Vol.2018-DBS-167 No.25
Vol.2018-IFAT-132 No.25
2018/9/14

# 物体の色や表情情報を利用した画像の印象にあった音楽推薦手法の提案

追木 智明[1,a)]　欅 惇志[1,b)]　宮崎 純[1,c)]

**概要**：本研究では画像中の物体や色，顔の表情情報を含む画像キャプションを用いて画像と音楽を感情を表現する共通の空間へプロットし，画像の印象にあった音楽を推薦するシステムを提案する．従来の手法では画像に対して抱く感情等に関するユーザスタディの結果に基づいて音楽の推薦を行っていたが，本研究では，代替の手法として一般物体認識，画像キャプション生成，画像中の表情推定などの手法を用いる．本論文では一般物体認識による抽出物体のみを用いた手法と，一般物体認識に加え画像キャプションや物体の色情報や顔の表情情報を用いた手法により，画像中の情報を抽出し，画像中の情報を印象語へ変換．重み付けを行い感情を表現する空間上への印象語のプロットを行う．さらに，各手法によってプロットされた印象語に対してクラスタリングを行い，各クラスタの中心の最近傍点となる音楽を推薦対象とし，画像から音楽の推薦を行った．また，推薦結果中の不適合な音楽の削減を目的とした推薦対象の絞り込みを行った．主観評価実験を行った結果，推薦対象の絞り込みによって，画像の印象に合った音楽が推薦されたとユーザが評価した割合が上がったことから推薦対象の絞り込みの有効性を示した．

**キーワード**：情報推薦　音楽推薦　一般物体認識　Word2Vec　画像キャプション

## 1　はじめに

近年，ユーザの選択した画像に対して相応しい音楽を推薦するシステムが着目されつつある [5][6][7]．展覧会などにおける絵画や写真などの画像に対して適切な音楽を BGM として流すことで，利用者は画像に対してより深い印象を持つことができる [1] と報告されていることからも，画像に対して適切な音楽を特定することは重要である．なお，既存の画像と音楽の対応付けを行う研究ではユーザスタディによって画像や音楽を感情を表現する共通の空間上に写像し，マッピングする手法が主流である．

それに対して我々は，画像から共通空間への写像において，1．一般物体認識や画像キャプション生成，顔の表情推定などによる画像中の情報抽出，2．画像中の情報から印象語へ Word2Vec を用いて変換し，共通空間へプロット，印象語間の角度に基づくクラスタリング，3．各象限にプロットされた印象語数の多数決による推薦対象象限の絞り込み，という手順によって，ユーザースタディを行うことなく画像を音楽と共通の空間上へ写像，対応付け，推薦を行う手法を提案する．本論文では画像中の情報に一般物体認識による抽出物体のみを用いた手法と，一般物体認識に加え，色が人に与える感情的な影響 [20] や，顔の表情情報や物体の動作が印象を決める上で重要であるという仮説から，画像キャプションや物体の色情報，顔の表情情報などを用いた手法の２つを提案し，評価を行った．

[1] 東京工業大学情報理工学院情報工学系 〒 1528552 東京都目黒区大岡山 2-12-1
[a)] tuiboku@lsc.cs.titech.ac.jp
[b)] keyaki@lsc.cs.titech.ac.jp
[c)] miyazaki@cs.titech.ac.jp

## 2　準備

### 2.1　物体認識

物体認識の分野では一般物体認識と特定物体認識の二つの技術に分けて研究が行われている．一般物体認識では画像上に存在する犬や車，人間といった物体を検知し分類する技術であるのに対して，特定物体認識は顔認証システムなどのように人間の中でも特定の人物などに特化して認識する技術のことである．従来，一般物体認識は物体領域候補の抽出，物体領域候補の物体認識，検出領域の絞り込みの 3 ステップに分かれて行われていたが，機械学習手法の発展とともに深層学習を用いた手法が主となった．深層学習を用いた手法では主に Convolutional Neural Network(CNN)[13] を用い，画像上の物体の領域 (バウンディングボックス) と物体名によってラベル付けしたデータセットを用いて学習させ，認識モデルを作成するのが主流となっている．

### 2.2　画像キャプション

ある画像に対して、画像の説明をするような文章を画像キャプションという．画像キャプションを生成する多くの手法 [16][17][18] では CNN と Recurrent Neural Network(RNN)[15] を組み合わせ，画像処理と文章生成を行うモデルを作成する．画像と人手で付加された画像の説明文が含まれるデータセットを用いてモデルを学習するため，付加された画像の説明文の性質によって出力されるキャプションの性質も大きく変わる．

