

Folksonomy による 階層構造画像データベースの構築

秋 間 雄 太^{†1} 柳 井 啓 司^{†1}

近年, Folksonomy の出現により, データベースにタグなどによって意味的な価値を付与することが進められてきたが, 階層構造のような概念間の関係を組み込んでいたデータベースは少ない. そこで, 本研究では, 意味的な階層構造を考慮した画像データベースの作成方法を提案する. 階層構造の構築方法は, 大量の画像データの各概念のノイズを除去した後に, 各概念を視覚特徴を用いたベクトル表現, タグを用いたベクトル表現, 視覚特徴とタグを統合したベクトル表現の3種類のベクトル表現で, JS ダイバージェンスによる距離尺度を用いて概念間の距離関係を推定し, さらに概念エントロピーを作成することで, 概念の広がりから上下関係を推測する. 最終的には, 作成した階層構造を, 視覚的な特徴のみで作成した場合とタグ特徴のみで作成した場合, そしてタグと視覚特徴を結合した場合での表現結果を考察した. 結果として, 視覚特徴での階層構造, タグ情報による階層構造のそれぞれにおいて特有の階層構造を確認することができ, また, 統合した階層構造は両方の階層構造を加味し, それぞれの特徴を内包した新しい階層構造を作り出すことに成功した.

Construction of Hierarchical Structure Image Database Based on Folksonomy

YUUTA AKIMA^{†1} and KEIJI YANAI^{†1}

Recently, Folksonomy attracts attentions as a new method to index large-scale image databases. In the Folksonomy-style image databases, they allow users to attach keywords to images as "tags". Since tag words are uncontrolled, they have various and many kinds of tags associated with images. This is much different from conventional image databases. In this paper, we propose a novel method to extract hierarchical structure on relations between tags from Folksonomy. The tag structure we extract can be used as an ontology for image database search which reflects both textual and visual relations between tags. In the proposed method, at first, we collect millions of tag-attached-images from Flickr which is the world-largest Folksonomy-style image database, and remove noise images from them. Next, we estimate concept vectors for highly-frequent tags based on only visual features, only tag word features and combined features

of both visual and textual features, and compute JS divergence and entropy for three kinds of concept vectors. Finally we estimate hierarchical structures between tags with three kinds of concepts. In the experiments, we show the constructed hierarchical structure, and it includes interesting relations which sometimes are difficult to be discovered by human. This indicates that the proposed method is promising and the structure is expected to help image search and some other applications.

1. はじめに

近年, Web 上に大量の画像が存在するようになり, その用途も多岐に渡るようになった. 一部のデータセットでは Folksonomy という利用者が自由にタグをつけることで対象に意味を付加する分類法を用いて, 意味的な情報が付加された画像データセットを手にいれることもできるようになり, Web 上の画像を用いた画像処理・認識の研究はより一層幅が広がっている.

しかし, Folksonomy を用いた画像データセットでは, 一つ一つの画像に付与されたタグを考慮しているだけで, 画像の視覚特徴やタグ概念間の関係で画像を結び付けているわけではない. 概念には意味があり, それぞれの概念に対して関連を持つ他の概念が存在する. 例えば, 動物にはライオン, トラ, 馬など様々な種類が存在し, それらを包含して動物という概念が存在している. これらの関係を自動的に推定することができれば, データベースに, より高度の情報を付加することができる. 既存のデータベースに自動的に階層構造などの意味的な関係を付与することができれば, データベースの持つ価値は飛躍的に上がり, 概念関係を用いた実験から概念検索の補助など様々な部分で貢献できることが期待される.

また, 画像情報によって自動的に階層構造を構築することで人の認識しえない新しい関係, またはデータベース固有の関係を見出せることが期待される.

そこで, 本研究では, 画像に付加されているタグ情報と視覚特徴を利用して, タグ概念間において, 意味的なあるいは視覚的な分布の広い概念ほど上位の概念であると仮定して上下位関係を推定し, 自動的に階層構造を構築して, 視覚的に表示することを目的とする.

2. 関連研究

階層構造を構築することは大変難しい課題で, 特に画像を用いた研究に関しては一時期に比べて, 減少傾向にある. その中で, 本研究に近い部分を持つ論文をいくつか紹介する.

^{†1} 電気通信大学大学院 電気通信学研究科 情報工学専攻
Department of Computer Science, The University of Electro-Communications

2.1 テキストベースの階層構造

単語辞書データベース 単語辞書データベースとして最も有名なのが WordNet である¹⁾。WordNet は専門家たちによって洗練された概念の意味関係を構築したデータベースで、近年、日本語に対応したデータベースも作成された。また、日本語テキストに限定した階層構造データベースとして日本語語彙大系²⁾が存在する。日本語語彙大系は 30 万語の日本語辞書と 14,000 件の文型パターンを収録した大規模日本語オントロジー辞書である。3,000 種の意味属性を用いて定義されており、最大規模の日本語シソーラスとなっている。

