DBSJ Japanese Journal
Vol. 15-J, Article No. 10, March 2017

画像・テキスト・感情語の潜在的な相関に基づく画像の感情分類

Image Sentiment Classification Using Latent Correlations Among Visual, Textual, and Sentiment Features

桂井 麻里衣♡ 佐藤 真一♢

Marie KATSURAI Shin'ichi SATOH

画像の感情分類は，情報検索・推薦やデータマイニングの高度化，マーケティングやヘルスケア応用など，多くの波及を見込める研究課題として注目を集めている．従来より，感情分類に適した画像特徴の設計方法が検討されてきたが，画像に付与されたテキスト情報の利用については議論が少ない．そこで本文では，画像・テキスト・感情語という三つの側面に基づく画像の感情分類手法を提案する．提案手法は，与えられた画像について利用可能なモダリティの特徴を共通空間で取り扱うために，モダリティ間の相関を最大とする埋め込み空間への射影を算出する．求めた埋め込み空間での新たな特徴表現に基づき画像の感情分類器を学習する．クラウドソーシングを通じて構築した画像データセットに基づく実験により，提案手法は従来の画像特徴のみを用いた手法に比べて高い感情分類精度を示した．

Image sentiment classification has received increasing research attention for several applications including affective image retrieval, opinion mining about social events, product marketing, and healthcare. Most conventional methods focus on the design of visual features, and the use of text associated to the images has not been sufficiently investigated. This paper presents an approach that exploits latent correlations among visual, textual, and sentiment features for image sentiment classification. In the proposed method, we find a latent embedding space in which correlations among the three features are maximized. The projected features in the latent space are used to train a sentiment classifier, which considers the complementary information from different views. Results of experiments conducted on Flickr and Instagram images demonstrate the effectiveness of our approach.

1. はじめに

スマートフォンやソーシャルメディアの普及に伴い，ユーザは日常生活で気軽に画像を撮影し，ウェブで体験を共有するようになった．例として，画像共有サイトの Flickr[1] では，2015 年に 100 億枚の画像保持を達成しており，Instagram[2] ではアクティブユーザが 4 億人を突破したことが報告されている．これらの大量の画像は世界中の出来事を視覚的に表す情報源としてみなすことができ，そこで表出されるユーザの感情はマーケティングや世論調査に重要な役割を担う [1]．また，感情に基づく画像検索 [2] や，ポジティブ心理学・ヘルスケアへの応用 [3] も期待されている．このように，画像の表す感情極性（ポジティブ，ネガティブ）の自動分類（以降，感情分類）は，多くの波及が見込める研究課題として近年急激に注目を集めている [4–8].

従来研究では，感情心理学や芸術理論に基づき，感情分類に適した画像特徴の設計方法が議論されてきた [9]．しかしながら，感情という高次概念と画像特徴の間に存在する affective gap により，画像特徴を直接感情に関連付けるのは困難である．一方，感情分類よりも広義な画像アノテーションの文脈では，画像がもつタグや説明文などのテキスト情報を学習時に相補的に用いることで，アノテーションを高精度化できることが報告されている [10]．このようなテキスト情報の利用は，画像の感情分類の枠組みでは未だ議論が少ない．特に，従来研究で用いられてきた芸術的な画像を含む小規模データセットに比べ，ソーシャルメディアに投稿された雑多な画像を対象とする場合は，画像の意味内容を捉える必要がある．

そこで本文では，画像特徴と感情の affective gap を低減させるために，画像・テキスト・感情語の潜在的な相関に基づく画像の感情分類手法を提案する．提案手法では，ソーシャルメディアから画像と周辺のテキストを収集し，それぞれから特徴量を算出する．このとき，テキストの感情表現を補助するために，外部の感情語辞書である SentiWordNet [11] を導入する．SentiWordNet は，英語テキストの感情分類に広く用いられており，画像タグの感情スコア算出にも適用可能である [12]．次に，正準相関分析 [13] を用いて各変量から埋め込み空間への射影行列を算出する．射影先における複数モダリティの特徴の距離を最小化することで，各モダリティの射影行列はその他のモダリティとの相関に基づき学習される．最後に，埋め込み空間における新たな特徴表現に基づき，テスト画像が複数のモダリティをもつ場合・もたない場合に応じて感情分類器を学習する．従来手法との比較実験には，Flickr および Instagram から収集した画像データセットに対し，クラウドソーシングにより感情ラベルを付与した．本文の最後には，構築したデータセットを用いた実験により，従来手法と比較した提案手法の有効性を示す．

以上をまとめると，本研究による主な貢献は次の通りである．

- 画像特徴・テキスト特徴・感情語特徴を導入することで，感情分類に適した埋め込み空間を設計する．
- クラウドソーシングを通じて独自に構築した感情ラベル付き画像データセットを用いて従来手法との比較実験を行い，テキストおよび感情語特徴導入の有効性を示す．