Vol.2018-DBS-167 No.25
Vol.2018-IFAT-132 No.25
2018/9/14

### 2.3 Word2Vec

Word2Vec とは Mikolov ら [2] によって提案された，ニューラルネットワーク構造のモデルによって単語や句を (語彙数と比較して) 低次元のベクトル (分散表現) として表現する技術である．Word2Vec は似通った文脈 (コンテキスト) にて出現する単語 · 句同士は似通った意味や属性 · 性質を持つ可能性が高いという分布意味論 [3] という仮説に基づいたものである．語間の類似度算出や，アナロジータスクにおいて多用される．

### 2.4 感情に関する空間

Russell ら [4] は Arousal と Valence の 2 次元によって感情を表す Arousal-Valence(AV) 空間を提案した．Arousal と Valence はそれぞれ [-1，1] の実数を取る．Arousal は-1 に近づくにつれて落ち着いた (calm) 感情を表し，+1 に近づくにつれて興奮した (excite) 感情を表す．Valence は-1 に近づくにつれて negative な感情を表し，+1 に近づくにつれて positive な感情を表す．図 1 に印象語を AV 空間上にプロットした例を示す．AV 空間上にプロットされる各印象語の座標はユーザスタディの結果によるものであり，原点からの角度と長さによって座標が決まる．

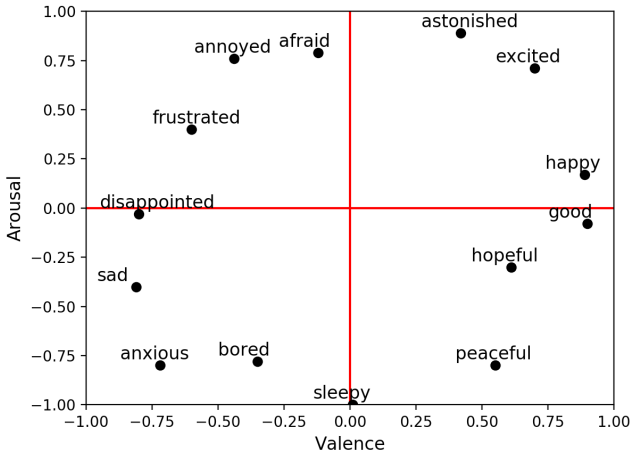


図 1 AV 空間上の印象語

### 2.5 角度によるクラスタリング

Dhillion らは非階層型クラスタリング手法の一つである k-means 法を超球面上のデータを用いることが出来るように拡張した手法である “spherical-k-means”[11] を提案した．この手法では k-means 法でクラスタの更新時に各クラスタの中心とデータのユークリッド距離を用いる代わりに角度を用いてクラスタの更新を行う．

## 3 関連研究

### 3.1 画像からの音楽推薦に関する既存研究

#### 3.1.1 風景特徴ベクトルによる推薦

糸井ら [5] はあらかじめ 6 種類の風景 (lakeside，mountain 等) を定義しておき，分類される風景画像と音楽のペアを提示しどの風景が一番音楽に近いと感じるかをクラウドソーシングによって調査した．その結果を用い，風景特徴ベクトルと呼ばれるベクトルをそれぞれの音楽に対して作成することで，入力画像に近い音楽の推薦を行った．

#### 3.1.2 多変量解析による推薦

新穂ら [6] は画像の色特徴量や画素の並びによる相関を考慮することで画像の特徴量をベクトル化，音楽からは音高を特徴量とし，音高の変化をパターン化することで音楽の特徴量をベクトル化している．これらのベクトル同士にどの程度の相関があるかを考慮する際に，画像と音楽それぞれに近いと感じる印象語をアンケートを行い画像と音楽への写像を構築することで推薦を行っている．

#### 3.1.3 画像特徴量と音楽特徴量による推薦

佐々木ら [7] は画像特徴量と音楽特徴量を用いて AV 空間上に画像と音楽をプロットすることで推薦を行っている．画像は色特徴量と形状特徴量を用いて，それぞれの特徴量に関して既存の式に基づいて AV 空間にプロットしている．音楽では主成分分析により特徴量を抽出しており．29 種類の音響特徴量を主成分分析することで得られた第一，第二主成分がそれぞれ Arousal と Valence に深く関連していることを用いて特徴量から AV 空間にプロットしている．

