Vol.2012-CVIM-181 No.8
2012/3/15

大規模 Web 画像データベースを用いた画像アノテーションシステムの構築

渡邉裕樹[†1]　秋良直人[†1]　廣池敦[†1]
松原大輔[†1]　平松義崇[†1]　永吉洋登[†1]
影広達彦[†1]　久光徹[†1]

本報告では，大規模 Web 画像データベースと類似画像検索技術を用いた画像アノテーションシステムについて述べる．本システムは，与えられた画像をクエリとして類似画像検索を行い，検索結果の画像に付随するテキスト中の単語を確率的指標により評価することで，特別な事前学習なしに画像を意味付けるキーワードを推定可能である．本システムを用いて 5 カテゴリ 30 概念の画像に対するアノテーションを行った結果，10 位内正解率がカテゴリ平均で 43〜75%，全概念の平均で約 59.1%であった．また，処理速度は，データベース構築時の画像とテキストの解析に 1 画像あたり 90ms，画像アノテーション処理に 643ms であった．

Development of Image Annotation System based on Large-scale Web Image Database

Yuki Watanabe,[†1] Naoto Akira,[†1]
Atsushi Hiroike,[†1] Daisuke Matsubara,[†1]
Yoshitaka Hiramatsu,[†1] Hiroto Nagayoshi,[†1]
Tatsuhiko Kagehiro[†1] and Toru Hisamitsu [†1]

In this paper, we propose an image annotation system based on large-scale Web image database and similar-image search engine. Given an unknown image, the system finds similar-images in the database, and obtains a set of text data associated with them. Then the system evaluates the representativeness of words in the text by using probabilistic measures, and outputs keywords for the image. The estimation accuracy of 5 category (30 concepts) images is between 43% and 75%, 59.1% on average. The processing time for database construction (image and text analysis) is 90 ms/image. The processing time for image annotation is 643 ms/image.

1. はじめに

制約のない実世界の画像中の物体やシーンを，計算機に認識させ，画像中の位置や物体名称などを一般的な表現で記述させる技術は一般物体認識と呼ばれ，画像認識の研究において最も困難な課題の一つとされている[1)]．一般物体認識の要素課題である画像アノテーションは，画像が表す内容に対応するメタデータを自動的に付与する技術であり，近年活発に研究がなされている[2)–6)]．

画像アノテーションの先駆的な研究として森らは，百科事典中の画像と説明文から，画像の部分領域と単語の対応を学習することで，未知の画像から関連単語を出力させる方法を 2001 年に提案している[2)]．しかし，当時の環境では，膨大な認識対象に対してデータを十分用意出来なかったため，認識精度は限定的なものであった．

近年では，インターネットの急速な発展を背景として，Web 上のデータを用いた画像認識の研究が発展している．例えば，ImageNet[4)] は，WordNet の階層構造を利用して，概念に属する画像を人手で収集したオントロジであり，2012 年 2 月の時点で 21,841 の概念，14,197,122 枚の画像が利用可能である． ImageNet は，誤分類の少ない高品質なデータであるため，機械学習を用いたカテゴリ分類に活用されている．

これに対して，ノイズを含む低品質なデータを大量に集め，それらを直接的に使用することで学習なしの画像認識を行う，事例ベースの手法が提案されている．例えば，Torralba らは，Web 検索エンジンを用いて 75,062 カテゴリ，約 8,000 万枚の画像を収集した TinyImages と呼ばれるデータベースを用い，単純な画像特徴量による k 近傍探索を行うことにより，画像認識が可能であることを示した[3)]．同様に，Wang らは ARISTA プロジェクトにおいて，20 億枚の画像データベースから準同一画像（Near Duplicate Image）を探索することで，製品名や著名人などのタグを推定可能であることを示した[5)]．

機械学習を用いたアプローチ，事例ベースのアプローチのいずれにおいても，大規模なデータを活用できるか否かが，近年の画像認識研究の発展の鍵となっている．

一方で，筆者らは，2008 年 9 月に大規模 Web 画像データベースを対象とした類似画像検索サービス「GazoPa」を立ち上げ，2011 年 6 月までの約 3 年間，実証実験を行なって

†1 株式会社 日立製作所 中央研究所
The Central Research Laboratory, Hitachi Ltd.