Folksonomy から自動的に作成される階層 単語辞書データベースのように人手で全て作成する方法に対して、Folksonomy を用いて自動的に階層構造を構築しようとする試みが存在する。Tang ら³⁾ はタグがついている映画情報などの Folksonomy に基づいたデータセットを用いて、4つの距離尺度を定義し、階層構造を構築している。彼らの構築した階層構造は概念の意味というより、取り扱っている対象の内容のつながりを示しているといえる。本研究での階層間のつながりもまた彼らの研究と似て、Flickr データベースの概念に含まれる画像の傾向を階層的に表現したものになった。

2.2 画像を用いた階層構造

教師無しによる画像データベースの木構造構築 Bart ら⁴⁾ や Sivic ら⁵⁾ は画像データセットから教師なしで木構造を構築し、まとめるモデルを提案しているが、視覚的な特徴のみによる分類であるため、そこには意味的なものは内在しにくいという点で、意味的な概念間の階層構造を構築することを目的とする本研究とは違いがある。

概念間の距離 概念間の関係によって階層構造を構築するにあたって大切なことは、概念間の距離を求めることである。Wu ら⁶⁾ は JS ダイバージェンスを使った新たな概念間の距離尺度 Flickr Distance を提案しており、JS ダイバージェンスを距離尺度として扱うことに関して共通点があるが、彼らは概念間の距離関係だけを考慮しており、概念間の階層構造（上下関係）を考慮していない点で、本研究と違いが存在する。

WordNet を利用した階層構造 WordNet は画像の分類⁷⁾⁸⁾ などにおいても利用されるが、Deng ら⁹⁾ は WordNet を利用することで階層構造を持った画像データセットの構築を実現した。彼らの作成した階層構造を持つデータベースは、既存の洗練されたデータベース WordNet を基にしているため、高質なデータベースであり、さらに、彼らは Amazon Mechanical Turk を使って、人手で画像がどの概念に属するかを決めているため、さらにデータベースの質は向上している。しかし、WordNet 含めて良くも悪くも人によるものであることで、人間が判断し得ない、見つけられない新しい関係を探し出すことは難しい。データセットによっても傾向があり、唯一の正解というものがありえない以上、WordNet という一つの視点に縛られてしまうことは、そのまま作成されるデータセットに限界を与えることにも繋がる。対して、本手法では、自動的に階層構造を構築していくことで、データセッ

トごとの傾向をうまく組み込む形で階層構造を作成することが可能で、さらに人が気づかないような階層構造を見いだす可能性も秘めている。また、Deng らが利用している WordNet はタイムリーな情報に対応しにくいという欠点があり、本研究では、Folksonomy の誰でも付けられるタグ情報を利用して、タイムリーな情報にも強く、より一般的なデータベースを自動構築することでこの問題を解決することができる。

3. 階層構造画像データベースの概要

本研究では、大量の画像を Flickr から収集し、そのデータを画像に付与されているタグと画像の持つ視覚特徴を分析することで、タグ概念間の関係を推定する。最終的に得られる結果は図 1 のような関係を持つ有向グラフとなる。これらの階層が表現するものは、扱うデータ、本実験では Flickr 画像データベース特有の上位下位関係や共起関係を示している。この階層構造は、人間の直感に当てはまらない部分も存在するかもしれないが、その関係性こそが、対応するデータベース特有の関係か、あるいは我々の知り得ない新しい関係性である可能性であるといえる。この図は概念のつながりが 2 段階までの概念を示しているが、データベースとしては下位概念の存在しない概念までのつながりが得られている。ここで重要な点は music のような視覚的な特徴を持たないような単語に対しても興味深い関係が抽出されているという点である。

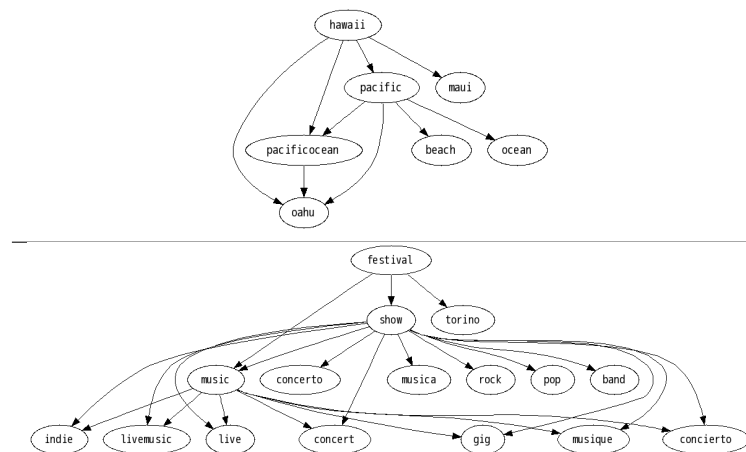


図 1 結果の例

4. 階層構造画像データベースの構築手法

データベースの構築方法の流れは図2に示したように進めた。

まず、Flickrからおおよそ200万枚の画像と画像に付与されたタグ（画像1枚につき平均8個のタグが付与されている）を収集した。次に、全ての画像に対して確率的なクラスタリングを行うpLSA（probabilistic Latent Semantic Analysis, 確率潜在的意味解析）を実行し、各画像の関連性を算出した。抽出した関連性から、各概念らしい画像を算出し、その上位を扱うことで、各概念に対するノイズ画像除去を行った。さらに各概念を表現する画像を抽出した後は、概念ごとに属する画像のpLSAによる確率分布から概念に対する確率分布を算出した。最後に、算出した概念の確率分布を用いて、概念間の距離関係、概念ベクトルから得られたエントロピーによる上下関係から概念間の階層構造を構築した。