本文の構成は以下の通りである．まず，2 章において画像の感情分類およびクロスモーダル検索の従来研究を説明する．3 章で画像・テキスト・感情語特徴の抽出方法を説明し，4 章で複数モダリティの相関に基づく画像の感情分類手法を提案する．5 章では，感情ラベル付き画像データセットを用いた比較実験を行い，提案手法の有効性を評価する．最後に，6 章において本文をまとめ，今後の方向性について検討する．

2. 関連研究

2.1 画像の感情分類

画像と感情の関係のモデル化は，次世代データマイニング・マルチメディア検索を支える重要なトピックの一つであり，画像特徴を入力・感情ラベルを出力とした教師あり学習が主流である．従来より，感情心理学や芸術理論の知見に基づき，感情分類に適した色特徴やテクスチャ特徴の設計方法が議論されてきた [9, 14]．これらの研究では，数百枚程度の芸術的な画像からなる比較的小さなデータセット [15] が性能評価に用いられている．一方，画像

♡ 非会員　同志社大学理工学部情報システムデザイン学科
katsurai@mm.doshisha.ac.jp
♢ 非会員　国立情報学研究所コンテンツ科学研究系
satoh@nii.ac.jp

[1] https://www.flickr.com/
[2] https://www.instagram.com/

検索や物体認識で広く用いられてきた色ヒストグラムや SIFT 特徴量などを感情分類に用いる試みもある [4]．しかしながら，これらの低次特徴量と感情の間には大きな affective gap が存在し，直接的な対応付けは困難である．特に，ソーシャルメディアに投稿された画像中の感情を分析する場合は，単なる印象評価ではなく，画像の意味内容を認識・考慮することが重要である．そこで，低次特徴量を一度物体やシーンの認識結果にマッピングし，その認識結果を中間表現 (mid-level representation) とする手法が提案されている [5,6]．例として，文献 [5] では，感情に関連する形容詞・名詞ペアからなる 1,200 個のフレーズ（例："cute kids", "disgusting food"）を Flickr のテキストから自動で選出し，低次特徴量を用いて各フレーズの識別器を学習する．さらに，各画像に対する識別器の出力値を並べた 1,200 次元のベクトルを中間表現として感情極性の分類器を学習する．同様に，文献 [6] では，顔画像データベースを用いて予め表情識別器を学習し，画像中の表情認識結果を中間表現とする．以上の従来研究は，いずれも感情に適した画像特徴・中間表現の算出に着目しており，トレーニング画像およびテスト画像がもつテキスト情報の利用については議論が少ない．そこで本文では，画像とテキストを相補的に用いた感情分類手法を提案し，その効果を検証する．

近年は様々な画像認識タスクで CNN による性能向上がめざましく，画像の感情分類においても CNN が利用され始めている [7,8]．具体的には，感情分類のための CNN の学習 [7] や，既存の CNN から得られる特徴量の利用 [8] が挙げられる．本文では，文献 [8] と同様のアプローチで CNN に基づく画像特徴を算出し，提案手法に導入する．

2.2 画像・テキストの相関算出

これまで画像アノテーションやクロスモーダル検索の研究において，画像とテキストの相関が示されてきた [16,17]．著者の以前の研究においても，形容詞や感情語を含む高次概念間の関係抽出に対する複数モダリティ導入の有効性を示した [18]．そこで本研究では，画像の感情分類を狭義の画像アノテーション問題として捉え，トレーニング画像およびテスト画像がもつテキスト情報を分類器に導入した際の性能検証に取り組む．

異なるモダリティからの特徴を同一空間で取り扱うための代表的な手法に正準相関分析 [19] がある．従来の正準相関分析が変数間の線形の関係のみをモデル化するのに対し，二変量間の非線形な関係を捉えるカーネル正準相関分析 [13] やディープ正準相関分析 [20] が提案されている．しかしながら，これらの手法は非常に多くの計算量やメモリを必要とするため，大規模データセットでの適用は困難である．そのため，文献 [16] では，一般化正準相関分析に対し explicit feature map を導入することで非線形性を近似している．一方，文献 [17] のように，線形の正準相関分析により CNN 特徴量とテキスト特徴の相関を捉えることに成功した例を鑑み，本文も通常の一般化正準相関分析を用いる．

3. 特徴算出

本章では，画像とテキストの対が与えられたとき，画像特徴，テキスト特徴，感情語特徴を算出する方法を説明する．

3.1 画像特徴の算出

提案手法では，画像特徴として CNN 特徴量 [21] を用いる．CNN 特徴量とは，あるタスクに向けて予め学習された CNN に画像を入力し，全結合層から得られる出力を要素にもつベクトルを指す．本文では，ILSVRC2012 データセット [3] を用いて学習された 8 層 CNN [22] に画像を入力し，7 層目の出力となる 4096 次元ベクトルを用いる．得られたベクトル集合に対し主成分分析を適用し，512 次元へ削減する．CNN 特徴量は近年画像検索やアトリビュート認識などの様々な画像認識タスクで性能向上を示しており [17]，本文の実験においても従来の画像特徴と性能を比較する．

