画像の印象に合った音楽の推薦システムの提案

追木　智明† 欅　惇志†,†† 宮崎　純†

† 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
†† 国立研究開発法人科学技術振興機構　ACT-I
E-mail: †{tuiboku,keyaki}@lsc.cs.titech.ac.jp, ††miyazaki@cs.titech.ac.jp

あらまし　本研究では同一空間上に画像と音楽をプロットすることで画像の印象に合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推薦を行っていたが，ユーザスタディ自体が高コストであるという問題点が挙げられる．そこで本研究では，代替の手法として一般物体認識による画像中の物体の抽出，Word2Vec を用いた物体から印象語への変換，各物体への重み付けによって同一空間上への印象語のプロットを行った．さらに，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，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い，ユーザによって画像の印象に合った音楽が推薦されたと評価された比率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 機械学習

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において，ユーザスタディによって画像や音楽を同一空間上に射影してマッピングする手法が主流である．

それに対して本研究では，画像から共通空間への射影において，1. 画像上の物体を一般物体認識を用いて抽出し，2. 単語間の意味的な類似度を測る手法として Word2Vec を用い，画像から抽出された物体から印象語への変換とプロット，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同一空間上へ射影，対応付け，推薦を行うことを目指す．

2. 基礎知識

本節では，本研究で用いられている機械学習の手法について説明する．

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 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．

2.3 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 も同様に-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 に印象語が AV 空間上にプロットされている例を示す．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，原点からの角度が感情の意味合い，長さが感情の強さを意味する．

2.4 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．

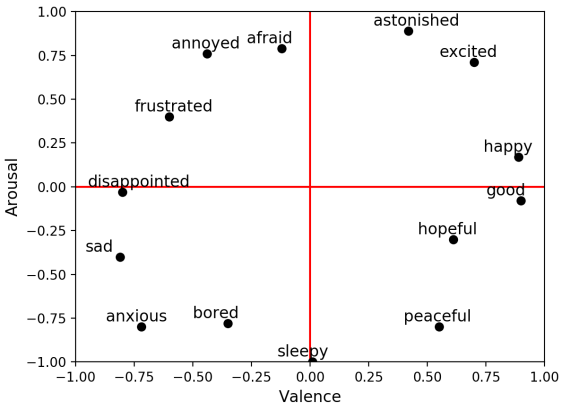

図 1　AV 空間上の印象語

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．具体的に，画像では色特徴量と形状特徴量を用いており，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，画像中の物体を抽出に一般物体認識，単語間の意味的な類似度を測る手法に Word2Vec といった機械学習の手法を用いている．本システムでは画像と音楽それぞれに対して共通空間である AV 空間に射影するための処理が行われる (図 2)．画像に対しては以下のステップによって画像から抽出された印象語を AV 空間にプロットする．また，本研究では，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音楽を AV 空間上にプロットする (注1)．

（1） 一般物体認識による画像上の物体抽出
（2） 物体の面積による各物体の重み付け
（3） Word2Vec による物体から印象語への変換
（4） 重みを考慮した印象語のプロットとクラスタリング
（5） 多数決法による象限の決定

ステップ (1) 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 (2) 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．また，ステップ (3) で抽出された物体に対して Word2Vec を用いて印象語の類似度を計算することで印象語への変換を行う．続いて，ステップ (4) にて各印象語の座標と重みを用いて AV 空間への印象語のプロットを行い，プロット結果に対するクラスタリングを行う．ステップ (5) でステップ (4) のクラスタリングの結果となる複数のクラスタの中心に対して各象限に含まれるプロットされた印象語の個数による多数決を行い，推薦対象を絞り込み，推薦結果に含まれる不適合な音楽の削減を行う．本研究の提案手法の利点として，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を AV 空間上へ射影することが可能になるということが挙げられる．以項で各ステップの詳細について説明する．

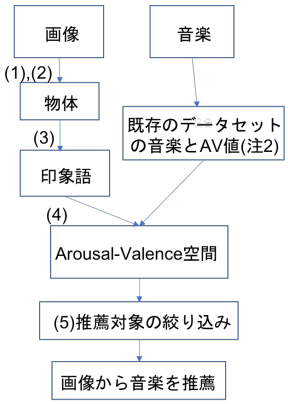

図 2　システムの流れ

4.1 画像上の物体の抽出

ステップ (1) 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(1),(2) 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

(注1)：本音楽データセットでは長さが 45 秒の音楽 1000 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．

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．

4.2 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車を検知し面積を計算している (図 3(2))．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 (図 3(3), (4))．

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

($dmax$: 物体ごとの総面積の最大値, $d(X)$: X の総面積)