4.1 画像表現方法

Bag-of-keypoints Bag-of-keypointsモデルは、相対位置を用いずに画像を局所特徴の集合と捉えた手法である。元々はテキストで用いられた手法で、各文書に現れる単語の集合を特徴としたbag-of-words表現を視覚特徴に適応したものである。局所特徴の特徴ベクトルをベクトル量子化したvisual wordsと呼ばれる特徴ベクトルをまとめたものをcodebookと呼び、それを記述子として画像の特徴ベクトルを生成する。よって、画像はvisual wordの集合として表現される。後の処理のためにも、画像を処理しやすい形にまとめる必要があり、bag-of-keypoints表現を行うことで画像は k 次元（本実験では $k=1000$ ）のベクトル表現にまとめることができる。本実験では、局所特徴としてSIFT特徴量を用いた。

Bag-of-Tags Flickrでは、各画像に複数のタグが存在する。Bag-of-keypointsやbag-of-

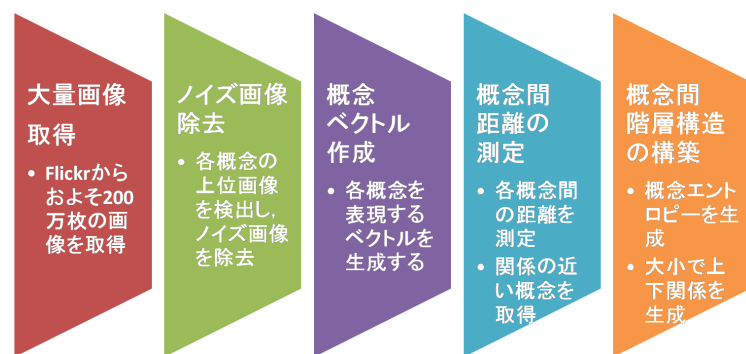


図2 データベース構築フロー

words表現と同様に、ある画像に付与されているタグの出現頻度によるヒストグラムによって表現されるのがbag-of-tags表現である。出現頻度と言っても、各画像に同じタグが2度付与されることは無いので、基本的には0か1の値をとるヒストグラムになる。本実験でのbag-of-tags表現のサイズは4345でおこなった。

結合pLSAベクトル 本研究では、視覚情報とタグ情報を統合した結果も求める。視覚情報はbag-of-keypoints表現、タグ情報はbag-of-tags表現で表現されるが、2つの画像特徴を統合する場合、単純に統合するだけでは次元数や取りうる値の違いが大きいためどちらかの情報が反映されにくくなる可能性がある。そこで本研究では、Lienhartらの手法¹⁰⁾を基に統合を行う。視覚情報のbag-of-keypoints表現とタグ情報のbag-of-tags表現を別々の隠れトピック z の次元の同じpLSAモデルで学習し、得られた各画像 d_i ごとの隠れトピックの確率分布 $P(z_i^{keypoints}|d_i)$, $P(z_i^{tags}|d_i)$ を結合することで得ることができる。隠れトピック z の次元は本研究では100としたので、この統合でできる特徴ベクトルは200次元のベクトルになる。

4.2 概念階層構造表現方法

4.2.1 pLSAとfold-in heuristics

pLSA（probabilistic Latent Semantic Analysis, 確率潜在的意味解析）は文書と単語など、離散2変数の計数データの生成モデルのことである。

文書： $d \in D = \{d_1, \dots, d_N\}$, 単語： $w \in W = \{w_1, \dots, w_M\}$, 隠れトピック： $z \in Z = \{z_1, \dots, z_K\}$ とする。 M サイズの語彙 w からなる単語を含んだ N 個の文書 d を持っていると仮定すると、これは文書を単語の集合とする表現からなるBag-of-Wordsで表現できる。 z は文書 d で単語 w それぞれの発生に関係する話題としての変数である。

文書： $d \in D$ と単語： $w \in W$ の生起は独立と考え、 d_i と w_j の同時確率 $P(d_i, w_j)$ を

$$P(d_i, w_j) = \sum_{k=1}^K P(d_i|z_k)P(w_j|z_k)P(z_k) \quad (1)$$

のように表す。これを画像分類に置き換えると、文書は画像、単語はvisual word、話題はカテゴリと考えることができる。文書と単語の表現方法としてはBag-of-Keypoints表現やBag-of-Tags表現を用いることになる。 $P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$ を反復的手法を用いて計算することにより最適解を推定する。

ただし、大規模なデータセットを一度にpLSAで計算するには、莫大な時間を要することになり、またメモリを大幅に使うことで計算できない事態になることも考えられる。そこで、fold-in heuristics¹¹⁾という手法を用いて、十分大きな一部のデータでpLSAを実験した後に特定のパラメータを固定して、全データでpLSAを計算する手法を利用する。これは、十分大きな一部のデータで求めたパラメータが全体のパラメータと同一であると仮定す