## 4 提案手法

3 節で紹介したように関連研究の多くの手法 [5][6] はある特徴量に対してユーザスタディの結果を用いて音楽とのマッチングを行うというものであった．また，佐々木ら [7] は画像から抽出可能である色特徴量や形状特徴量から AV 値への変換を行うことで推薦を行った．

本研究では，提案手法として画像を AV 空間にプロットする際に一般物体認識や画像キャプション生成，画像中の表情推定などの手法を用いることでユーザスタディによる人的コストを費やすことなく画像と音楽を感情を表現する共通の空間上に写像することが可能になる．

提案手法では，図 2 に示す通り，下記の三つの手順を踏まえて画像を入力とした音楽推薦を行う．まず，1. 一般物体認識や画像キャプション生成，顔の表情情報等による画像中の情報抽出 (図 2(1))，2. 画像中の情報を印象語へ変換 (図 2(2))，3. AV 空間上へ印象語のプロット，推薦対象象限の絞り込み，音楽の推薦 (図 2(3),(4))，である．また，音楽は AV 空間上における座標が付与されている音楽データセットを用いて AV 空間上に音楽をプロットする[*1]．本論文では画像中の情報抽出に一般物体認識のみを用いた手法 (一般物体認識手法) と，更に画像キャプションや顔の表情情報，物体の色情報を用いることで画像中のより詳細な情報を用いた手法 (キャプション手法) を提案する．以降で各手法の詳細を述べる．

### 4.1 画像中の情報の抽出

ステップ (1) で一般物体認識手法とキャプション手法を用いた画像中の情報抽出を行う．

[*1] 本音楽データセットでは長さが 45 秒の音楽約 2000 曲に対してユーザスタディを行い 0.5 秒毎に AV 値を調査することによって AV 空間上での座標が取得されている．

Vol.2018-DBS-167 No.25
Vol.2018-IFAT-132 No.25
2018/9/14

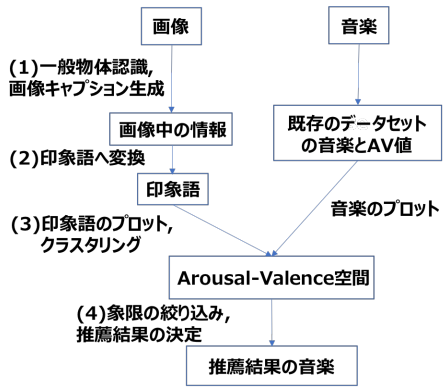


図 2　システムの概要図

#### 4.1.1　一般物体認識による抽出物体を用いた手法

一般物体認識手法は以下のステップ (a)，(b) で画像中の情報を抽出する．

(a)一般物体認識による画像上の物体抽出
(b)物体の面積による各物体の重み付け

ステップ (a) で一般物体認識によって画像上の物体の抽出を行う．その後，ステップ (b) で，画像中の物体の面積は画像における物体の印象への影響力の大きさと相関を持つという仮定のもとに物体の重み付けを行う．以降で各ステップの詳細を述べる．

**(a) 画像上の物体の抽出**

ステップ (a) では画像に対して一般物体認識を行うことで，画像上の物体の抽出を行う．図 3(a) では，入力画像に対して一般物体認識を行うことで画像中の物体の種類と画像上での面積を抽出している．一般物体認識には Liu らによって提案された深層学習を用いた手法の一つである SSD と呼ばれる手法を用いている．SSD では初期段階で画像をグリッド分割し，教師データと比較することで高速かつ高精度なバウンディングボックスの調整が可能な手法である．

**(b) 物体の面積による各物体の重み付け**

一般物体認識の結果に対して，画像上で面積の大きい物体が印象を決める上で重要であると仮定し各物体に重み付けを行う．図 3 の例では一般物体認識の結果人間と本，ベンチを検知し面積を計算している (図 3(b))．さらに同じ種類の物体同士の面積を足し合わせた値を総面積とし総面積の最大値と各物体の総面積を用いて以下の式によって各物体の重みを計算する．

$$weight(X) = \frac{d(X)}{dmax}$$

(dmax:物体ごとの総面積の最大値，d(X):X の総面積)

