図カメラ，ガイドカメラ，PoseCam である．ガイドカメラが買い切りで 100 円，それ以外は無料であるが，本研究が提案する構図推薦カメラはより高い金額的価値を持つと評価された．

7.　おわりに

本研究では，写真の美的評価を 3 段階で分類する美的評価エンジンを構築した．美的評価エンジンを Jetson TX1 に実装して構図推薦カメラを開発した．また，美的評価の高い写真が得られる方向へ構図を誘導し，撮影を指示する処理がリアルタイムに動作することを確認した．この構図推薦は初心者に見られるようなミスを指摘したり，熟練者が意識するような構図を推薦して新たな気づきや構図を追求する機会を与えた．ユーザ評価によって構図推薦機能への高い満足度，撮影結果の美的評価の向上，製品の金額的価値の上昇を確認した．

さらに撮影技術の習得効率を向上させるためには，美的評価に対する詳細な理由づけが必要と考えられる．今後の発展としては評価値の理由，および画像の部分領域ごとの改善点を明らかにすることを可能にしていきたい．また，露光やピント，色彩，コントラストが美的評価に与える影響も明らかにしていきたい．

参考文献

[1] Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q., Wang, J.Z., Li, J. and Luo, J.: Aesthetics and Emotions in Images, *IEEE Signal Processing Magazine*, Vol.28, No.5, pp.94–115 (online), DOI: 10.1109/MSP.2011.941851 (2011).

[2] Murray, N., Marchesotti, L. and Perronnin, F.: AVA: A large-scale database for aesthetic visual analysis, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp.2408–2415 (online), DOI: 10.1109/CVPR.2012.6247954 (2012).

[3] Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J.Z.: RAPID: Rating Pictorial Aesthetics Using Deep Learning, *Proc. 22nd ACM International Conference on Multimedia, MM '14*, pp.457–466, ACM (online), DOI: 10.1145/2647868.2654927 (2014).

[4] Kao, Y., He, R. and Huang, K.: Deep Aesthetic Quality Assessment With Semantic Information, *IEEE Trans. Image Processing*, Vol.26, No.3, pp.1482–1495 (online), DOI: 10.1109/TIP.2017.2651399 (2017).

[5] ブライアン・ピーターソン：ナショナルジオグラフィック プロの撮り方 構図を極める，ナショナル・ジオグラフィック，日経ナショナルジオグラフィック社 (2013).

[6] 内池秀人，福井麻衣子：写真構図のルールブック，マイナビ (2012).

[7] 山田芳文：写真は「構図」でよくなる! すぐに上達する厳選のテクニック 23，エムディエヌコーポレーション (2018).

[8] 家田　暁, 琴　智秀, 萩原将文：感性を反映した構図修正による写真品質向上システム, 芸術科学会論文誌, Vol.9, No.4, pp.163–172（オンライン），DOI: 10.3756/artsci.9.163 (2010).

[9] 志津野之也，濱川　礼：構図マッチング手法を用いた写真撮影時の自動構図決定手法，マルチメディア，分散協調とモバイルシンポジウム 2014 論文集，Vol.2014, pp.646–656 (2014).
[10] Bhattacharya, S., Sukthankar, R. and Shah, M.: A Framework for Photo-quality Assessment and Enhancement Based on Visual Aesthetics, *Proc. 18th ACM International Conference on Multimedia, MM '10*, pp.271–280, ACM (online), DOI: 10.1145/1873951.1873990 (2010).
[11] Krizhevsky, A., Sutskever, I. and Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25*, Pereira, F., Burges, C.J.C., Bottou, L. and Weinberger, K.Q. (Eds.), Curran Associates, Inc., pp.1097–1105 (2012) (online), available from ⟨http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf⟩.
[12] 内田祐介，山下隆義：[サーベイ論文] 畳み込みニューラルネットワークの研究動向，パターン認識・メディア理解研究会 (2017).
[13] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol.abs/1409.1556 (2014) (online), available from ⟨http://arxiv.org/abs/1409.1556⟩.
[14] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going Deeper with Convolutions, *CoRR*, Vol.abs/1409.4842 (2014) (online), available from ⟨http://arxiv.org/abs/1409.4842⟩.
[15] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CoRR*, Vol.abs/1512.03385 (2015) (online), available from ⟨http://arxiv.org/abs/1512.03385⟩.

付　　録

A.1　構図の探索範囲

構図推薦は，一定のサイズの 9 枚の画像から最適な構図を探索している．より広い領域，狭い領域，あるいは回転させた領域，等の探索には現状では対応していないが，探索する画像を増やすことで可能になる．また，レンズの絞り，ピント，ホワイトバランス，露出等についても現状は対応できていないが，いずれも画像の内容を大きく変える要素であり，撮影前に設定値を変更することで事前に画像を想定できるため，探索対象に含めることは可能である．しかしながら，いずれの場合も探索範囲を広げると計算量は増加する．

井上 義隆　（正会員）

2011 年東京工業大学大学院理工学研究科修士課程修了．同年株式会社 NTT ドコモ入社．時空間行動予測，画像認識に関する研究開発に従事．

松村 択磨

1998 年仙台電波工業高等専門学校（現，仙台高等専門学校）情報通信工学科卒業．2002 年ドコモ・テクノロジ株式会社入社．携帯電話交換機の仕様検討・設計，新規サービスの検討・開発業務等に従事．

深澤 佑介　（正会員）

2002 年東京大学工学部卒業．2004 年東京大学大学院工学研究科修士課程修了．同年株式会社 NTT ドコモ入社．2011 年東京大学大学院工学研究科博士後期課程修了．同年 10 月東京大学人工物工学研究センターにて協力研究員，2017 年より客員研究員兼任．Web マイニング，レコメンデーション，実世界行動予測に関する研究開発に従事．IEEE，人工知能学会各会員．博士（工学）．

山田 和宏

1999 年慶應義塾大学大学院政策・メディア研究科修士課程修了．同年株式会社 NTT ドコモ入社．Java アプリダウンロードサービス「i アプリ」のサービス/技術企画，スマートフォン向けサービス基盤企画，新規事業の創出等に従事．