ことで実現され、全データで pLSA を計算する段階ではデータを小分けにして計算することが可能になる。

4.2.2 pLSA による各概念のノイズ画像除去

まず、得られた $P(d|z_k)$ を用いて各トピック z_k に帰属したものが各概念に所属する確率 $P(\text{Concept}|z_k)$ を計算する。ただし、 Concept は概念に属する画像集合（概念と対応するタグが付与されている画像集合）を、 NotConcept は概念に属していない画像集合（概念と対応するタグが付与されていない画像集合）を表している。

$$P_{\text{Concept}}^{z_k} = \sum_{d \in \text{Concept}} P(d|z_k) / |\text{Concept}| \quad (2)$$

$$P_{\text{NotConcept}}^{z_k} = \sum_{d \in \text{NotConcept}} P(d|z_k) / |\text{NotConcept}| \quad (3)$$

$$P(\text{Concept}|z_k) = \frac{P_{\text{Concept}}^{z_k}}{P_{\text{Concept}}^{z_k} + P_{\text{NotConcept}}^{z_k}} \quad (4)$$

$P(\text{Concept}|z_k)$ と $P(z_k|d_i)$ を用いて、画像 d_i の各概念への帰属確率 $P(\text{Concept}|d_i)$ を計算する。

$$P(\text{Concept}|d_i) = \sum_{k=1}^K P(\text{Concept}|z_k) P(z_k|d_i) \quad (5)$$

この値から各概念に属する画像の順位が判明するため、上位何枚と指定することで、ノイズ画像を除去することが可能となる¹²⁾。

4.2.3 概念確率分布の作成

概念を表現する確率分布を作成するために、pLSA を計算することで得られた $P(z|d)$ を用いる。 $P(z_k|d_i)$ は各画像 d_i がどのトピック z_k に帰属するかの確率であり、同様に各概念の画像集合 $d \in \text{Concept}$ が隠れトピック $z \in Z$ に帰属する確率を求めることで、 $P(z|\text{Concept})$ を得ることが可能である。概念 Concept に対する各隠れトピックへの帰属確率 $P(z|\text{Concept})$ の求め方は次式のようにになる。ただし、全画像の集合を D とする。

$$P(z, \text{Concept}) = \left(\sum_{d \in \text{Concept}} p(z|d) \right) / |D| \quad (6)$$

$$P(\text{Concept}) = |\text{Concept}| / |D| \quad (7)$$

$$P(z|\text{Concept}) = P(z, \text{Concept}) / P(\text{Concept}) \quad (8)$$

$$= \left(\sum_{d \in \text{Concept}} p(z|d) \right) / |\text{Concept}| \quad (9)$$

4.2.4 JS ダイバージェンスによる概念間の距離推定

JS ダイバージェンス（Jensen-Shannon divergence）は2つの確率分布間の距離尺度である。JS ダイバージェンスは、対称性がないために距離尺度としては不完全な KL ダイバージェンス（Kullback-Leibler divergence）に対称性を与えた尺度となっている。2つの確率分布間の類似性が高いほど値は小さくなる。2つの確率分布 P, Q に対してそれぞれの算出方法は次式のようにになる。

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (10)$$

$$D_{JS}(P||Q) = \frac{D_{KL}(P||(P/2 + Q/2))}{2} + \frac{D_{KL}(Q||(Q/2 + P/2))}{2} \quad (11)$$

4.2.5 エントロピーによる概念間の上位下位関係

各概念の確率分布 P に対するエントロピー $H(P)$ は次式で得ることができる。

$$H(P) = - \sum_{z \in Z} P(z|\text{Concept}) \log(P(z|\text{Concept})) \quad (12)$$

このとき、エントロピーの値は様々な隠れトピックに分類される可能性のある概念ほど大きな値になり、逆にある一定の隠れトピックにしか属さないような概念ほど小さな値になる。上位概念であるほど様々な画像、特徴を持つことが考えられるために、結果的にエントロピーが大きくなることが期待される。逆に下位概念はある一定の特徴に特化することが考えられるために、結果的にエントロピーが小さくなることが期待される。このエントロピーの大小を利用して、概念間の階層構造を構築する。

4.2.6 階層構造の構築方法

算出した概念距離とエントロピーからの階層構造の構築方法として、Heymann らの手法¹³⁾ や江田らの手法¹⁴⁾ が参考になるが、Heymann らの扱う拡張グリーディ法はノードを必ずある一つの親の子供として配置しなければならなくなり、関係に無理が生じることがある。そこで、本研究では江田らの非巡回有向グラフ（Directed Acyclic Graph, DAG）を用いた手法を参考に実験を行った。

DAG は閉路を持たないことで、上位概念から下位概念までの道筋を容易に辿ることができる。また、Heymann らが扱っている拡張グリーディ法は一度全体として構築しなければならないのに対して、DAG は動的に部分構造を展開できる点で拡張が簡単である。