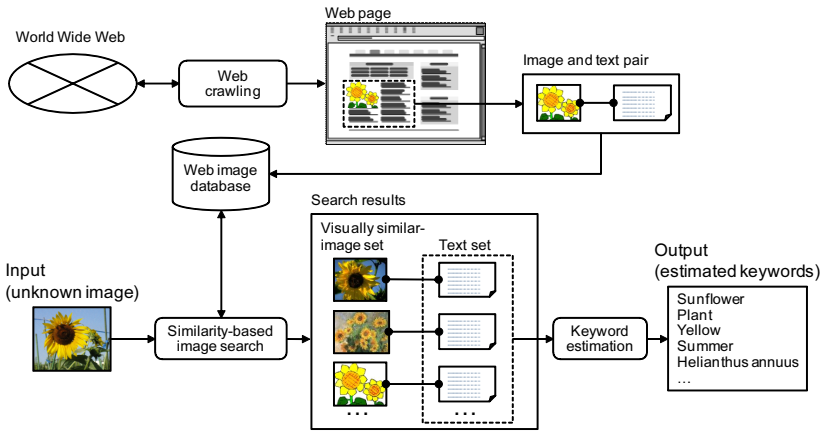

図 1 Web 画像データベースを用いた画像アノテーションシステム
Fig. 1 Image annotation system using Web image database.

きた[7]．期間中に，断続的に自動クローリングを実施し，現在約 1 億件の画像と画像周辺テキストがデータベース化されている．Web スケールの大規模データを扱うことにより，これまでに，画像特徴量の改良，データベースの並列管理や高速類似ベクトル検索などの技術開発を行なってきた[8]．これらの技術は，類似画像検索プラットフォーム EnraEnra^{TM} として製品化されている[9]．

本報告では，Web 画像データベースと類似画像検索を用いた画像アノテーションを提案する．提案手法は，入力画像をクエリとして類似画像検索を行い，検索結果に付随するテキスト中の単語を確率的指標により評価することで，画像を特徴付けるキーワードを推定する．また本研究では，「GazoPa」で収集した約 1 億件の Web 画像データベースと，EnraEnra を用いた画像アノテーションシステムを構築し，提案手法の有効性を示す．

2. Web 画像データベースを用いた画像アノテーション

2.1 提案手法の概要

Web 画像データベースを用いた画像アノテーションシステムを図 1 に示す．本システムにおける処理は，クローリングによってデータベースを構築する前処理と，データベースを用いて画像認識する処理に分けられる．

前処理においては，Web クローリングによって自動的に取得した Web ページから，画像とその周辺テキストを抽出し，それらを関連付けて画像データベースに保存しておく．周辺テキストとしては，例えば，ページのタイトルや，<img>タグの alt 属性，<img>タグの前後のテキストなどが利用できる．このようにして機械的に抽出されたテキストは，必ずしも画像を説明するものではないが，関連する単語が含まれる可能性は高い．

認識処理においては，まず，ユーザから入力された未知の画像をクエリとして類似画像検索を行う．類似画像検索は，画像そのものが持つ色や形状などの特徴による検索であり，この結果，入力画像と「見た目」の類似した画像が得られる．また，画像データベースには，画像とテキストが関連付けて保存されているため，検索結果からテキストの集合が得られる．次に，得られたテキスト集合を 1 つの文書としてみなし，この文書を特徴付ける重要語を抽出する．重要語の抽出では，文書に含まれるすべての単語に対して，後述する重み付け指標によるスコアでソーティングし，その上位またはスコアが閾値以上の単語を出力する．

以上の機能が実現されると，未知の画像に対して自動的にタグを付与し，データ解析や検索に利用したり，ユーザが詳細なタグ付作業を行う際の補助ツールとして利用したりできる．本手法は，事前に認識対象を定義する必要がないため，膨大な概念が認識対象となる一般物体認識における学習コストの問題を解消することができる．また，継続したクローリングにより，時代の流れと共に生まれる新たな概念の画像に自動的に対応することができる．

一方で，Web ページのテキストは画像の説明文としては S/N が非常に低いため，十分なデータ量が集まらないと認識精度が上がらないという問題がある．また，検索対象が大規模であるため，高速かつスケーラブルな類似画像検索プラットフォームが必須である．

筆者らは，Web 画像検索サービス「GazoPa」で収集した約 1 億件の Web 画像データベースと，EnraEnra を用いて，提案手法の基づく画像アノテーションシステムを構築した．本システムは，一般的な PC サーバ（CPU 2.40 GHz，メモリ 4GB）を 13 台使用しており，各サーバに約 400 万件の画像データを管理する DB サーバプロセスが 2 つずつ存在する．それらに対して並列に検索処理を行うことで大規模類似画像検索を実現している．

以下では，類似画像検索の概要と，テキストからの重要語抽出について詳細に説明する．

2.2 類似画像検索

類似画像検索は，与えられた画像と「見た目」の類似した画像を引き出す処理である．

類似画像検索においては，まず，与えられた画像の「見た目」の特徴を表す数値データを抽出する．このような数値データを「特徴量」と呼び，通常は固定長のベクトルデータとして扱う．類似画像検索の結果は，特徴量に大きく左右されるが，その作成方法に王道はなく，統一的な抽出手法を実現することは不可能であるといわれている．特に，一般画像を対

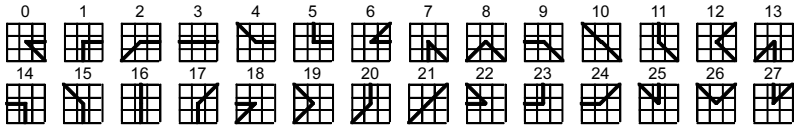

図 2 局所エッジパターン特徴量
Fig. 2 Local edge pattern features.