#### 4.1.2　画像キャプションと顔の表情情報を用いる手法

キャプション手法は以下のステップ (a)，(b) で画像中の情報を抽出する．

(a)画像キャプションの生成と表情情報の抽出
(b)キャプションに対する色情報と表情情報の付加
(b-1)Mask R-CNN による物体のセグメンテーション
(b-2)ドミナントカラーの計算と英単語への変換
(b-3)キャプションに対する色情報と表情情報の付加

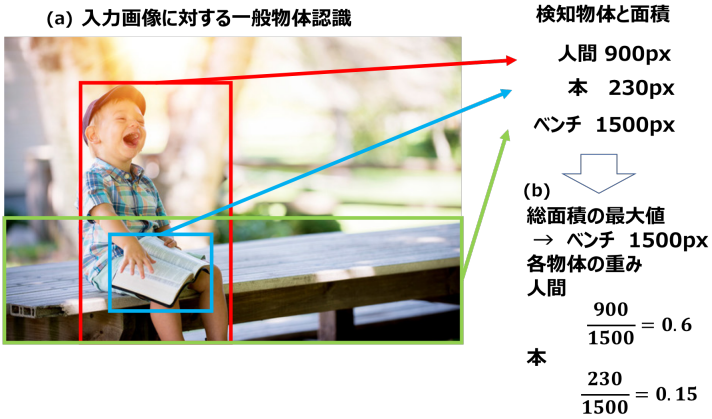


図 3　各物体の重みの計算

ステップ (a) で入力画像に対する画像キャプションの生成と表情情報の抽出を行う．次に，ステップ (b) でキャプションへの表情情報の付加と，入力画像に対する一般物体認識により抽出された物体の印象を支配する色 (ドミナントカラー) を計算しキャプションに色情報を付加する．以降で各ステップの詳細を述べる．

**(a) 画像キャプションの生成と表情情報の抽出**

画像キャプションの生成と画像中の人の表情情報の抽出に Microsoft 社の Computer Vision API[*2] と Face API[*3] を用いる．図 4 に画像キャプション生成と表情推定の例を示す．表情情報は 8 つの表情クラスに対してそれぞれ確信度が出力される．本手法では最も確信度が高いクラスの表情を採用した．


図 4　画像キャプションの生成と表情推定の例

**(b) キャプションに対する色情報と表情情報の付加**

ステップ (b-1)〜(b-3) で画像中の物体の色情報と顔の表情情報をステップ (a) で取得したキャプションに対して付加する．以下に各ステップの詳細について示す．

**(b-1)Mask R-CNN による物体のセグメンテーション**

入力画像に対して一般物体認識の一手法である Mask R-CNN[19] を用いる．Mask R-CNN の特徴として，物体の領域を矩形のみでなくピクセル単位で分類することでより物体の形状に沿ったセグメンテーションが可能であることが挙げられる (図 5(b-1))．

**(b-2) ドミナントカラーの計算と英単語への変換**

[*2] https://azure.microsoft.com/ja-jp/services/cognitive-services/computer-vision/
[*3] https://azure.microsoft.com/ja-jp/services/cognitive-services/face/

Vol.2018-DBS-167 No.25
Vol.2018-IFAT-132 No.25
2018/9/14

ステップ (b-1) で抽出した各物体の領域情報からそれぞれの物体に対してドミナントカラーを計算する．領域の各ピクセルの RGB 値を K-means 法によって 3 つのクラスタにクラスタリングし，各クラスタの中心の座標をドミナントカラーの RGB 値とする．次に，英単語と RGB 値のペアの中で最もドミナントカラーに近い色を計算し，RGB 値から英単語への変換を行い，物体中の割合が最も大きい英単語をキャプションへ付加する (図 5(b-2))．色間の距離の計算には人の知覚を考慮した色差式である CIEDE2000[14] を用いた．

**(b-3) キャプションに対する色情報と表情上情報の付加**

人に関する英単語を辞書として用意し，キャプション中に辞書に含まれる単語があれば直前に表情情報を付加する．次に，Mask R-CNN による抽出物体とキャプション中の各単語の類似度を Word2Vec を用いて計算し，類似度が 0.5 以上の単語の中で最も類似度が高い単語の前にステップ (b-2) で変換した色を表す英単語を付加する (図 5(b-3))．