JS ダイバージェンスで概念間の距離が求まり、各概念のエントロピーが求まっているとしたとき、DAG による概念間の階層関係は以下の順序にそって実行される。

(1) 全ての概念それぞれに対して

(a) 距離の近い概念上位 k 個、概念間距離が $threshold$ 以下の概念を取り出す

表 1 実験画像データセット情報

全画像枚数	1,991,349 枚
全オーナー数	95,140 人
全タグ種類数	872,597 個
取得不可能画像数	9,536 枚
タグの付与なし画像数	343,219 枚

(b) 取り出した概念と現在注目している概念のエントロピーを比較し、大きい方を親ノード、小さい方を子ノードと設定する

以上を実行することで、各概念に階層構造が生じる。

本研究では、 $k = 20$ 、タグ情報でのしきい値は $threshold = 0.15$ 、視覚情報でのエントロピーは $threshold = 0.05$ 、統合情報でのエントロピーは $threshold = 0.1$ とした。実際のグラフ構築には Graphviz¹⁵⁾ を利用した。

5. 実験

実験では、ノイズ画像除去を行った後に、視覚特徴のみ、タグ特徴のみ、両方を統合した特徴の 3 種類の方法で階層構造を構築し、比較を行った。

5.1 実験データセット

実験用の画像データセットは Flickr から取得した。最終的に、1,991,349 枚のおよそ 200 万枚の画像を Flickr API を用いて取得した。Flickr API に用意されているメソッドを用いて、約 200 万枚分の画像情報を取得した後に、その情報をもとに画像を取得した。ただし、同じ ID を持つ画像情報の取得は行わないように設定した。収集したデータセットの情報は表 1 のようになった。画像取得において情報は取得できたが画像が取得できないものが存在するが、この要因は Flickr の一般ユーザー一人あたりの所持上限が決まっていることで情報取得後に上限を越えたユーザの画像がアクセスできなくなっていることが考えられる。

また、図 3 はそれぞれのタグ概念に所属する画像数ごとのタグの種類数を両対数グラフで表したものである。この図からは、大部分のタグが 100 枚以下の画像でしか用いられていないことがわかる。

図 4 は利用オーナー数の上限ごとの利用タグの種類数を表現している。利用オーナー数が多いということは一般的なタグ概念であり、利用オーナー数が少ないことはオリジナルなタグかあるいはかなり専門的な内容のタグであることが考えられる。図から、利用オーナー数の多い一般的なタグとなるような概念の数は少なく 5000 個程度である。一方、利用者数の少ないような固有なタグは数が多いということがわかった。

図 4 の結果から、タグでの実験を行うのであればおよそ 5000 個の上位タグを用いれば、固有なタグの影響を受けにくい実験になることがわかった。

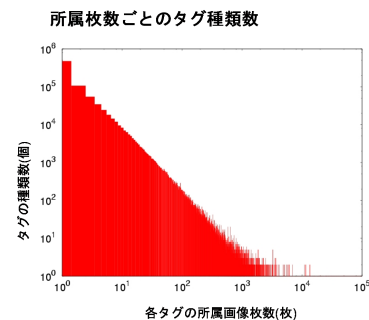


図 3 所属枚数ごとのタグ種類数

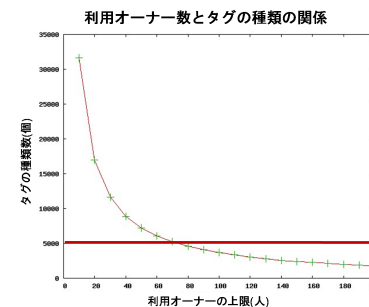


図 4 利用オーナー数の上限ごとの利用タグ種類数

5.2 実験の手順

階層構造構築の手順は以下のようになった。

- (1) 取得不可能画像、タグの付与なし画像を除いたおよそ 150 万枚の画像に対して画像ベクトル表現を行う
- (2) ベクトル表現された画像群における pLSA (fold-in heuristics) モデルを算出
- (3) ここで、オーナーが 1 人 1 タグに最大画像 20 枚と設定して、画像をふるい落とし、タグ数は図 4 での結果を反映させるためにオーナー 1 人 1 タグに最大 10 枚としたときに 50 枚以上画像が存在する 4345 個で以降の実験を行う（ただし、画像のタグ表現はあらかじめこの 4345 という値でヒストグラム表現をおこなう）
- (4) 制限された画像の中で、さらに各概念らしい画像でソートして上位 n 枚で実験を行う ($n = 5 \times \sqrt{\text{各概念の取得枚数}}$)、画像枚数の少ない概念でもおよそ 100 程度になるような設定を選んだ)
- (5) 上位 100 枚で各概念のエントロピーを求め、さらに JS ダイバージェンスを求める
- (6) 求めたエントロピーと概念距離から DAG を生成 (DAG の生成段階では各概念で 200 枚より少ない概念は除去)

以上の処理を、視覚特徴のみ、タグ特徴のみ、視覚特徴とタグ特徴を統合した特徴の 3 種類の画像表現に対して実行した。まずノイズ画像除去結果を示し、その後、各特徴表現における各概念階層構造の構築結果を比較して考察をおこなう。