[3] http://image-net.org/challenges/LSVRC/2012/

表 1: SentiWordNet に収録されている synset と感情スコアの例．

POS	Synset ID	PosScore	NegScore	Synset terms
v	02708707	0.125	0.000	vacation#1 holiday#1
a	00534250	0.000	0.375	obscure#2 dark#8
n	09376198	0.000	0.000	ocean#1
n	10112591	0.125	0.000	friend#1
a	00013887	0.000	0.250	abundant#1
n	07126383	0.000	0.625	moan#1 groan#1
n	06696483	0.500	0.000	laurels#1 honour#2 honor#1 award#2 accolade#1

3.2 テキスト特徴の算出

ソーシャルメディアの画像には，タイトルやタグ，投稿者からの説明文，画像閲覧者からのコメントなど様々なテキストが存在する．ここで，コメントには画像の芸術性に対する感想が混在するため [23]，画像内容に対する感情表現のみを抽出することは困難である．したがって提案手法では，画像投稿者以外からのコメントを除外し，画像タグと説明文からテキスト特徴を算出する．

ユーザが付与した画像タグや説明文には，単語の欠落やノイズが多いことがよく知られている．この問題を解決するために，テキスト集合から得られる単語間の関係を導入する．具体的には，以下の二通りの方法でテキスト特徴を次元削減する．

1. **BoW+SVD.** 各画像のテキストから単語セットを抽出し，Bag-of-Words (BoW) ベクトルを算出する．得られるベクトルは非常に高次元となるため，次元削減のためにスパース行列のための特異値分解 (Singular Value Decomposition: SVD) [24] を適用する．具体的には，テキスト特徴の行列 $\mathbf{X}_t$ を $\mathbf{X}_t = \mathbf{U}_t\mathbf{S}_t\mathbf{V}_t$ の形で分解し，次元削減後の行列を $\mathbf{U}_t\mathbf{S}_t$ で表す．本文では，SVD で得られるテキスト特徴の次元数を 1,500 とする．
2. **Skip-gram.** データセット中の単語をベクトル空間で表すために Skip-gram [25] を用いる．Skip-gram は各単語に対し文中の周辺単語の予測に有用な表現を高速に獲得するモデルである．2015 年 12 月 30 日時点の英語版 Wikipedia 全記事 [4] をダンプして得られたコーパスにおいて，出現回数が 5 回以上の単語のみを選択する．特徴ベクトルの次元数を 400 に設定して Skip-gram を学習し，単語 w に対し意味ベクトル $\boldsymbol{y}(w)$ を算出する．次に，画像のテキストから単語セット W を抽出し，次式の平均ベクトル $\boldsymbol{t} \in \mathbb{R}^{400}$ をテキスト特徴とする．

$$\boldsymbol{t} = \frac{1}{|W|}\sum_{w \in W} \boldsymbol{y}(w). \quad (1)$$

ここで，$|W|$ は単語セット W に含まれる単語の総数を表す．

本文の実験では，上記の BoW+SVD および Skip-gram によるテキスト特徴の性能をそれぞれ評価する．

3.3 感情語特徴の算出

提案手法では，テキストからの感情表現の抽出を補助するため，外部の知識源として SentiWordNet を用いる．SentiWordNet とは，WordNet [26] 内の synset と呼ばれる同義語集合に対しポジティブスコア，ネガティブスコアを付与した感情語辞書であり，テキストの感情分析で広く用いられている [27]．SentiWordNet に収録されている synset と感情スコアの例を表 1 に示す．この synset の感情スコアを用いて単語の感情スコアを算出する．具体的に，$|\Phi_w|$ が単語 w が属する synset の集合，pos_s，neg_s をそれぞれ synset s に付与されたポジティブスコアおよびネガティブス

[4] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

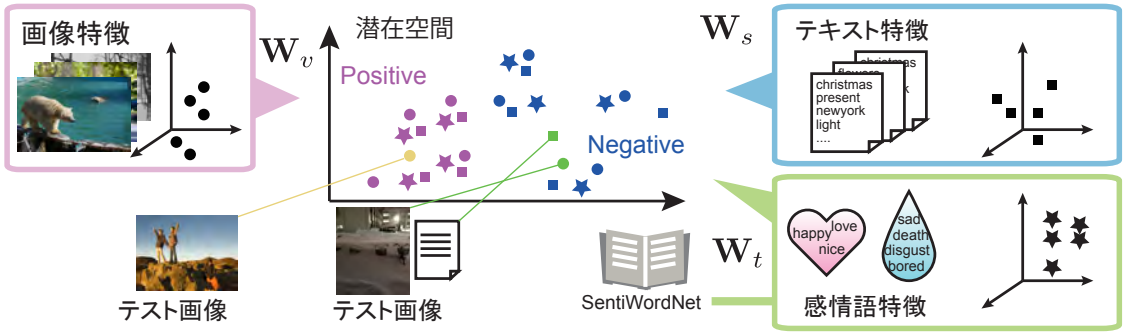

図 1: 提案手法の概要．特徴抽出，埋め込み空間への射影，新たな特徴表現に基づく感情分類という三つのステップから構成される．

表 2: SentiWordNet から得られる感情語の例．式 (2) で算出されたスコアによりポジティブまたはネガティブとみなされた単語を上位 10 個ずつ示す．