象とした類似画像検索においては，現在ではまだ研究者の経験と直感による所が大きい．

本システムでは，Web 画像検索サービス「GazoPa」の経験を通して作成された，以下の 4 種類の特徴量を用いる．特徴量の次元数は検索速度に影響を与えるため，PCA（Principal Component Analysis）によって数 100 次元程度まで圧縮して用いている．

(1) 色特徴量
RGB カラーヒストグラム特徴量．RGB 色空間を分割し，入力画像の画素がどの領域に属するかを判定し，ヒストグラムを生成する．

(2) 形状特徴量
画素周辺のエッジパターンのヒストグラム特徴量．グレースケール画像の各画素に対して，周辺画素の輝度勾配から，図 2 に示される 28 種類のエッジパターンとの一致度を求める．入力画像を格子状に構図分割し，各領域に含まれるエッジパターンの一致度を集計し，28 ビンのヒストグラムを領域数だけ計算し特徴量とする．また，解像度に応じてパターンが変化するため，複数の解像度の画像から抽出した特徴量を結合して使用する．さらに，パターン間の統計的な偏りを減らすための正規化を行う．

(3) 色＋形状特徴量 1
色特徴量と形状特徴量を組み合わせた特徴量．色に関しては，(1) のように構図分割を行わない画像全体のヒストグラムに加えて，構図分割で領域ごとに集計したヒストグラムを用いる．形状に関しては，(2) と同様の特徴量を用いる．

(4) 色＋形状特徴量 2
色特徴量，形状特徴量，および縦横比を組み合わせた特徴量．(3) と異なり，構図分割された特徴量のみで構成され，より細部に着目した検索結果になる．

画像データベースには，各画像から計算された特徴量が保存される．類似画像検索では，入力画像に対して，同様の方法で特徴量を計算し，データベースに保存されている各画像の特徴量とのベクトル間距離を求め，距離の小さい順にソーティングして出力することで，特徴の類似した画像が得られる．

基本的には，データベース中の全ての画像に対して距離を求めればよいが，データベースの規模が大きくなると，距離計算に時間がかかってしまう．特に，高次元の特徴量を扱う場合は，計算機のメモリ上に全データを置くことが難しく，HDD（Hard Disk Drive）へのデータアクセスが頻繁に生じてしまうため，処理時間は膨大になる．

そこで，本システムにおいては，クラスタリングを用いた高速類似ベクトル検索技術[8] を導入している．本手法では，データベースの構築時に類似する画像の集合（クラスタ）をひとまとめに保存しておく．計算機のメモリにはクラスタの平均ベクトルのみを保存しておき，画像特徴量は HDD 上に記憶するため，大規模なデータベースも一般的な PC サーバ上で動かすことができる．

検索を行う際には，まず，クラスタの平均ベクトルとの距離計算を行い，類似クラスタを見つける．次に，類似クラスタに属する画像特徴量を HDD から読み出し，距離計算，ソーティングを行い，出力する．このような 2 段階の探索によって，データ規模が増加した際の距離計算やディスクアクセスの増加を抑えることができる．

2.3 重要語抽出

重要語抽出は，与えられたテキストから，代表的な単語を特定する処理である．

テキスト解析，情報検索の研究分野においては，単語の重要性を表す指標として以下のものがよく用いられている．

- **TF (Term Frequency)**
  ある文書 d 中に出現する単語 t の頻度であり，$tf(t,d)$ と表す．画像アノテーションにおいては，検索の結果得られたテキスト集合をひとつの文書 d とみなし，頻度を数える．TF は，「何度も繰り返し言及される概念は重要な概念である」という仮定のもとに使われる指標であるが，一般にあまりに頻度の高い単語は文書を特徴付ける上では役に立たない場合が多いため，なんらかの制約を加えることが望ましい．
- **DF (Document Frequency)**
  全文書集合中で単語 t の出現する文書頻度であり，$df(t)$ と表す．DF が大きい単語は，どの文書にも一様に現れるため特定の文書を特徴付けない．
- **IDF (Inverse Document Frequency)**
  逆文書頻度 $idf(t)$ は，単語頻度の重み付けに用いられる指標であり，N を文書総数とすると，以下の式で定義される．

$$idf(t) = \log \frac{N}{df(t)} \tag{1}$$

IDF は，ある単語が少数の文書にしか出現しないとき大きくなり，どの文書にも出現すると最小になる．

- **TF-IDF**

  TF と IDF を組みわせた指標であり，以下の式で表される．

  $$tf \cdot idf(t,d) = tf(t,d) \times idf(t) \tag{2}$$

  TF-IDF は，計算が容易であり，よく用いられる指標であるが，高頻度語を過大評価する傾向が知られている．

以上の指標は，テキスト解析の研究分野において，ヒューリスティックに定義された指標であり，一般的な文書に対しては有用である．一方で，本研究で扱うテキストは，Web ページの画像周辺の断片的なテキストであるため，書籍やニュース記事などの通常文書の解析を想定して考案された指標は，適切ではない可能性がある．そこで本研究においては，以下に示すとおり，数学的により健全な，確率的指標を導入した．

