画像・単語列間の確率的な概念獲得による高速かつ高精度な汎用的画像認識・検索手法

原田達也† 中山英樹† 國吉康夫† 大津展之‡
†東京大学大学院情報理工学系研究科 ‡産業技術総合研究所

1 はじめに

近年，画像認識に関連する膨大な数の研究が行われているが，実用レベルに達しているものは顔認識，歩行者認識のように特定の認識に特化したものが大半である．これに対し，一般的な画像を想定した場合，認識対象は非常に多岐に渡り，画像には多種多様な物体が映った極めて雑然としたものとなる．このような一般的な画像認識・検索手法として画像アノテーション・リトリーバル分野がある．この従来研究として，高い認識性能を持つ SML [1] では，画像と単語を直接結びつけたモデルを用いているが，学習に多くの時間を必要とする．また，画像と単語の上位に位置する潜在変数を仮定した研究として，MBRM [2] が挙げられるが，これは潜在変数として学習サンプルのインスタンスを用いており，その認識性能は劣ってしまう．

そこで本研究では，複数の事物が写っている画像から，速度と精度を両立させた画像アノテーション・リトリーバル手法の開発を目的とする．

2 提案手法

2.1 概念獲得による画像と単語の関係のモデル化

画像アノテーション・リトリーバルを実現するためには，画像特徴 $\boldsymbol{x}$ と単語特徴 $\boldsymbol{w}$ の関連性を学習しなければならないが，これは両者の同時確率 $p(\boldsymbol{x},\boldsymbol{w})$ を学習する問題と定式化できる．画像特徴と単語特徴から得られる概念を $\boldsymbol{l}$ とし，概念が与えられた時に，画像特徴と単語は条件付き独立であると仮定する．抽象的な概念を導入すると $p(\boldsymbol{x},\boldsymbol{w})$ を次のように表現できる．

$$p(\boldsymbol{x},\boldsymbol{w}) = \sum_{i=1}^{N_l} p(\boldsymbol{x}|\boldsymbol{l}_i)p(\boldsymbol{w}|\boldsymbol{l}_i)P(\boldsymbol{l}_i), \tag{1}$$

ここで，N_l は潜在変数の状態数である．

次に，画像と単語列から得られる概念の選択が重要となる．本研究では，概念の獲得に正準相関分析を用いる．正準相関分析では二つの変量の直接的な関係を求めるのではなく，二つのベクトル間の相関をもっともよく表すような新しい変量に変換し，その変量によって二つのベクトル間の相関について理解しようとする．画像の特徴量 $\boldsymbol{x}$ を変換した特徴量を $\boldsymbol{s}$，単語の特徴量 $\boldsymbol{w}$ を変換した特徴量を $\boldsymbol{t}$ とする．最も高い相関から得られる $\boldsymbol{s}$ や $\boldsymbol{t}$ が概念 $\boldsymbol{l}$ に相当すると考える．ここでは，$\boldsymbol{l}$ として $\boldsymbol{s}$ を利用する．$\boldsymbol{s}_i$ の起きる確率は全て同じであると考えると式 (1) は次のようになる．

$$p(\boldsymbol{x},\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_i)p(\boldsymbol{w}|\boldsymbol{s}_i). \tag{2}$$

また，$p(\boldsymbol{x}|\boldsymbol{s}_i)$ の計算には，画像特徴の空間ではなく，概念空間での距離を利用する．概念空間は，画像と単語群の関係性から得られた空間であるので，ここでの距離を利用することにより，ノイズに依らないアノテーションに本質的な空間で各画像を比較することができる．ここでは，$p(\boldsymbol{x}|\boldsymbol{s}_i)$ には $\boldsymbol{s}_i$ を中心としたガウス分布を利用する．概念空間の次元を M，$\boldsymbol{x}$ を概念空間に変換した点を $\boldsymbol{s}$ とすると，$p(\boldsymbol{x}|\boldsymbol{s}_i)$ は以下のようになる．

$$p(\boldsymbol{x}|\boldsymbol{s}_i) = \frac{1}{\sqrt{(2\pi)^M|\Sigma|}} e^{\left(-\frac{1}{2}(\boldsymbol{s}-\boldsymbol{s}_i)^T\Sigma^{-1}(\boldsymbol{s}-\boldsymbol{s}_i)\right)}, \tag{3}$$

ここで Σ は共分散行列であり，$\Sigma = \beta I_M$ とする．また，画像特徴に比べ次元が大きく圧縮された正準変量のベクトル計算で済むため，計算コストが大きく削減できる．

$p(\boldsymbol{w}|\boldsymbol{s}_i)$ は，MBRM [2] と同様の言語モデルを用いてトップダウンに設計することを考える．

$$p(\boldsymbol{w}|\boldsymbol{s}_i) = \prod_{w\in\boldsymbol{w}} p_W(w|\boldsymbol{s}_i), \tag{4}$$

$$p_W(w|\boldsymbol{s}_i) = \mu\delta_{w,\boldsymbol{s}_i} + (1-\mu)\frac{N_w}{N_W}, \tag{5}$$

ここで，N_W は全学習データのラベル総数，N_w は全学習データにおける単語 w の出現回数，$\delta_{w,\boldsymbol{s}_i}$ は w が $\boldsymbol{s}_i$ にラベル付けされていたら 1，そうでなければ 0 をとる．

2.2 画像特徴と単語特徴

ここで利用する画像特徴量としては高次局所自己相関特徴（HLAC）[3] を色画像に拡張した Color-HLAC [4] を利用する．一般に画像のどの部分にアノテーションを行うべき対象が存在するか分からないし，対象物体がいくつあるかも分からない．画像のセグメンテー