5.3 ノイズ画像除去結果

ノイズ除去の結果を定量的に判断するために、dog, bee, sheep, coast, waterfall, singer, piano, blue, dark, lightning, sunset, bw, fire, airplane, soccer, luna, icecream の 17 種類において、ノイズ除去を行っていない画像群から対応するタグの付与された画像を

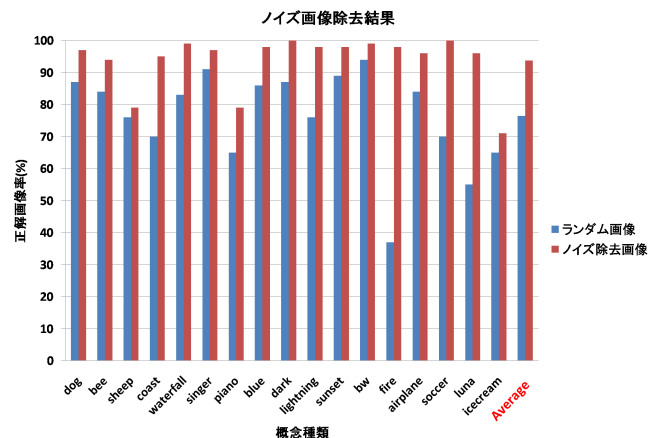


図5 ノイズ除去の精度評価

ランダムに 100 枚取得してきたものと、ノイズ除去で各タグに所属する可能性の高い上位 100 枚の画像で各タグでの正解画像の割合を調べた。基本的には、そのタグに最もふさわしいであろう対象を含む画像を正解とすることにしたが、全体が写っていないなどはある程度まで許容した。

結果は図5のようになった。平均してノイズ除去なしのランダム画像群では正解率 76.4%で、ノイズ除去ありの画像群では正解率 93.8% となった。もともとそのタグが付与されている画像なので、完全に違う画像というものはそれほど多くはない。しかし、図5の結果だけでは示せないが、ノイズ除去した画像セットは比較してかなり良質なものになっている。ノイズ除去後の結果が悪い概念は、元々の画像データセットの枚数が少なかったり、質が悪いことが要因で起こっているようで、基本的にノイズ除去後の結果が悪いデータセットは、ランダムで抽出したときの結果もかなり悪いことがわかった。

5.4 概念階層構造の構築結果

まず、視覚特徴による階層構造とタグ特徴による階層構造の結果を比較しながら、それぞれの特徴での階層構造の傾向を示す。その後、視覚特徴とタグ特徴を統合した階層構造の結果を示し、視覚特徴の利点とタグ特徴の利点を兼ね備えた階層構造になっていることを示す。

5.4.1 視覚特徴による階層構造とタグ特徴による階層構造

Cute の下位構造の結果が視覚特徴による結果が図 6 となり，タグ特徴による結果が図 7 となった．矢印の刺されている方が下位概念になっている．

まず先にタグ特徴による結果を見てみると、cute と聞いて思いつくような単語が多く、特

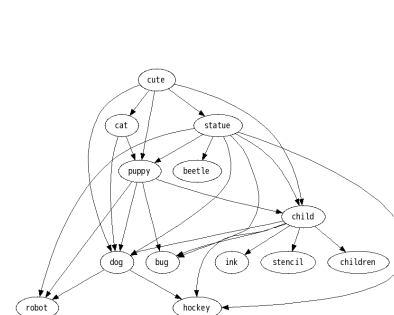


図 6 視覚特徴で作成した cute の階層構造

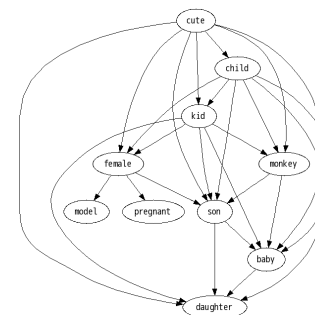


図 7 タグ特徴で作成した cute の階層構造

に female, kid, daughter など人関連の概念が多い。

次に、視覚による結果を見てみると、タグ特徴とは違った関係性が見て取れる。実際の cute の画像例である図 8 を見てみると犬、猫、子供などの画像が見て取れた。この画像例をふまえて視覚特徴による階層構造を見てみると、cute という単語がふさわしい子犬という意味の puppy が結び付いている。puppy の画像例である図 9 では、犬の画像という点で cute の画像例と共通している部分が存在した。この puppy は cute とのタグ特徴の関係ではあまり上位には来ない。つまり、cute というタグがつけられていないが、視覚的に cute である概念が検出されている。加えて、cat, dog などの動物を表す概念が多いのを見て取れる。これは cute の画像例で示した内容とも一致している。また、彫像を表す statue も cute と結び付いている。statue と cute は感覚的にはそれほど近い概念ではないかもしれない。しかし、人間が気づかないような cute の要素が statue に見て取れる可能性があり、少なくとも Flickr で集めたデータセットにおいて statue は視覚的に cute であると判断されたことになる。statue での画像例である図 10 を見てみると、子供の彫像も存在し、その部分に cute とのつながりがあるのかもしれない。このように、人間が見て取れる要素に加えて、人間が判断できない新しい関係性を見つけ出せる可能性を秘めていることがわかった。