- **KL (Kullback-Leibler divergence)**

  2 つの確率分布 P_i と Q_i の差異を計る尺度であり，以下の式で表される．

  $$kl = \sum_i P_i \log \frac{P_i}{Q_i} \tag{3}$$

  単語 t を評価する際には，部分文書集合から無作為に文書を取り出したときに単語 t を含む文書である確率を $p(t)$，全文書集合から無作為に文書を取り出したときに単語 t を含む文書である確率を $q(t)$ とすれば，以下の式を用いることができる．

  $$kl(t) = p(t) \log \frac{p(t)}{q(t)} + (1-p(t)) \log \frac{1-p(t)}{1-q(t)} \tag{4}$$

  画像アノテーションにおいて，全画像数を N，検索によって得られた画像数を M，検索結果に付随する文書集合における単語 t を含む文書数を $df'(t)$ とすれば，$p(t)$，$q(t)$ は，以下の式で与えられる．

  $$p(t) = \frac{df'(t)}{M} \tag{5}$$

  $$q(t) = \frac{df(t)}{N} \tag{6}$$

- **HGS**[10)]

  HGS は，超幾何分布（hypergeometric distribution）に基づく確率的指標であり，以下の式で表される．

  $$hgs(W,V,w,v) = \sum_{u \geq v} hg(W,V,w,u) \tag{7}$$

  $$hg(W,V,w,u) = \frac{{}_V C_u \times {}_{(W-V)} C_{(w-u)}}{{}_W C_w} \tag{8}$$

  $hgs(W,V,w,v)$ は，非復元抽出を厳密に表現した関数であり，「W 個の玉の中に V 個の赤い玉があるときに，任意に取り出した w 個の玉の中に赤い玉が v 個以上含まれる確率」を意味し，$hg(W,V,w,u)$ は，「W 個の玉の中に V 個の赤い玉があるときに，任意に取り出した w 個の玉の中に赤い玉がちょうど u 個含まれる確率」を意味する．

  画像アノテーションにおいては，W は全画像の文書に含まれる重複を許可した全単語数，V は全文書中での単語 t の頻度，w は検索結果の文書に含まれる重複を許可した全単語数，v は検索結果の文書中の単語 t の頻度，として考える．

画像アノテーションを効率的に行うために，本システムでは，取得したテキストに対して形態素解析をかけて，単語単位に分解し，単語の ID 列としてデータベースに保存しておく．

3. 性能評価実験

3.1 評 価 方 法

以下では，開発した画像アノテーションシステムのキーワード推定精度と処理速度を評価する．評価に用いたテストセットは，「GazoPa」のデータベースから報告者が抽出した画像であり，現時点で，6 種類のカテゴリ，30 種類の概念に関して，各 100 枚の画像，計 3,000 枚が集められている．以下に，評価セットに含まれるカテゴリと単語（概念）を示す．括弧内は，各単語を含むテキストの頻度（DF）であり，単位は「千枚」である．

- **animal:** cat (461), dog (525), dolphin (56), panda (47), elephant (23), penguin (61), horse (207)
- **articraft:** watch (1271), camera (881), phone (745), car (1610), motorcycle (109), washer (24), boot (176), tv (1768)
- **person:** Obama (321), Putin (9), Einstein (17)
- **scene:** sunset (181), firework (48), beach (599), snow (214), night (938), aurora (31)
- **plant:** sunflower (21), cacti (3), rose (264), tulip (31), apple (93), mushroom (45)

類似画像検索における特徴量のパラメータは，経験的に設定した以下のものである．

(1) 色特徴量（**color**）

RGB の各成分の分割数をそれぞれ 12，21，6 としており，得られた $12\times21\times6=1,512$ 次元の特徴量を PCA によって 100 次元に圧縮して用いる．

(2) 形状特徴量（**shape**）
領域分割数 5×5，2 段階の解像度，2 種類の正規化特徴量を組み合わせて得られる，$2\times5\times5\times28\times2=2,800$ 次元の特徴量を 200 次元に圧縮して用いる．

(3) 色＋形状特徴量 1 （**color+shape 1**）
入力画像を縦横それぞれ 4 つに構図分割して，領域ごとに RGB カラーヒストグラムを生成し，それぞれ PCA で 200 次元に圧縮する．得られた $4\times4\times200=3,200$ 次元の特徴量を，さらに PCA で 200 次元に圧縮し，上位 50 次元のみを取り出す．これに加え，(1) の色特徴量の上位 50 次元，(2) の形状特徴量の上位 100 次元を組み合わせ，200 次元の特徴量とする．