図 5　キャプションに対する色情報と表情情報の付加

## 4.2　画像中の情報から印象語への変換

ステップ (2) では各手法から抽出された画像中の情報と関連する印象を推定するために Word2Vec を用い，画像中の情報を印象語へ変換する．

一般物体認識手法では，Word2Vec を用いて各物体と印象語の類似度を計算，ソートし，上位 5 個を各物体に対して近いと考えられる印象語とする．図 6 の例では，図 3 の例の抽出結果である人間，本，ベンチに対して Word2Vec を用いて AV 値が既知である印象語[*4]との類似度を計算し，類似度順でソートしている．

キャプション手法では，キャプション中の内容語 (名詞，動詞，形容詞，副詞) を抽出し，関連する印象語へ変換する．キャプション中の各単語と印象語[*4]との類似度を Word2Vec によって計算，ソートし，上位 5 個の印象語を単語と関係のある印象語とする．更に最も類似度が高い印象語の重みを 1 として正規化し，各印象語の重み付けを行う．

Word2Vec では，Google News を学習用コーパスとした学習済みモデル[*5]を使用した．

[*4] 印象語は Georgios ら [10] の論文で AV 値が公開されているものを使用．
[*5] https://code.google.com/archive/p/word2vec/

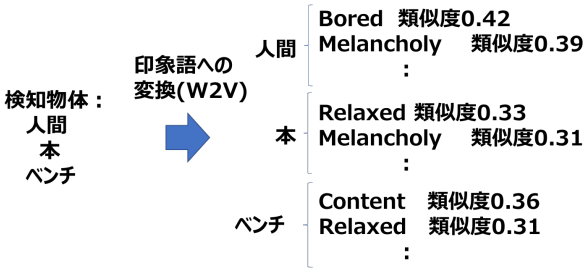


図 6　画像中の情報から印象語への変換

## 4.3　印象語のプロットとクラスタリング

ステップ (3) は以下の手順で行われる．
(3-1)重みを考慮した印象語のプロット
(3-2)印象語のクラスタリング
ステップ (3-1) で物体やキャプション中の単語から変換された印象語の座標 (x,y) に対して各印象語の重み weight(X) を掛け，印象語の原点からの長さを調整することで，重みを考慮した印象語の座標 (x*weight(X),y*weight(Y)) を用いて AV 空間上に印象語をプロットする (図 7)．さらに，ステップ (3-2) で AV 空間にプロットされた印象語の座標が原点からの角度と長さによって決まるという性質から各印象語間の角度によるクラスタリング手法である “spherical-k-means” [11] を用い，クラスタリングを行った．AV 空間では各象限で異なる 4 つのクラスタに分類し，クラスタの中心から最も近い音楽を推薦対象とする．今回，クラスタ数を 4 としているが，今後クラスタ数に関しても考える必要がある．

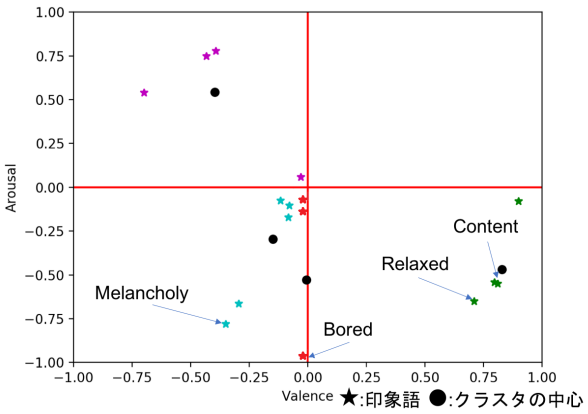


図 7　印象語のプロット

## 4.4　推薦対象象限の絞り込みと音楽の推薦

ステップ (4) は以下の手順で行われる．
(4-1)各象限にプロットされた印象語の個数の集計
(4-2)集計結果に基づく推薦対象象限の絞り込み
(4-3)推薦結果音楽の決定
ステップ (4-1) では図 7 のような印象語のプロットとクラスタリングの結果に対して，各象限にプロットされた印象語数を集計する (図 8)．両手法を用いて画像 60 枚に対してプロットされた印象語を集計し，各象限にプロットされる平均印象語数を計算する．次にステップ (4-2) で，画像 1 枚からある象限にプロットされた印象語数を平均印象語数で割ることで，各象限にプロットされる印象語数の偏りを考慮した正規化を行い，正規化結果中の最大値となる象限を推薦対象象限とする (図 8)．