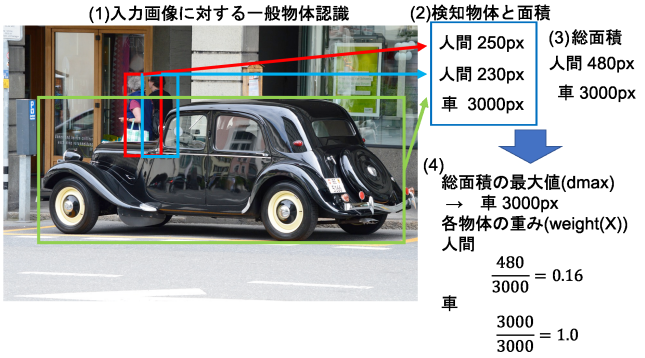

図 3　各物体の重みの計算

4.3 印象語への変換

ステップ (1) で抽出された物体に対して Word2Vec を用いて各物体と各印象語の類似度を計算しソートすることで各物体に対して近いと考えられる印象語を抽出する．図 4 の例では，図 3 の例の抽出結果である人間と車に対して Word2Vec を用いて AV 値が既知である印象語 (注2) との類似度を計算し，類似度順でソートしている．Word2Vec の学習用コーパスは Jure Lec ら [9] の “MemeTracker Raw Phrases Data” を用いている．このコーパスは過去のブログやニュース記事を集めたものであり，全体で 1700 万種類の異なるフレーズが含まれる．

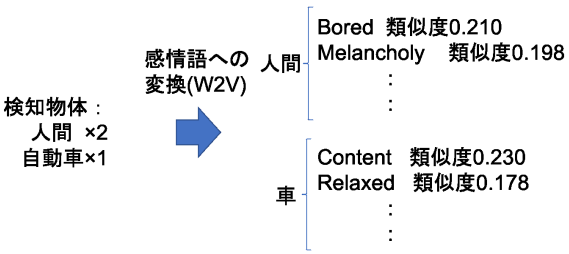

図 4　物体から印象語への変換

(注2)：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．

4.4 重みを考慮した印象語のプロット

ステップ (1),(2) によって物体 X から変換された印象語の座標 (x,y) に対してステップ (3) で算出した各印象語の重み weight(X) を掛け，印象語の座標を (x*weight(X),y*weight(Y)) とすることで印象語の原点からの長さ (感情の強さ) を調整し，AV 空間上にプロットする (図 5)．さらに，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，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．なお，AV 空間の象限数は 4 であるためクラスタの個数を 4 つに設定している．

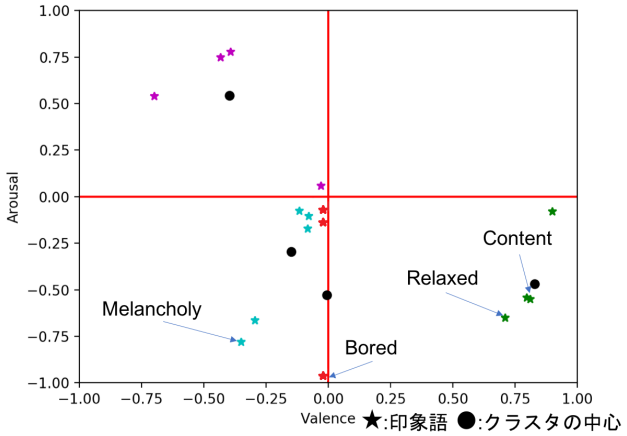

図 5　印象語のプロット

4.5 印象語数の多数決による推薦対象象限の絞り込み

ステップ (5) は以下の手順で行われる．

(A)　各象限にプロットされた印象語数の集計

(B)　集計結果に基づく推薦対象象限の絞り込み

ステップ (A) では図 5 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 6). ここで，画像 162 枚に対してステップ (4) までの手法を実行し，各象限にプロットされた印象語の累計結果すると，第 3,4 象限にプロットされる印象語数が多く，第 1 象限にプロットされる印象語数が極端に少ないことがわかる (表 1)．そこで，1 枚の画像から各象限にプロットされる印象語数の平均値を計算 (表 1) し，ステップ (B) で集計結果を各象限の画像 1 枚あたりの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 6)．