(4) 色＋形状特徴量 2 （**color+shape 2**）
(3) の色特徴量 3,200 次元，(2) の形状特徴量 2,800 次元，計 6,000 次元の特徴量を PCA で 200 次元に圧縮し，縦横比について 1 次元を加えた 201 次元の特徴量とする．

各特徴量は，ランダムに選択した 10 万画像の特徴量の分散値で正規化しており，任意の 2 体間の 2 乗距離の期待値は 2 である．

単語評価の指標としては，TF，TF-IDF，KL，HGS を用いた．

本評価では，テストセットの画像を 1 枚選んでシステムに入力し，出力されたキーワードの上位 10 位以内に，入力画像の概念と同じ単語が含まれていれば正解とする．各概念 100 枚の画像を評価し，正解数/100 をその概念の推定精度とする．なお，DF が画像総数の 0.002%以下または 2%以上の単語はノイズとして除去してから評価した．

3.2 キーワード推定の精度評価

以下では，画像特徴量や単語の評価指標が推定精度に与える影響について述べる．

図 3 は，類似画像検索に用いる画像特徴量を変えた時のキーワード推定精度を 30 種類の概念毎に比較したものである．表 1 は，カテゴリ毎と全体の平均精度である．画像アノテーションに用いる類似画像数 $M=100$ 枚とし，単語の評価指標には KL を使用した．

全体の平均精度は，色のみや形状のみの特徴量を用いた場合に比べて，色と形状を組み合わせることにより大きく向上していることがわかる．類似画像検索において，例えば，照明条件が異なる画像を探す場合は，形状特徴のみでの検索が有用である．また，解像度が低く，輪郭がはっきりしない場合は，色特徴のみでの検索が有用な場合がある．一方で，今回の用いたテストセットの概念を特徴付けるためには色と形状の両方が必要な場合が多く，

表 1 画像特徴量とキーワード推定のカテゴリ平均精度
Table 1 The effects of image features on estimation accuracy (average of category).

	色	形状	色+形状 1	色+形状 2
animal	15.7	21.1	32.0	31.7
material	35.5	73.1	76.1	74.8
person	38.3	46.0	49.3	51.7
scene	28.3	30.0	50.5	48.2
plant	23.3	17.0	33.3	32.2
all	27.9	38.2	50.2	49.0

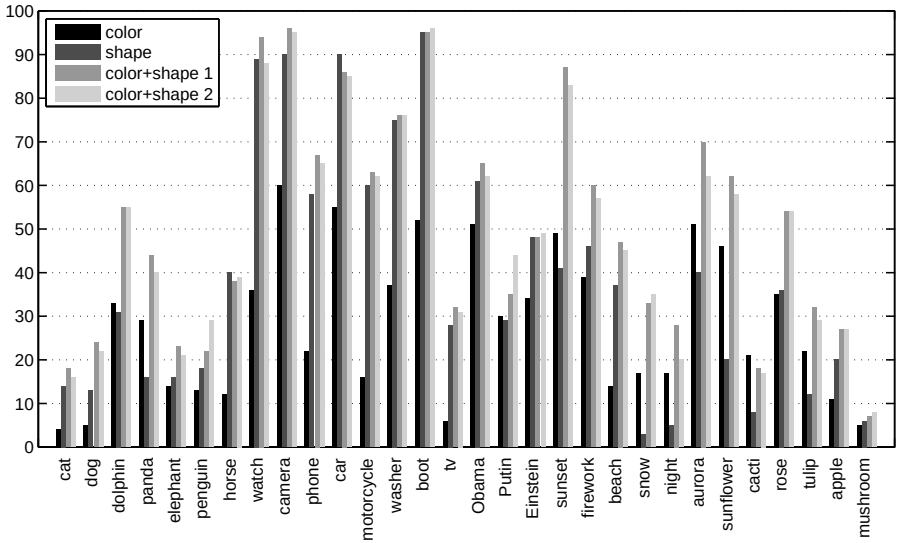

図 3 画像特徴量とキーワード推定精度
Fig. 3 The effects of image features on estimation accuracy.

どちらかの特徴を無視して広く探すより，両方の特徴を用いて絞り込んだほうが，ノイズを減らすことができるため，精度が高くなったと考えられる．

また，2 種類の「色+形状特徴量」に関しては，特徴量として画像の縦横比を含めない方の精度が若干高い．検索結果の印象としては，画像の縦横比は重要な要因になりうるが，scene のように様々なサイズで撮影されやすい画像については，縦横比を無視した方が良い検索結

表 2 単語の評価指標とキーワード推定のカテゴリ平均精度
Table 2 The effects of evalutation functions on estimation accuracy (average of category).