ポジティブ単語	スコア	順位	ネガティブ単語	スコア
estimable	1.50	1	miserable	-1.91
healthy	1.48	2	wretched	-1.82
gracious	1.47	3	unsound	-1.80
happy	1.45	4	deplorable	-1.73
fortunate	1.44	5	nasty	-1.68
lucky	1.44	6	unlawful	-1.55
majestic	1.42	7	atrocious	-1.50
superiority	1.33	8	painful	-1.50
brilliant	1.32	9	lowly	-1.50
urbanity	1.31	10	unhappy	-1.50

コアとしたとき，単語 w の感情スコアを次式で定義する．

$$Pos(w) = \left(\sum_{s \in \Phi_w} \frac{1}{r_s^w} \right)^{-1} \sum_{s \in \Phi_w} \frac{1}{r_s^w} (pos_s - neg_s). \quad (2)$$

ここで，r_s^w は単語 w が synset s 中で現れた順位を表す [5]．表 1 の例では，s = “02708707”，w = “holiday” としたとき，$r_s^w = 2$ となる．$Pos(w)$ の値が大きいほど単語 w がポジティブであることを意味し，値が小さいほどネガティブであることを意味する．式 (2) を用いて SentiWordNet 中の単語の感情スコアを算出した際に，値が上位 10 個および下位 10 個となる単語を表 2 に示す．

次に，画像の単語セット W が与えられたとき，テキスト全体の感情スコアを次式で算出する．

$$SentiScore = \sum_{w \in W} Pos(w). \quad (3)$$

今回は，単語セット W 中に感情語が多いほど信頼性が高いとみなし，単語数がスコアに与える影響を大きくした．得られた値は次式の閾値処理に基づきポジティブスコア s_{pos} へと変換する．

$$s = \begin{cases} \tau & \text{if } SentiScore \geq \tau, \\ SentiScore & \text{if } \tau > SentiScore > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

ここで，τ は閾値のパラメータであり，本文では実験的に $\tau = 1.5$ と設定した [6]．s の上限を設定する理由は，感情語が大量に付与された画像などの外れ値を防ぐためである．同様の閾値処理により，式 (3) が負の値をとる場合にネガティブスコア s_{neg} も算出する．最終的に，画像の感情語特徴は $\boldsymbol{s} = [s_{pos}, s_{neg}]$ で表される．

以降，画像，テキスト，および感情語をそれぞれインデックス v, t, s で表す．

[5]著者らの予備実験では，単語の順位を考慮しない場合に比べて効果的に感情語を抽出できたため，本文では順位による重み付けを採用する．

[6]τ が提案手法の性能に与える影響の調査は今後の課題とする．

4. 複数モダリティの相関に基づく画像の感情分類

本章では，画像・テキスト・感情語という三つの側面の潜在的な相関に基づく画像の感情分類手法を提案する．提案手法の概要を図 1 に示す．はじめに，トレーニング画像から画像特徴，テキスト特徴，感情語特徴をそれぞれ算出する．次に，正準相関分析を用いて各モダリティから埋め込み空間への射影を算出し（4.1 節），得られる新たな特徴表現に基づき画像の感情分類器を学習する（4.2 節）．以降，各ステップの詳細を説明する．

4.1 画像，テキスト，感情語を用いた埋め込み空間の推定

いま，n_1 枚の感情ラベル付き画像セット Ω_1 と，n_2 枚のラベルなし・テキスト付き画像セット Ω_2 が与えられているとする．これらを足し合わせた n ($n = n_1 + n_2$) 枚の画像から，3 章の方法で画像特徴ベクトル，テキスト特徴ベクトル，および感情語特徴ベクトルをそれぞれ算出する．このとき，感情ラベル付き画像セットについては，感情語特徴 $\boldsymbol{s}$ を正解ラベルに応じて最大値で置き換える．具体的には，感情ラベルがポジティブのときは $s_{pos} = \tau$，ネガティブのときは $s_{neg} = \tau$ とする．この操作により，感情ラベル付き画像のテキストから感情語が欠落している場合にも，感情表現を補って画像とテキストの相関を捉える．以上により得られた各モダリティの特徴行列をそれぞれ $\mathbf{X}_v \in \mathbb{R}^{n \times d_v}$，$\mathbf{X}_t \in \mathbb{R}^{n \times d_t}$，$\mathbf{X}_s \in \mathbb{R}^{n \times d_s}$ で表す．