表 1　画像 162 枚から各象限にプロットされた印象語数

象限	1	2	3	4
印象語数	46	152	624	448
画像 1 枚あたりの印象語数	0.29	0.94	3.88	2.78

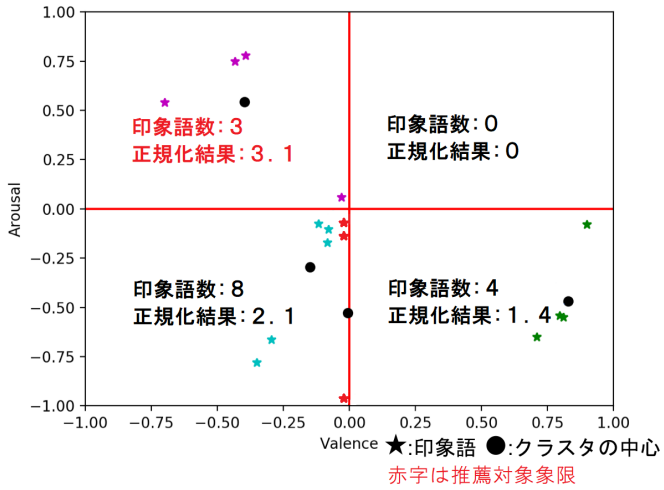

図 6　推薦対象象限の絞り込み

4.6　推薦結果音楽の決定

ステップ (4) 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 7 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．

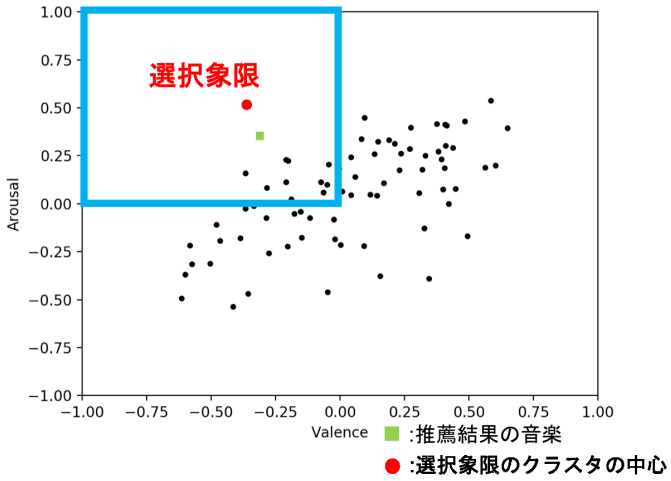

図 7　推薦象限の絞り込みと音楽の推薦結果

5.　評 価 実 験

提案手法による性能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

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 “Emotion in Music Database (1000 songs)” [12] を用い，音楽を AV 空間上にプロットした．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 1000 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

提案手法の評価のため，20～30 歳台の計 11 人に対して画像を 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．その結果と提案手法によって選択された象限が一致した物を正解とした時，正解率は 68/110(61.8 %) という結果になった．

表 2 に提案手法による象限の選択割合とユーザによる象限の選択割合，提案手法による選択象限とユーザの選択象限の一致率を示す．結果から，第 2 象限の選択率が提案手法，ユーザによる選択の両方で低いことがわかる．理由として，画像に対して AV 空間の第 2 象限における感情の意味合いである Excited + Negative(怒り) に相当する印象を抱く事が少ないからであると考えられる．また，本論文では画像から物体を抽出し印象語に変換するというステップによって AV 空間上にプロットするという手法を用いているが，画像に対する物体認識に加え，画像中の人間の顔の表情による感情推定が可能である Google Cloud Vision API(注3) や，画像の内容を表した文章を生成することが可能である Microsoft Azure の Computer Vision API(注4) を用いることで，画像中に写っている人間の評定による感情推定や画像から変換された文章に対する感情推定を行うことによる，物体の状態や物体間の関係性を考慮した感情推定や，画像中の色や彩度を考慮した感情推定手法を今後検討していく．

表 2　提案手法による象限の絞り込みに関する評価 (データ数:110)

象限	1	2	3	4
ユーザの選択率 (%)	36.4	5.4	22.7	35.5
提案手法による選択率 (%)	30.0	10.0	30.0	30.0
正解率 (%)	65.0	0.0	76.0	59.0

5.2　推薦結果に関する評価

本節では下記の 3 つの手法に対して評価を行う．全推薦手法とそれ以外の手法を比較することで推薦対象の絞り込みの有効性，提案手法とユーザ指定手法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．

- 推薦対象の絞り込みを行わない（全推薦手法）
- 多数決による象限の決定 (提案手法)
- ユーザが推薦対象の象限を選択 (ユーザ指定手法)

評価として 5.1 節での実験に用いた画像 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．アンケートの結果として各画像に対して推薦された音楽への各ユーザの評価値の平均 (以降，推薦スコアとする) のヒストグラムを示す (図 8)．また，各手法に関して推薦スコアが 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の 3 項目を表 3 に示す．なお，推薦の失敗と