	TF	TF-IDF	KL	HGS
animal	22.7	32.3	32.0	32.9
artifact	79.8	78.3	76.1	75.1
person	32.0	46.0	49.3	49.3
scene	54.3	57.7	54.2	54.3
plant	17.0	29.0	33.3	33.0
all	44.0	50.3	50.2	50.1

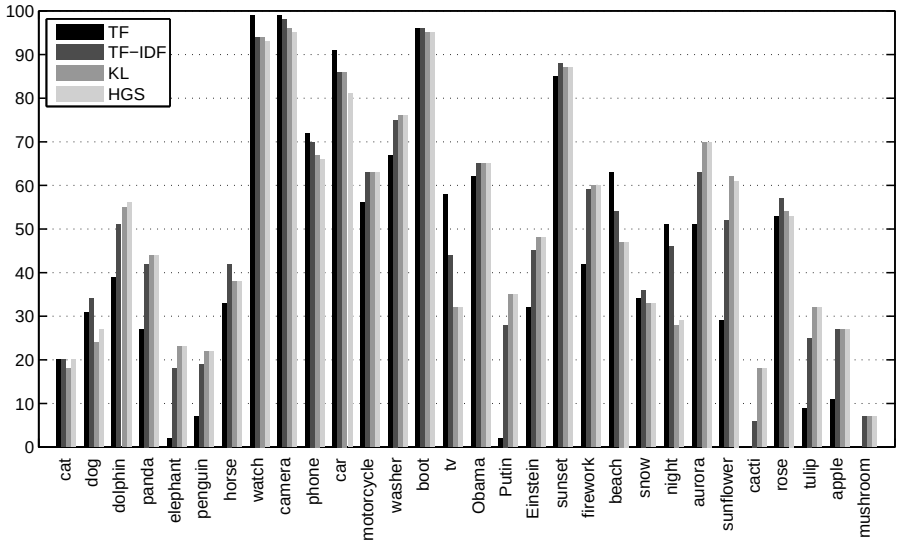

図 4 単語の評価指標とキーワード推定精度
Fig. 4 The effects of evalutation functions on estimation accuracy.

果になるためだと考えられる．

図 4 は，単語の評価指標を変えた時のキーワード推定精度を 30 種類の概念毎に比較したものである．表 2 は，図 4 の結果のカテゴリ毎の平均精度と全体の平均精度である．類似画像数 $M = 100$ 枚とし，特徴量は「色＋形状特徴量 1」を用いた．

全体の平均精度は，単純な単語頻度である TF に比べて，統計情報によるヒューリスティクスな重み付けが加わった TF-IDF，確率的指標である KL，HGS が明らかによい結果となった．TF 以外の 3 指標の平均精度にはほぼ差がないが，図 4 で概念毎に詳細に比較すると，TF-IDF は高頻度語の精度が高く，KL，HGS は低頻度語の精度が高いことがわかる．単語頻度と概念の抽象度に相関があり，低頻度語がより抽象度の低い概念であると仮定すると，確率的手法は画像をより詳細に特徴づける単語を抽出できていることになる．ただ，用途によっては抽象度の高い単語が求められる場合もあるため，今回の結果からは一概にどの指標がよいとは言えず，目的に合わせた評価指標の選択が重要である．

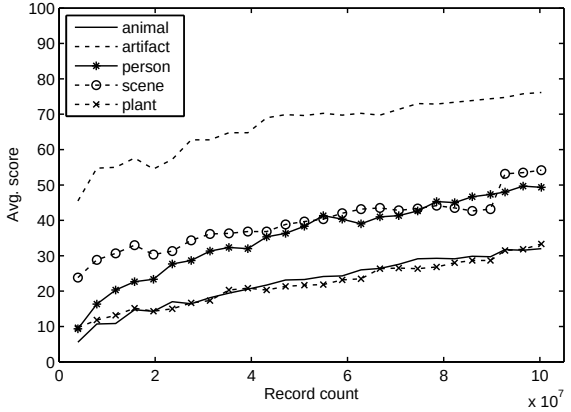

図 5 データベース規模のキーワード推定精度への影響
Fig. 5 The effects of database size on estimation accuracy.

3.3 データベースの規模とキーワード推定精度

提案手法は，大規模なデータベースから多数の類似画像を集めてくることで，S/N の低いテキスト集合から重要単語を抽出することを可能としている．そのため，母体となるデータベースの規模が推定精度に与える影響は大きいと予想される．

以下では，データベースの規模がキーワード推定精度に与える影響を評価する．本評価では，類似画像検索の対象となる画像数を変えながら，キーワード推定精度を評価した．特徴量は「色＋形状特徴量 1」を使用し，単語の評価指標には KL を用いた．

図 5 に，データベースの規模とキーワード推定精度の関係を示す．横軸は，使用したデータベースの登録画像数，縦軸はカテゴリ毎のキーワード推定の平均精度である．どのカテゴリも，データベースの規模を増やすほど精度が向上していることがわかる．animal，plant

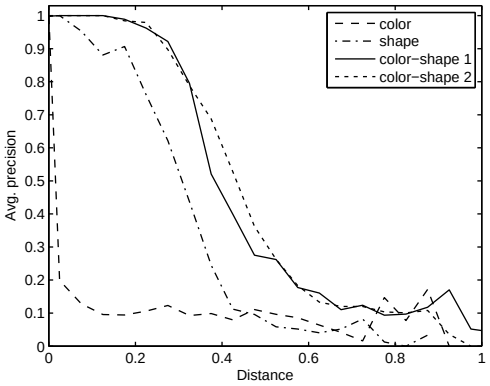

図 6 類似画像検索結果の適合率
Fig. 6 Precision curve of image search result.