Vol.2018-DBS-167 No.25
Vol.2018-IFAT-132 No.25
2018/9/14

#### 4.4.1 推薦結果音楽の決定

ステップ (4-3) でステップ (4-2) によって選択された象限に含まれる推薦対象を入力画像に対する推薦結果とする．図 9 中の青枠は選択された象限を表し，その象限に含まれるクラスタの中心の最近傍点となる音楽を推薦結果とする．

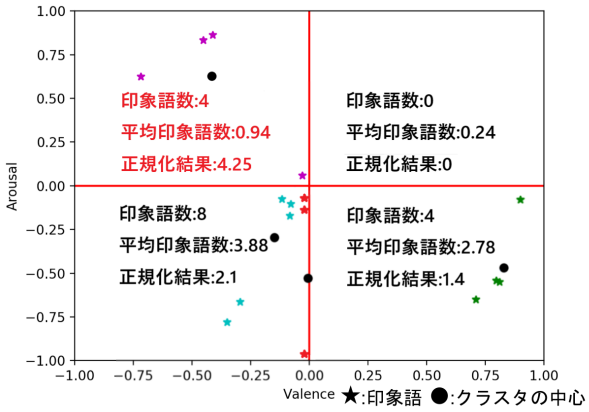


図 8 推薦対象象限の絞り込み

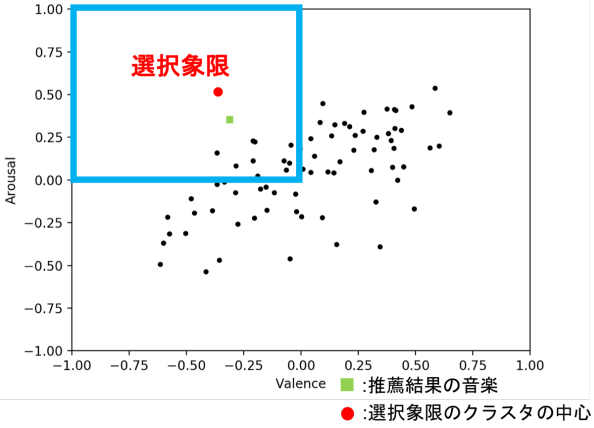


図 9 推薦象限の絞り込みと音楽の推薦結果

## 5 評価実験

提案手法による推薦精度を評価するために以下の二点について評価を行った．

(1)画像から印象語への変換に関する評価
(2)推薦結果に関する評価

これらの評価のため，主観評価実験としてクラウドソーシングサイト上の計 20 人に対して次の内容の質問を行った．(1) では各ユーザに対して画像を 10 枚提示し画像に対して抱いた印象に基づいて AV 空間上の一点を選択してもらった．(2) では (1) と同一の画像 10 枚に対して推薦された音楽と画像の印象がどの程度合っているかをアンケートによる 5 段階評価 (1:合っていない，5:合っている) を行った．また，AV 空間上の座標が付与された音楽データセットである “DEAM: MediaEval Database for Emotional Analysis in Music” [12] を用い，音楽を AV 空間上にプロットしている．このデータセットは 45 秒間の音楽の 15 秒〜45 秒の区間対してユーザスタディによって 0.5 秒毎に AV 値を調査したものであり，全 2000 曲からなる．今回は各区間の AV 値の平均をその音楽の AV 値として AV 空間にプロットした．

### 5.1 画像から印象語への変換に関する評価

実験結果より，一般物体認識手法とキャプション手法により選択された象限とユーザが選択した象限の一致率を計算した結果，表 1 のようになった．結果から，各手法の全体の正解率は低い値となった．理由としては，ユーザが第 2，3 象限を選択する確率が極めて低く，第 1，4 象限に偏っているのに対して，各手法では各象限にプロットされた印象語数による正規化を行うことで各象限の選択率に大きな偏りないことが影響していると考えられる．各象限の印象語のプロット数だけでなくユーザによる象限の選択の傾向を考慮することで精度の改善が見込めると考えられる．

### 5.2 推薦結果に関する評価

本節では下記の 3 手法について評価を行う．

- **推薦対象象限の絞り込みを行わない (絞り込みなし)**
  推薦対象象限の絞り込みを行わず各クラスタの中心から合計 4 曲の音楽を推薦する手法
- **印象語数による象限の決定 (絞り込みあり)**
  4.3 節の手順から決定された音楽のみを推薦する手法
- **ユーザが推薦対象の象限を選択 (ユーザ選択)**
  (1) の結果より，ユーザが選択した象限から音楽を推薦する手法．他の手法に対する参考手法とする．