High speed and high accuracy image annotation/retrieval based on probabilistic conceptual learning between images and labels
†T. Harada, †H. Nakayama, †Y. Kuniyoshi and ‡N. Otsu
†The University of Tokyo
‡National Institute of Advanced Industrial Science and Technology

ションを行ってから，各領域ごとにアノテーションを行うことも考えられるが，一般に画像のセグメンテーションは非常に難しいため，これがボトルネックとなりアノテーションの性能低下の大きな要因となる可能性が高い．画像のセグメンテーション性能に依存しないアノテーションを行うためには，取り出した画像特徴量に位置不変性や加法性があることが重要な要件となるが，HLAC はこの二つの性質を備えるものである．

各画像には一つ以上の単語が付与されている．例えば，ある画像に対して，"空"，"飛行機"，"雲"などである．最も単純な特徴量としては，記号が割りあたっているときには 1，割りあたっていない場合は 0 となる特徴量に変換するものである．

2.3 画像アノテーション・リトリーバル

画像アノテーションを行うには，未知画像から画像特徴量 $\boldsymbol{x}_{new}$ を取り，単語群の事後確率 $p(\boldsymbol{w}|\boldsymbol{x}_{new})$ を求める．ただし，$p(\boldsymbol{x}_{new})$ はどの単語に対しても同じ値をとるので，以下の式を計算するだけでよい．

$$h = \sum_{i=1}^{N} p(\boldsymbol{x}_{new}|\boldsymbol{s}_i)p(\boldsymbol{w}|\boldsymbol{s}_i). \tag{6}$$

各単語の h の高い順に画像に単語を割り当てることで，未知画像に単語群を付与することが可能となる．

画像のリトリーバルには尤度を用いる．画像を引き出す単語群の特徴量を $\boldsymbol{w}_{new}$ とする．$\boldsymbol{w}_{new}$ から尤度 $p(\boldsymbol{w}_{new}|\boldsymbol{x})$ を計算するために次の式を利用する．

$$g = \frac{\sum_{j=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_j)p(\boldsymbol{w}_{new}|\boldsymbol{s}_j)}{\sum_{j=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_j)}. \tag{7}$$

全ての画像に関して g を計算し，この g の大きい順に画像を引き出してくることでランク付けされた画像を取り出すことが可能となる．

3 実験

画像アノテーション・リトリーバルの精度，速度を従来手法と比較を行う．データセットとして，Corel5K [5] を用いる．これは，5000 枚の画像から構成され，この分野の標準的なベンチマークとなっている．アノテーションの評価は，単語ごとの Recall，Precision の平均値と F 値を用いる．リトリーバルの評価手法に関しては，Mean Average Precision を用いて評価する．

図 1 左にアノテーションの性能比較の結果，図 2 に未知画像にアノテーションした例を示す．このように，提案手法は従来手法を凌駕し，多種多様な画像に対して柔軟に認識が可能である．図 1 右にリトリーバルの性能比較を示す．このように，リトリーバルにおいても，提案手法は従来手法を大きく上回っている．

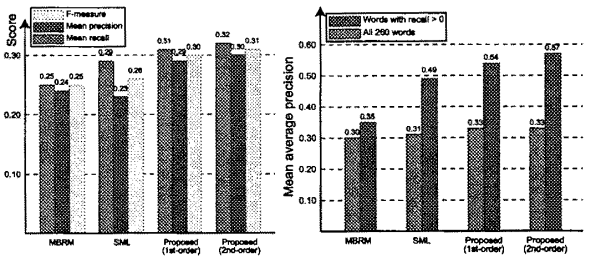

図 1: Performance comparison of the proposed and previously published methods on Corel5K bench mark test.

Corel5K において，従来の最高手法である SML [1] では LinuxPC 3000 台を用い，学習に約 1 時間，500 枚のサンプルの認識に約 280 秒を要した．提案手法は高々 1 次の HLAC を用いた場合に，市販のデスクトップ PC 1 台で学習は遅くとも 1 時間で終了し，500 枚のサンプルが約 10 秒程度で認識可能であり，性能で優位を保ちかつ圧倒的に高速なアルゴリズムとなっている．

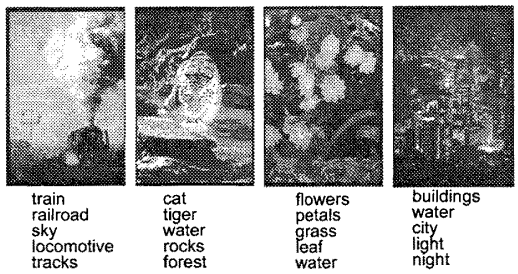

図 2: Annotation examples.

4 まとめ

本研究では，高次局所自己相関特徴と確率的正準相関分析，言語モデルの組み合わせにより，画像・単語間の概念対応の確率構造を柔軟に学習する新しい画像アノテーション・リトリーバル手法を提案し，標準的なデータセットを用い，本手法が精度・速度の両面で既存手法を圧倒的に上回ることを示した．

参考文献

[1] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. PAMI*, Vol. 29, No. 3, pp. 394–410, 2007.

[2] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proc. IEEE Conf. CVPR*, Vol. 2, pp. 1002–1009, 2004.

[3] N. Otsu and T. Kurita. A new scheme for practical, flexible and intelligent vision systems. In *Proc. IAPR Workshop on Computer Vision*, 1988.

[4] T. Kato, T. Kurita, N. Otsu, and K. Hirata. A sketch retrieval method for full color image database –query by visual example–. In *Proc. of 11th ICPR*, Vol. 2, pp. 213–216, 1992.

[5] P. Duygulu, K. Barnard, and D. F. N. Freitas. Object recognition as machine translation: Learning a lexicon for a fixed image vocabular. In *Proc. ECCV*, pp. 349–354, 2002.