のカテゴリは現状では 30%程度の精度しか出ていないが，クローリングを継続することにより，今後自動的に精度が向上していく可能性がある．また，データベースの規模の増加につれて，より詳細な概念や，認知度の低い概念の画像まで認識可能になることが期待される．

3.4 類似度を考慮した単語の重み付け

以上の精度評価においては，類似画像検索の結果の上位 $M = 100$ 件に付随するテキストから重要語を抽出した．しかし，現状のデータベースの規模では，入力画像と同一の概念を表す画像が，上位数件にしか現れないことが多く，残りの画像に付随するテキストは単にノイズになってしまう可能性が高い．類似画像検索において，検索結果がクエリと同一の概念を表しているかどうかは一概には判定できないが，特徴量距離の近い画像ほど，同一の概念の画像であることが期待できる．図 6 は，特徴量距離と平均適合率の関係を評価したものである．テストセットの各概念から 10 枚を選び，その類似画像検索結果が同一概念であるかどうかを目視で確認し，距離区間毎の平均適合率をプロットした．例えば,「色+形状特徴量 1」の場合は，距離が 0.3 以上になると急激に適合率が低下し，距離 0.6 以上の画像はキーワード推定にはほとんど役に立たないことがわかる．

以上の考察と評価結果から，特徴量距離を用いて単語の頻度に重みを付ければ，精度が向上すると考えられる．検索結果のテキスト集合をひとつの文書 d とみなし，そこに含まれる単語 t の重み付き頻度 $tf'(t,d)$ を求める式は以下である．

表 3 類似度による単語頻度の重み付けを用いたキーワード推定のカテゴリ平均精度
Table 3 The effect of weighted-TF on estimation accuracy (average of category).

	重み付けなし	重み付けあり
animal	32.0	43.6
artifact	76.1	75.3
person	49.3	66.7
scene	54.2	59.5
plant	33.3	51.7
all	50.2	59.1

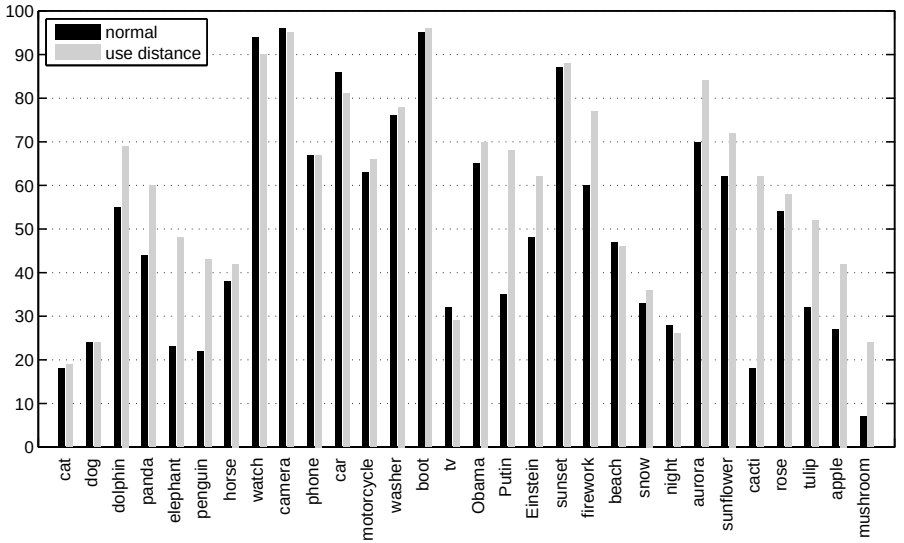

図 7 類似度による単語頻度の重み付けを用いたキーワード推定精度
Fig. 7 The effect of weighted-TF on estimation accuracy.

$$tf'(t,d) = \sum_{i}^{M} tf(t,d_i) \times f(s_i) \tag{9}$$

ここで，d_i は検索結果 i 番目のテキスト，s_i は特徴量距離であり，$f(s_i)$ は距離を用いた重み付け関数である．例えば，$f(s_i) = e^{-\sqrt{s_i}}$ とした時の精度を図 7 と表 3 に示す．画像特徴