象限の絞り込みの有無を比較することで推薦対象の絞り込みの有効性，象限絞り込みの有無とユーザ象限選択を比較することで提案手法による象限の選択の有効性を確認する．ここで各手法に関して各画像に対して推薦された音楽への各ユーザ評価値の平均 (以降，推薦スコアとする) が 4 以上である割合，各画像の推薦スコアの平均 (以降，全体平均とする)，推薦に失敗する割合 (推薦失敗割合) の 3 指標に関してランダム，一般物体認識手法，キャプション手法 (色/表情なし)，キャプション手法 (色/表情あり) の 4 手法について評価した結果を表 2 に示す．なお，推薦失敗とは選択した象限に対する推薦対象が存在せず推薦が行えないことを指す．

結果から，象限の絞り込みを行うことで推薦スコア 4 以上の割合が増加していることから，推薦された音楽には印象に合わないものが多く含まれ，象限の絞り込みによって不適切な推薦結果を削減できていると言える．また，全体平均のユーザ象限選択と象限絞り込みありを比較すると，参考値であるユーザ象限選択のほうが一般物体認識手法を除いて良い結果となっているが，これは各手法の象限絞り込みの精度の低さが大きく影響していると考えられる．推薦失敗割合に関しては，一般物体認識手法やキャプション手法を用いた象限の絞り込みでは発生せず，ユーザが象限を選択した場合にのみ発生するという結果になった．これらのことから，各手法の象限の絞り込みの精度を改善することで低い推薦失敗割合を維持しつつ，推薦スコア 4 以上の割合や全体平均の向上が見込めるのではないかと考えられる．また，キャプションに付加された表情情報のみを用いて象限の絞り込みを行った際，キャプション手法 (色/表情あり) の全体平均が 2.79 から 3.09 へ上昇した．よって，表情情報は画像を見たときの印象に対して大きく影響すると考えられ，表情情報を重視した場合に各指標の改善が見込めると考えられる．

**表 1** ユーザ選択象限と各種手法による選択象限の一致率

| | 一般物体認識手法 | | | | | キャプション手法 (色/表情なし) | | | | | キャプション手法 (色/表情あり) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 象限 | 1 | 2 | 3 | 4 | 全体 | 1 | 2 | 3 | 4 | 全体 | 1 | 2 | 3 | 4 | 全体 |
| ユーザによる象限の選択率 (%) | 56.0 | 6.0 | 4.0 | 34.0 | - | 56.0 | 6.0 | 4.0 | 34.0 | - | 56.0 | 6.0 | 4.0 | 34.0 | - |
| 各手法による象限の選択率 (%) | 30.0 | 30.0 | 30.0 | 10.0 | - | 10.0 | 40.0 | 30.0 | 20.0 | - | 10.0 | 40.0 | 30.0 | 20.0 | - |
| 一致率 (%) | 42.0 | 41.7 | 25.0 | 17.6 | 33.0 | 15.1 | 58.3 | 37.5 | 23.5 | 19.5 | 12.5 | 50.0 | 37.5 | 23.5 | 21.5 |

**表 2** 各種手法の比較

| | 推薦スコア 4 以上 | | | 推薦スコアの全体平均 | | | 推薦失敗割合 | |
|---|---|---|---|---|---|---|---|---|
| | 絞り込みなし | あり | ユーザ選択 | 絞り込みなし | あり | ユーザ選択 | 絞り込みあり | ユーザ選択 |
| ランダム | 1.5 % | - | - | 2.37 | - | - | - | - |
| 一般物体認識手法 | 0.5 % | 19.5 % | 12 % | 2.37 | 2.48 | 1.87 | 0 % | 12.5 % |
| キャプション手法 (色/表情なし) | 2 % | 30.5 % | 46.5 % | 2.56 | 2.59 | 3.14 | 0 % | 3 % |
| キャプション手法 (色/表情あり) | 0 % | 34.5 % | 40.5 % | 2.50 | 2.79 | 3.01 | 0 % | 1 % |

## 6 まとめ

本論文では，一般物体認識による抽出物体や，画像中の顔の表情情報や物体の色情報を付加した画像キャプションを用い，画像と音楽を AV 空間上へ写像し，対応付けを行うことで画像の印象にあった音楽の推薦システムを提案した．