これらの結果から、視覚特徴による階層構造の傾向は必ずしも人間が感覚的に判断するような関係性を持ってはいないが、その分だけ人間が判断し得ないような新しい関係性または利用するデータセット特有の関係性を内包している可能性をもっていることがわかった。また、タグ特徴による階層構造の傾向は人間が自主的に付与した意味を用いているために、基本的には概念間の距離関係や上下関係は、人間の感覚に近いものが得られることがわかった。

5.4.2 統合特徴による階層構造の利点

視覚特徴による階層構造とタグ特徴による階層構造、そして、統合特徴による階層構造を



図 8 cute の画像例



図 9 puppy の画像例



図 10 statue の画像例

比較することで、視覚特徴による階層構造とタグ特徴による階層構造の欠点と統合特徴による階層構造の利点を以下の実験結果から示す。

Mountain の下位階層構造を見てみると、視覚特徴による階層構造は図 11、タグ特徴による階層構造は図 12、統合特徴による階層構造は図 13 になった。

統合特徴による mountain の階層構造を見てみると、alps, alpen, berge (ドイツ語の mountains) といった視覚特徴による階層構造やタグ特徴による階層構造に表れず、mountain と関係がある概念が見て取れた。

ここで alps という概念が視覚特徴による階層構造とタグ特徴による階層構造のそれぞれに表れない理由を調べてみた。視覚特徴の階層構造では、mountain を表す画像例である図 14 と alps を表す画像例である図 15 の画像が山の写っている似たような画像であることから見て取れるように、mountain と alps の関係は深い。alps が上位階層と認識されてしまっていた。このように、視覚特徴による階層構造は良くも悪くもデータセットの視覚的な傾向を大きく反映してしまうために、各概念のデータセットの傾向によってはエントロピーによる上下関係が人間の感覚とずれてしまうことがある。一方、タグ特徴による階層構造に alps が表れない理由としては、単純に mountain と alps との関係が遠くなっていることが原因となっていた。タグ特徴による階層構造では視覚特徴による階層構造のように上位下位関係が崩れることは少ないが、視覚的な類似性が考慮されていないために、当然の結果として、視覚的な類似性よりも単純な関係性が優先される。普通に扱う分には問題ないのかもしれないが、画像検索などへの応用を考える場合には、視覚的な類似性は必要不可欠である。

統合特徴による階層構造では、視覚特徴による階層構造の利点とタグ特徴による階層構造の利点が組み合わせられることで、視覚特徴による階層構造の上下関係のずれとタグ特徴による階層構造の視覚的な弱さといった欠点を克服することが可能であり、より実用的な階層構造となっていることがわかった。

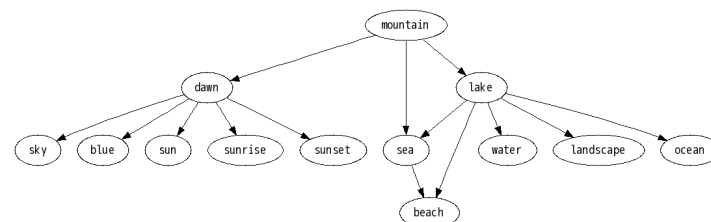


図 11 視覚特徴で作成した mountain の階層構造

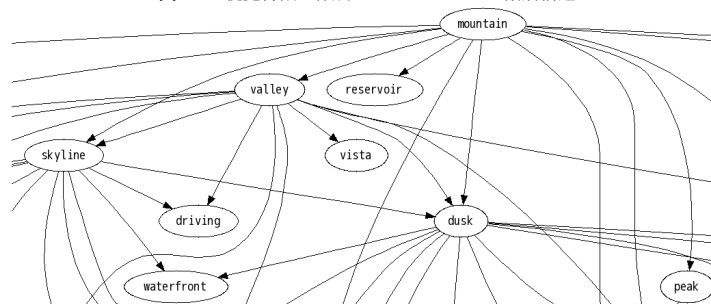


図 12 タグ特徴で作成した mountain の階層構造 (一部抜粋)

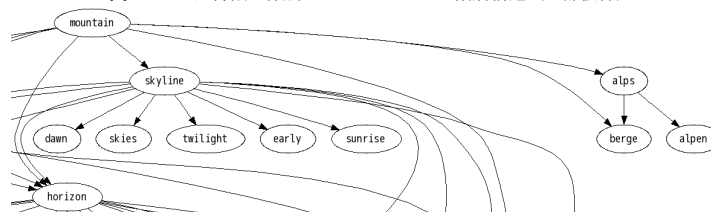


図 13 統合特徴で作成した mountain の階層構造 (一部抜粋)

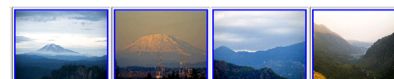


図 14 mountain の画像例



図 15 alps の画像例

6. おわりに

本研究では、大規模な画像データベースを用いて概念間の階層構造を自動的に構築した。構築する階層構造は視覚特徴、タグ特徴、視覚特徴とタグ特徴の統合特徴の三種類で作成し、それぞれの結果を比較した。

概念間の関係を階層を用いた視覚的なアプローチで表現できるようになったことで、概念一つ一つを見るよりも相対的な判断ができるようになり、データセットに対する考察が大幅容易になった。