提案手法では，三つのモダリティからの埋め込み空間を推定するために，一般化された正準相関分析を用いる．具体的には，次式のように，同一の画像からの特徴間の距離を射影先で最小化することで，i 番目のモダリティに対する射影行列 $\mathbf{W}_i \in \mathbb{R}^{d_i \times d}$，$(d = d_v + d_t + d_s)$ を算出する [13]．

$$\min_{\mathbf{W}_v, \mathbf{W}_t, \mathbf{W}_s} \sum_{i,j \in \{v,t,s\}} \|\mathbf{X}_i \mathbf{W}_i - \mathbf{X}_j \mathbf{W}_j\|_F^2 \quad (5)$$

$$\text{subject to } \mathbf{W}_i^T \mathbf{\Sigma}_{ii} \mathbf{W}_i = \mathbf{I}, \boldsymbol{w}_{ik}^T \mathbf{\Sigma}_{ij} \boldsymbol{w}_{jl} = 0, \quad (6)$$

$$i, j \in \{v, t, s\}, i \neq j, k, l = 1, \cdots, d, k \neq l. \quad (7)$$

上式において，$\|\mathbf{A}\|_F$ は行列 $\mathbf{A}$ のフロベニウスノルムを表し，$\mathbf{\Sigma}_{ij}$ は $\mathbf{X}_i$，$\mathbf{X}_j$ の間の共分散行列，$\boldsymbol{w}_{ik}$ は行列 $\mathbf{W}_i$ の k 番目の列ベクトルを表す．式 (7) の最小化は一般化固有値問題に帰着する．

三つ以上のモダリティに基づく正準相関分析は，近年のクロスモーダル検索や画像アノテーションで，画像特徴と二種類のテキスト特徴の相関を算出するために用いられている [16, 17]．本研究においても，テキストから得られる二種類の特徴を導入し，感情分類に適した埋め込み空間を求める．

4.2 埋め込み空間における画像の感情分類

で算出した射影行列により，埋め込み空間において特徴行列 $\mathbf{X}_i (i \in \{v, t, s\})$ は次式で表される．

$$\mathbf{P}_i = \mathbf{X}_i \mathbf{W}_i \mathbf{D}^p. \quad (8)$$

ここで，$\mathbf{D}$ は固有値を対角成分にもつ対角行列である．各次元の重要性は対応する固有値の大きさによって示されるため，パラメー

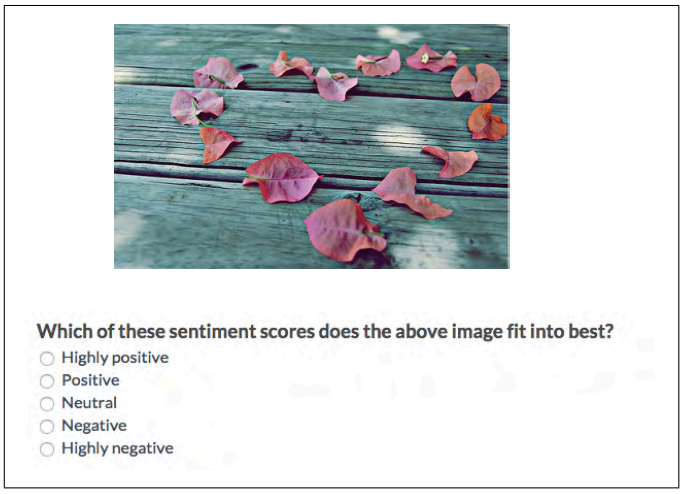

図 2: クラウドソーシングで用いたインタフェースのスクリーンショット．ユーザは提示された画像に対し五段階で感情極性を評価する．

タ p によって重み付けする．本文では，文献 [16,17] で良い精度を得た設定と同様，$p=4$ とする．

式 (8) を用いて，感情ラベル付き画像セット Ω_1 から $\mathbf{P}_v \in \mathbb{R}^{n_1 \times d'}$，$\mathbf{P}_t \in \mathbb{R}^{n_1 \times d'}$，および $\mathbf{P}_s \in \mathbb{R}^{n_1 \times d'}$ $(d' \leq d)$ を算出し，これらを結合したものを最終的な特徴行列とする．得られたトレーニング画像の新たな特徴表現と感情ラベルを用いて感情分類器を学習する．本文では，従来手法 [4,5,8] と同様に線形 SVM を用いる．テスト画像が与えられたときは，利用可能なモダリティから算出された特徴のみを埋め込み空間に射影すればよい．つまり，とりうる射影の組み合わせに応じて特徴行列の結合を変更し，線形 SVM を学習する．

5. 実験

本章では，提案手法の有効性を確認するために実験を行う．5.1 節では，実験用データセットの構築方法を述べる．5.2 節で比較手法を説明し，5.3 節で各手法の感情分類性能を検証する．

5.1 データセット構築

現在ウェブで公開されている画像の感情分類に関するデータセットはいずれも数百枚規模である [5,7,15]．文献 [4] では画像タグと SentiWordNet に基づき付与した擬似的な感情ラベルを正解データとみなしているが，画像タグの欠落やノイズを考えるとこの方法は信頼性が低い．そこで本研究では，クラウドソーシングを通じて人手でデータセットを構築する．まず，写真共有サイト Flickr および Instagram から，以下のように画像を収集した [28]．

- **Flickr データセット**
  文献 [29] で提供されている Flickr の画像 ID にしたがって画像およびテキスト情報を収集した．各ユーザにつき画像枚数を最大 70 枚に限定し，105,587 枚の画像を得た．このデータセットでの最頻出単語は “view”，“black”，“photo”，“canon”，“nikon”，“film” であった．
- **Instagram データセット**
  SentiWordNet に収録されている感情スコアの高い単語（例：“congratulations”，“terrible”）をクエリキーワードとし，Instagram API[7] を用いて 120,000 枚の画像を収集した．各ユーザにつき画像枚数を最大 10 枚と限定した．このデータセットの最頻出単語には “love”，“like”，“life”，“day”，“new” などがあり，Flickr データセットに比べユーザの日常生活をよく反映しているといえる．