提案手法の評価として，画像から印象語への変換による AV 空間への画像のプロットとプロットされた画像による音楽の推薦結果の二点について評価実験を行った．画像から印象語への変換に関する評価として，ユーザが実際に画像を AV 空間上にプロットし，プロットした象限と提案手法により選択された象限がどの程度一致するかを評価した．結果から，ユーザの象限選択は第 1，4 象限に偏る傾向があり，現在の手法では正しくユーザのプロットを推定することができないことが判明した．推薦結果に関する評価として，推薦された音楽に対してユーザが 5 段階評価を行った結果に対し，象限絞り込みなし，象限絞り込みあり，ユーザ選択手法の 3 つの手法に分けて評価した．その結果，提案手法やユーザによる象限の選択によって推薦対象を絞り込んだ場合に推薦スコアが 4 以上になる割合が増えたことから象限の決定の有効性を示した．しかし，提案手法とユーザが選択した象限の一致率はかなり低い結果となり，今後，一致率を改善することで全体の評価値の改善が見込めると考えられる．

今後の課題として，ユーザによる象限の選択の傾向を提案手法による象限の選択と組み合わせることや，表情情報を重要視するような象限の選択を行うことで提案手法とユーザによる象限の選択の一致率や推薦スコア 4 以上の割合，全体平均といった各指標の改善が必要である．また，深層学習を用いた文章の感情推定を応用し，画像キャプションから AV 値を直接推定する手法も検討していく．

## 7 謝辞

研究に関する助言を頂いた，お茶の水女子大学理学部教授伊藤先生に感謝いたします．本研究の一部は，JSPS 科研費 (16H02908，18H03242，18H03342，15K20990，17K12684)，JST ACT-I の助成を受けたものである．ここに記して謝意を表す．

## 参考文献

[1] 岩宮眞一郎． オーディオ・ヴィジュアル・メディアを通しての情報伝達における視覚と聴覚の相互作用に及ぼす音と映像の調和の影響． 日本音響学会誌， 1992， 48．9: 649-657.

[2] MIKOLOV, Tomas, et al. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111-3119.

[3] M. Baroni et al., “DistributionalMemory: A General Framework for Corpus-Based Semantics”, Journal Computational Linguistics, Vol. 36, No. 4, pp. 673721, 2010.

[4] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology，39(6)，1161-1178.

[5] 糸井勇貴，奥健太，山西良典，楽曲の風景特徴化に基づく風景アウェア楽曲推薦システム，DEIM Forum，2017，A8-3

[6] 新穂 龍太朗， 齋藤 康之， “画像の印象に合う楽曲の自動推薦システムに関する研究”；映像情報メディア学会 メディア工学研究会技術報告， ME2013-7， pp. 23-26， Feb. 2013.

[7] 佐々木将人，平井辰典，大矢隼士，森島繁生，”入力画像に感性的に一致した楽曲を推薦するシステム”，情報処理学会第 75 回全国大会，2D-5， 2013．3．6-8

[8] LIU, Wei, et al. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer，Cham，2016. p. 21-37.

[9] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the nes cycle. In Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[10] PALTOGLOU, Georgios; THELWALL, Michael. Seeing stars of valence and arousal in blog posts. IEEE Transactions on Affective Computing, 2013, 4. 1: 116-123.

[11] DHILLON, Inderjit S. ; MODHA, Dharmendra S. Concept decompositions for large sparse text data using clustering. Machine learning, 2001, 42. 1: 143-175.

[12] ALAJANKI, Anna; YANG, Yi-Hsuan; SOLEYMANI, Mohammad. Benchmarking music emotion recognition systems. PLOS ONE, 2016, 835-838.

[13] LECUN, Yann, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86.11: 2278-2324.

[14] Luo MR, Cui G, Rigg B. The development of the CIE 2000 colourdifference formula: CIEDE2000. Color Res Appl 2001;26:340350.

[15] Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48:7199.

[16] VINYALS, Oriol, et al. Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3156-3164.

[17] KARPATHY, Andrej; FEI-FEI, Li. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3128-3137.

[18] YOU, Quanzeng, et al. Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 4651-4659.

[19] HE, Kaiming, et al. Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017. p. 2980-2988.

[20] 相馬一郎. 色彩と感情. テレビジョン, 1967, 21.12: 858-865.