(注3)：https://cloud.google.com/vision/?hl=ja

(注4)：https://azure.microsoft.com/ja-jp/services/cognitive-services/computer-vision/

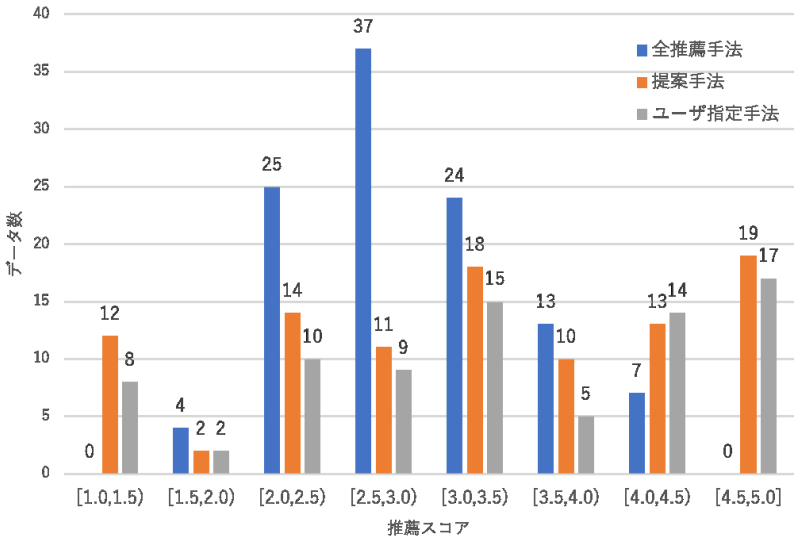

図 8　各種法の推薦スコアのヒストグラム

は提案手法とユーザ選択手法によって決定された象限に推薦対象が存在せず推薦が行えないことを指す．

表 3　各手法の比較

	推薦スコア 4 以上 (%)	全体平均	推薦失敗割合 (%)
全推薦手法	6.4(7/110)	2.79	0
提案手法	32.3(32/99)	3.06	10
ユーザ指定手法	38.6(31/80)	3.2	27.3

結果の考察として，図 8 から全推薦手法では推薦スコアが [2.0,3.5) の範囲に含まれる割合が約 78 %と非常に大きいことがわかる．理由として，全推薦手法では象限の絞り込みを行わずに推薦を行うため，推薦結果に適合度の高くない音楽が含まれることで推薦スコアが下がっていると考えられる．また，提案手法とユーザ指定手法によって象限の絞り込みを行うことで推薦スコア 4 以上の音楽の割合が全推薦手法を大きく上回っていることから，象限の絞り込みは有効であると言える．次に，提案手法とユーザ指定手法を比較すると推薦スコア 4 以上の割合と全体平均の両指標においてユーザ指定手法が上回っているが，推薦失敗時の推薦スコアを 0 とした場合，提案手法の全体平均は 2.76，ユーザ指定手法の全体平均は 2.31 であり，現状の手法においてユーザ指定手法が最適な手法であると結論付けるのは難しい．

6. ま と め

本論文では，画像に対して一般物体認識を行い，Word2Vec を用いることで，画像の持つ印象にふさわしい音楽推薦システムの提案を行った．また，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，提案手法による象限の選択とユーザが選択した象限が一致するかどうかを調査し，約 6 割の割合で提案手法による象限の選択とユーザによる象限の選択が一致することを示した．また，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，全推薦手法，提案手法，ユーザ選択手法の 3 つの手法に分けて評価し，提案手法とユーザ選択手法において推薦スコアが 4 以上になる割合が全推薦手法を上回ったことから象限の決定の有効性を示した．

今後の展望として，画像から生成された文章や画像中の人の表情といった要素による物体の状態や物体間の関係を考慮した感情抽出を行う手法や，画像中の物体に加え，画像上の色や彩度を考慮した手法を今後考えていく予定である．また，本論文では音楽を AV 空間にプロットする際に既存のデータセットを用いているが，最終目的として音楽に関してもユーザスタディを用いない手法でのプロットを目指しているため，関連研究にも用いられていた音響特徴量や歌詞などといった基本的な情報から感情推定を行うこと等を行うことによって画像と同様にユーザスタディを用いない手法を検討する．

7. 謝　　辞

本研究の一部は，JSPS 科研費 (JP15H02701，JP16H02908，JP15K20990，JP17K12684)，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．

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] SOLEYMANI， Mohammad， et al． 1000 songs for emotional analysis of music． In: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia． ACM， 2013． p． 1-6．
[13] LECUN, Yann, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86.11: 2278-2324.