データベース特有あるいは人間の知り得ない新しい関係を抽出したい場合には人の判断の入らない視覚特徴のみの階層構造を用いるのが有効であり、人の認識と視覚的な関係を合わせた階層構造を構築するためには統合特徴による階層構造を用いることが有効であることがわかった。

さらに、今回生成した階層構造から画像検索などに利用するために必要な人の判断に近づける場合の課題を発見することができた。

以上から、本研究では大規模画像データセットから自動的に階層構造を作り出すことで、データベース特有の関係性を抽出できるような階層構造の作成が可能になり、そのことで、データベースに対する考察のアプローチが容易になり、そして、人の判断しえないような関係を抽出できる可能性や画像検索などのアプリケーションに組み込むための課題を見つけることができた。

7. 今後の課題

本研究で作成した階層構造をより実用的なものにしていくためには、いくつかの解決しなければならない問題が存在する。ここでは、それらの問題点と解決案を示す。

視覚特徴による階層構造において、本実験では SIFT 特徴量のみを用いているために、形状のみの判断しか行われていない。そのため、例えば色に関して言えば、色の区別が無いために形状的な近さで関係が遠いものでも近くなってしまうたり、概念の大きさなどが色の種類などに依存するような概念は上下関係が不安定になりやすい傾向がみてとれた。そのため、より概念の特徴や広さを忠実に表現するためにも、より多様な特徴を用いて階層構造を作成する必要があると考えられる。

また、pLSA によるノイズ画像除去において、それぞれの概念らしい上位画像を取ってことで、確かにノイズとなる画像は除去できているが、 $P(\text{Concept}|z)$ の偏り方によっては、類似した画像ばかりが上位に来ることになり、概念のばらつきが制限されてしまう可能性がある。解決策としては $P(\text{Concept}|z)$ の上限値を決めておくことで、単独の特徴に重みが行かないようにすることなどが考えられる。

データセットにも問題点があり、本実験で用いた 200 万枚のデータセットは、Flickr から日付を用いて取得してきたデータセットであり、タグについては指定していない。そのため、多種多様なタグが付いた画像が取得できるが、実験データセットで画像枚数の少ない概念も存在する。もちろん、ある程度の数以上の概念のデータセットを扱っているが、概念の広さを表すには不適當な枚数のデータセットもあるかもしれない。そのため、今後の実験で

は、各概念ごとに画像を取得することで、きちんとした枚数の概念データセットを作成していく必要があると思われる。

本研究では、作成した階層構造の数値的な評価を行うことができていない。階層構造自体を評価することは難しいため、今後、作成した階層構造を画像検索の補助などに生かすことができれば、数値的な評価も可能になると考えられる。

参 考 文 献

- 1) Fellbaum, C.(ed.): *WordNet: An Electronic Lexical Database*, The MIT Press (2000).
- 2) 池原 悟, 宮崎正弘, 白井 論, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林 良彦: 日本語彙大系 CD-ROM 版, 岩波書店 (1999).
- 3) Tang, J., Leung, H., Luo, Q., Chen, D. and Gong, J.: Towards Ontology Learning from Folksonomies, *Proc. of International Joint Conferences on Artificial Intelligence* (2009).
- 4) Bart, E., Porteous, I., Perona, P. and Welling, M.: Unsupervised Learning of Visual Taxonomies, *Proc. of IEEE Computer Vision and Pattern Recognition* (2008).
- 5) Sivic, J., Russell, B., Zisserman, A., Freeman, W. and Efros, A.: Unsupervised Discovery of Visual Object Class Hierarchies, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp.1-8 (2008).
- 6) Wu, L., Hua, X.S., Yu, N., Ma, W.Y. and Li, S.: Flickr Distance, *Proc. of ACM International Conference Multimedia* (2008).
- 7) Fan, J., Gao, Y. and Luo, H.: Hierarchical Classification for Automatic Image Annotation, *Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval* (2007).
- 8) Marszalek, M. and Schmid, C.: Semantic Hierarchies for Visual Object Recognition, *Proc. of IEEE Computer Vision and Pattern Recognition* (2007).
- 9) Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Fei-Fei, L.: ImageNet: A Large-scale Hierarchical Image Database, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp.710-719 (2009).
- 10) Lienhart, R., Romberg, S. and Horster, E.: Multilayer pLSA for Multimodal Image Retrieval, *Proc. of ACM International Conference on Image and Video Retrieval* (2009).
- 11) Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol.43, pp.177-196 (2001).
- 12) Monay, F. and Gatica-Perez, D.: Modeling Semantic Aspects for Cross-media Image Retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.29, No.10, pp.1802-1817 (2007).
- 13) Heymann, P. and Garcia-Molina, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems, Technical report, Stanford, <http://dbpubs.stanford.edu:8090/pub/2006-10> (2006).
- 14) 江田毅晴, 吉川正俊, 山室雅司: 非巡回有向グラフによるフォークソノミータグの局所拡張可能な配置方法, 電子情報通信学会 第 19 回データ工学ワークショップ論文集 (2008).
- 15) Graphviz: <http://www.graphviz.org/>.