[7] https://instagram.com/developer/

表 3: Flickr および Instagram データセットに対するクラウドソーシングによるアノテーション結果．ポジティブ，ニュートラル，またはニュートラルと評価した人数の組み合わせと，対応する画像枚数を示す．また，アノテータ間の極性の一致率を最下段に示す．

評価人数			データセット	
ポジティブ	ニュートラル	ネガティブ	Flickr	Instagram
3	0	0	21549	16364
2	1	0	20001	13351
2	0	1	6586	3361
1	2	0	11651	9897
1	1	1	9449	5270
1	0	2	3701	2302
0	3	0	3290	3858
0	2	1	5010	3558
0	1	2	4914	3260
0	0	3	3988	4218
アノテータ一致率			78.10%	83.29%

表 4: 各データセットから選出したポジティブ・ネガティブ画像の枚数．

	Flickr	Instagram
ポジティブ	41,552	29,715
ネガティブ	8,902	7,478

各データセット中の画像に対し，CrowdFlower[8] のクラウドソーシングプラットフォームを利用して一枚の画像につき三名のユーザからのアノテーションを得た．具体的には，図 2 に示すインタフェースのように，画像を一枚ずつ提示し，当てはまる感情を五段階（Highly positive, Positive, Neutral, Negative, Highly negative）で選択させた．ポジティブ，ニュートラル，またはネガティブと評価した人数の組み合わせと，対応する画像枚数を表 3 に示す．アノテータ三名のうち，ポジティブとニュートラルのみ，またはネガティブとニュートラルのみが選択された場合に一致したとみなしたとき，一致率は Flickr データセットで 78.10%，Instagram データセットで 83.29% となった．構築した感情ラベル付きデータセットはウェブで公開する（`http://mm.doshisha.ac.jp/senti/CrossSentiment.html`）．

感情心理学の研究では，性別や文化の違いが感情評価に影響を及ぼすとの議論があるが [30]，本研究ではマルチメディア検索での従来研究にならい，単純にアノテータ間で極性が不一致となった画像を除外し，意見が合致した画像のみを実験に用いる．得られたデータセットの画像枚数を表 4 に示す．表からも読み取れるように，ソーシャルメディアにはポジティブと評価される画像の方が多い傾向があった．実験ではポジティブ・ネガティブともに同じ枚数をサンプリングする．

5.2 比較手法

本実験では，以下の手法との性能比較を行う．

- **Random:** テスト画像をランダムに分類する．
- **Low [4]:** HSV 色ヒストグラムと，SIFT 特徴の Bow を結合したベクトルを用いて線形 SVM を学習する．
- **SentiBank [5]:** 低次特徴量に基づく 1,200 個のフレーズの認識結果を中間表現とし，線形 SVM を学習する．
- **CNN [8]：** CNN の第 7 層から得られる 4096 次元ベクトルを用いて線形 SVM を学習する．
- **CNN+PCA：** CNN の第 7 層から得られる 4096 次元の特徴に対し主成分分析を適用し，128 次元へ削減したあと線形 SVM を学習する．
- **BoW+SVD:** BoW+SVD により算出された 1,500 次元のテ

[8] http://www.crowdflower.com/

表 5: 各データセットにおける感情分類の正解率．5 回の試行の平均および標準偏差を示す．

	データセット	
手法	Flickr	Instagram
Random	50.39 ± 0.69%	50.54 ± 0.65%
Low [4]	66.99 ± 0.52%	64.24 ± 0.60%
SentiBank [5]	71.61 ± 0.18%	68.50 ± 0.61%
CNN	69.80 ± 0.34%	66.48 ± 0.55%
CNN+PCA	77.51 ± 0.38%	74.35 ± 0.72%
BoW+SVD	72.46 ± 0.48%	73.32 ± 0.17%
Skip-gram	73.01 ± 0.23%	72.76 ± 0.28%
CNN+PCA+BoW+SVD+S	80.43 ± 0.31%	79.15 ± 0.31%
CNN+PCA+Skip-gram	78.58 ± 0.29%	75.89 ± 0.32%
CNN+PCA+Skip-gram+S	79.30 ± 0.33%	78.45 ± 0.24%
LC(V+T)+P(V)	77.54 ± 0.30%	75.01 ± 0.39%
LC(V+S)+P(V)	77.84 ± 0.30%	74.72 ± 0.38%
LC(V+T+S)+P(V)	78.38 ± 0.34%	75.57 ± 0.50%
LC(V+T)+P(V+T)	79.28 ± 0.51%	78.60 ± 0.71%
LC(V+T+S)+P(V+T)	81.20 ± 0.43%	80.04 ± 0.67%
LC(V+T+S)+P(V+T+S)	**81.25 ± 0.34%**	**80.17 ± 0.32%**