量は「色+形状特徴量 1」，評価指標は KL である．

評価結果では，もともと検索結果の上位 100 件がほとんど同一概念の画像である場合に多少の精度低下が見られたが，総じて見ると平均精度は 50.2%から 59.1%に大きく向上している．適切な重み付け関数の設定については議論の余地があるため，今後テストセットを拡充させ，詳細な検討を実施する予定である．

3.5 処理速度

以下では，本システムにおけるデータベース構築，画像アノテーション処理の実行時間について述べる．評価環境は，Intel Core^{TM}i7 CPU 2.93 GHz，メモリ 16 GB である．

データベースの構築時間については，収集済みの約 393 万件の画像に対して，特徴量抽出，クラスタリング，テキスト処理の処理時間を測定した．

測定の結果，画像特徴量抽出に 20ms，4 種類の特徴量のクラスタリングに 60ms，テキスト解析に 10ms，合計で 1 画像あたり 90ms と非常に高速であった．一方で，データベース構築においては，データ解析よりも，データ取得がボトルネックになる．例えば，1 プロセス 1 スレッドでクローリングを行った場合，約 20 万件/日，1 画像あたり 432ms であった．

画像アノテーションの処理時間については，テストセットの 3,000 枚の画像を入力し，平均処理時間を求めた．類似画像検索に用いた特徴量は「色＋形状特徴量 1」であり，単語の評価指標は KL とした．1 回の画像アノテーションに用いる類似画像数を 100 枚とした時，平均で 2,453 個の異なり単語が評価対象になった．

測定の結果，類似画像検索に 463ms，テキスト情報の読み出しに 153ms，単語評価に 26ms であり，大部分が類似画像検索の処理時間となった．今後，データベース規模の更なる拡大のため，検索方式やパラメータ調整などを検討していく予定である．

4. おわりに

本報告では，Web 画像データベースと類似画像検索技術を用いた画像アノテーション手法を提案した．本手法は，入力画像をクエリとして類似画像検索を行い，検索結果に付随するテキスト中の単語を確率的指標により評価することで，画像を特徴付けるキーワードを推定する．本研究では，類似画像検索サービス「GazoPa」で収集した約 1 億件の Web 画像データベースと，類似画像検索プラットフォーム EnraEnra を用いることで，提案手法に基づく画像アノテーションシステムを構築した．

本研究では，構築したシステムを用いて，画像アノテーションの精度および処理速度を評価した．独自に作成したテストセットを用いた精度評価では，5 カテゴリ 30 概念の画像に対する 10 位内正解率が，カテゴリ平均で 32〜76%，全概念の平均で約 50.2%であった．単語の評価指標として TF や TF-IDF のようなヒューリスティクスな指標の代わりに，確率的モデルに基づく指標を用いることで，より抽象度の低い概念のキーワードを推定できることを述べた．また，類似度に応じて単語頻度に重み付けをすることにより，全体平均精度が 59.1%まで向上することを示した．処理速度については，データベース構築時の画像とテキストの解析に 90ms，画像アノテーション処理に 643ms であった．

今後は，より多くの概念や，詳細な概念の画像を対象とした画像アノテーションの実現に向けて研究を進めていく．そのためにも，画像データベースの拡充が必須であり，大規模データベース管理や自動クローリング技術を継続開発していく予定である．

参考文献

1) 柳井啓司：一般物体認識の現状と今後，情報処理学会論文誌: コンピュータビジョンとイメージメディア， Vol.48, No.SIG16 (CVIM 19) (2007).
2) 森靖英，高橋裕信， 岡隆一：単語群付き画像の分割クラスタリングによる未知画像からの関連単語推定，電子情報通信学会論文誌. D， Vol.84, No.4, pp.649–658 (2001).
3) Torralba, A., Fergus, R. and Freeman, W.: 80 million tiny images: A large data set for nonparametric object and scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.1958–1970 (2008).
4) Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255 (2009).
5) Wang, X., Zhang, L., Liu, M., Li, Y. and Ma, W.: ARISTA - image search to annotation on billions of web photos, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2987–2994 (2010).
6) 中山英樹，原田達也，國吉康夫：大規模 Web 画像のための画像アノテーション・リトリーバル手法，電子情報通信学会論文誌. D，Vol.93, No.8, pp.1267–1280 (2010).
7) 廣池敦，小林秀幹：類似画像検索のこれまでとこれから〜Web 画像検索サービス「GazoPa」の経験を踏まえて〜，信学技報， Vol.111, No.222, pp.21–23 (2011).
8) Matsubara, D. and Hiroike, A.: High-Speed Similarity-Based Image Retrieval with Data-Alignment Optimization Using Self-Organization Algorithm, *11th IEEE International Symposium on Multimedia*, pp.312–317 (2009).
9) ：画像検索ソリューション. `http://www.hitachi-solutions.co.jp/sis/`.
10) Hisamitsu, T. and Niwa, Y.: A measure of term representativeness based on the number of co-occurring salient words, *Proceedings of the 19th international conference on Computational linguistics*, pp.320–326 (2002).