キスト特徴を用いて線形 SVM を学習する．

- **Skip-gram:** Skip-gram による 400 次元のテキスト特徴を用いて線形 SVM を学習する．
- **CNN+PCA+BoW+SVD:** CNN+PCA(128 次元) と BoW+SVD(1,500 次元) を結合したベクトルで線形 SVM を学習する．2 つのモダリティを用いた手法である．
- **CNN+PCA+Skip-gram:** CNN+PCA（128 次元）と Skip-gram（400 次元）を結合したベクトルで線形 SVM を学習する．2 つのモダリティを用いた手法である．
- **CNN+PCA+BoW+SVD+S:** CNN+PCA+BoW+SVD に，提案手法で用いた感情語特徴（2 次元）を結合したベクトルで線形 SVM を学習する．三つのモダリティを用いた手法である．
- **CNN+PCA+Skip-gram+S:** CNN+PCA+Skip-gram に，提案手法で用いた感情語特徴（2 次元）を結合したベクトルで線形 SVM を学習する．三つのモダリティを用いた手法である．

各手法の線形 SVM の学習には Liblinear[9] を用いた．SVM のパラメータ C はトレーニングデータに基づくクロスバリデーションで決定した．

さらに，提案手法における複数のモダリティの効果を検証するために，とりうる組み合わせごとに性能を評価する．例として，画像特徴およびテキスト特徴を用いて埋め込み空間を求めた場合を LC(V+T)，三種類の特徴すべてで埋め込み空間を求めた場合を LC(V+T+S) として表す．同様に，テスト画像の画像特徴のみを射影した場合を P(V)，画像特徴およびテキスト特徴を射影した場合を P(V+T) と表記する．

5.3 感情分類の性能評価

表 4 に示した画像から，トレーニング用またはテスト用画像をランダムに 5 回選出した．具体的に，Flickr データセットでは，各感情極性に対し 6,000 枚の画像をトレーニングセット，2,500 枚の画像をテストセットとした．また Instagram データセットでは，各感情極性に対し 5,000 枚の画像をトレーニングセット，2,400 枚の画像をテストセットとした．手法の性能評価の指標として，一回の試行につき次式の正解率を算出する．

$$\text{正解率} = \frac{\text{正しく感情分類が行えた画像の枚数}}{\text{テスト画像の総数}} \quad (9)$$

[9] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

全ての試行における正解率の平均および標準偏差を表 5 に示す．表より，同じ数のモダリティを用いた場合，提案手法は比較手法よりも高精度に感情を分類できていることがわかる．特に LC(V+T+S)+P(V) の結果から，テスト画像にテキスト情報が全く存在しない場合であっても，テキストおよび感情語特徴を用いた埋め込み空間が性能向上に貢献するといえる．三変量すべてが利用可能な場合はいずれの手法も精度が大きく向上したが，提案手法の LC(V+T+S)+P(V+T+S) が最も高い精度を示した．

提案手法で従来の画像特徴 [4,5] を用いた場合の性能は文献 [28] で検証したが，本実験では CNN 特徴量を用いることで正解率が大きく向上することを確認した．こうした CNN 特徴量の有用性は，CNN+PCA が Low や VSO を圧倒していることからも明らかである．今後は，ImageNet で事前学習済みの 8 層 CNN をファインチューニングし，感情分類に特化した画像特徴を算出する予定である．

Flickr データセットの一回目の試行でテスト用に選出されたクリエイティブ・コモンズ画像のうち，LC(V+T+S)+P(V+T+S) によってポジティブまたはネガティブと分類された上位 24 枚の画像を図 3 に示す．画像キャプションは Flickr ユーザ ID に対応し，赤枠は誤分類された画像を表す．いずれの画像もポジティブまたはネガティブな感情が表出されており，青空を背景にもつものの墓石がメインとなる画像もうまく分類できている．一方で，図 3 (b) にはポジティブのラベルをもつ画像も混在した．今後は，提案手法で推定した埋め込み空間を用いることで，各モダリティの特徴からの推定結果に一貫性がみられる画像・そうでない画像を分別し，より分類性能を向上させる．

6. おわりに

本文では，画像・テキスト・感情語という三つの側面の潜在的な相関に基づく画像の感情分類手法を提案した．提案手法では，各モダリティからの特徴から埋め込み空間への射影を算出し，共通空間での新たな特徴表現を用いて感情分類器を学習した．本文の最後には，提案手法の性能を評価するために，クラウドソーシングを通じて構築した感情ラベル付きデータセットを用いて実験を行った．実験では，同数のモダリティを用いた感情分類器の学習に比べ，提案手法が最もアノテータの評価に近い分類結果を示した．

本論文では，従来研究にならいポジティブとネガティブのニクラス分類のみを検証した．今後は，Plutchik が提唱する感情の輪 [31] に基づく複数クラスの分類へと提案手法を拡張させる予定である．

提案手法で用いる特徴量の設計についても検討を重ねる予定である．特に，文献 [6] のように表情などを表す特徴量や，文字認識なども分類精度向上につながるといえる．また本文の実験では，画像の感情分類における CNN 特徴量の有用性が明らかとなった．今後は，ImageNet で事前学習された CNN をファインチューニングした際の分類性能を検証するとともに，より感情に特化した特徴抽出が可能であるかを議論する必要がある．

画像の感情分類は，情報検索・推薦やデータマイニング，ヒューマンコンピュータインタラクションに有用な技術である．今後は提案手法の応用として，感情に基づく画像検索手法も検討する予定である．

[謝辞]

本研究の一部は，公益財団法人大川情報通信基金研究助成（課題番号：15-12）によって行われた．

[文献]

[1] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 216–223, June 2014.

(a) ポジティブと分類された上位 24 枚の画像.

(b) ネガティブと分類された上位 24 枚の画像.

図 3: Flickr データセットの一回目の試行でテスト用に選出されたクリエイティブ・コモンズ画像に対し，提案手法 LC(V+T+S)+P(V+T+S) によってポジティブまたはネガティブと分類された上位 24 枚の画像．画像のキャプションは Flickr ユーザ ID に対応する．赤枠で囲まれた画像は誤分類を表す.

[2] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Proc. Int. Conf. Systems, Man and Cybernetics (SMC)*, Vol. 4, pp. 3534–3539, Oct 2006.

[3] G. Coppersmith, M. Dredze, and C. Harman. Quantifying mental health signals in Twitter. In *Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 51–60. Association for Computational Linguistics, June 2014.

[4] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. Analyzing and predicting sentiment of images on the social web. In *Proc. Int. Conf. Multimedia (MM)*, pp. 715–718, 2010.

[5] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proc. Int. Conf. Multimedia (MM)*, pp. 223–232, 2013.

[6] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: Image sentiment analysis from a mid-level perspective. In *Proc. Int. Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, pp. 10:1–10:8, 2013.

[7] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proc. Int. AAAI Conf.Artificial Intelligence (AAAI)*, 2015.

[8] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou. Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction. In *Proc. Int. Workshop on Affect & Sentiment in Multimedia (ASM)*, pp. 57–62, 2015.

[9] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proc. Int. Conf. Multimedia (MM)*, pp. 83–92, 2010.

[10] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 902–909, June 2010.

[11] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, pp. 417–422, 2006.

[12] M. Katsurai. Estimating sentiment polarity of web images based on user-generated tags and SentiWordNet.

In *Proc. Int. Workshop on Multimedia Big Data Analytics (MBDA)*, 2014.

[13] D Hardoon, S Szedmak, and J Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, Vol. 16, No. 12, pp. 2639–2664, Dec 2004.

[14] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek. Emotional valence categorization using holistic image features. In *Proc. Int. Conf. Image Processing (ICIP)*, pp. 101–104, Oct 2008.

[15] P. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, University of Florida, Gainesville, 2008.

[16] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, Vol. 106, No. 2, pp. 210–233, 2014.

[17] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 37, No. 11, pp. 2332–2345, Nov 2015.

[18] M. Katsurai, T. Ogawa, and M. Haseyama. A cross-modal approach for extracting semantic relationships between concepts using tagged images. *IEEE Trans. Multimedia*, Vol. 16, No. 4, pp. 1059–1074, June 2014.

[19] H. Hotelling. Relations between two sets of variates. *Biometrika*, Vol. 28, No. 3/4, pp. 321–377, December 1936.

[20] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 1247–1255, 2013.

[21] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2014.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. Int. Conf. Multimedia (MM)*, pp. 675–678, 2014.

[23] S. Kisilevich, C. Rohrdantz, and D. Keim. "Beautiful picture of an ugly place". Exploring photo collections using opinion and sentiment analysis of user comments. In *Proc. Int. Multiconf. Computer Science and Information Technology (IMCSIT)*, pp. 419–428, oct 2010.

[24] R. M. Larsen. Lanczos bidiagonalization with partial reorthogonalization. Technical Report 537, Department of Computer Science, Aarhus University, 1998.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 3111–3119, 2013.

[26] G. A. Miller. WordNet: A lexical database for English. *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, November 1995.

[27] K. Denecke. Using SentiWordNet for multilingual sentiment analysis. In *Proc. Int. Conf. Data Engineering Workshop (ICDEW)*, pp. 507–512, 2008.

[28] M. Katsurai and S. Satoh. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2837–2841, 2016.

[29] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, and J. Tang. How do your friends on social media disclose your emotions? In *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, pp. 306–312, 2014.

[30] H. R. Markus and S. Kitayama. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, Vol. 98, No. 2, pp. 224–253, Apr 1991.

[31] R. Plutchik. The nature of emotions. *American Scientist*, Vol. 89, No. 4, pp. 344–350, 2001.

桂井 麻里衣 Marie KATSURAI
同志社大学理工学部情報システム学科助教．2014 年北海道大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．マルチメディア検索，学術情報マイニング等の研究に従事．

佐藤 真一 Shin'ichi SATOH
国立情報学研究所コンテンツ科学研究系教授．1992 年東京大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．画像理解，画像データベース，映像データベース等の研究に従事．