自らの視覚記憶を言葉で検索可能とする AI Goggles

AI Goggles: Retrieving Visual Memories by Words

原田達也*1　中山英樹*1　國吉康夫*1
Tatsuya Harada　Hideki Nakayama　Yasuo Kuniyoshi

*1東京大学　大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

In this paper, we present the AI Goggles, which can instantly describe objects and scenes in the real world and retrieve visual memories about them using keywords input by the users. This is a stand-alone wearable system working on a tiny mobile computer. The system can quickly learn unknown objects and scenes by teaching and learn to label and retrieve them on site, without loss of recognition ability for previously learnt ones. As the core algorithm of the system, we propose and implement a new method of multi labeling and retrieval of unconstrained real-world images. This system is expected to be a memory aid system for the elderly.

1. はじめに

インターネットの普及により世界中の情報を瞬時に検索可能になり，一気に世界の距離が縮まってはいるが，自らの記憶を検索する手段は未だに存在しない．自らの視覚記憶を検索する手段が確立されれば，過去の自分との距離を縮め，記憶支援のみならず幅広い応用が期待できる．

そこで我々は，視覚情報から環境中の事物を柔軟に認識・検索を行う AI Goggles [原田 08, Nakayama 09] の研究を進めている．これは，カメラを備えたゴーグルと携帯型 PC からなるシステムである（図 1）．本システムは，装着者の視界と同期したカメラ画像を取得し，リアルタイムにその画像にラベル付けを行い，その結果を Head Mount Display（HMD）に出力する．同時に，認識されたラベルと取得画像を合わせて視覚ログとして計算機に保存する．検索時には，言葉を用いて視覚ログから該当する画像を選び，HMD へ出力する．同様のシステムとして Torralba らの研究 [Torralba 03] が挙げられるが，大規模な確率モデルの学習を必要とするため，認識対象数は限られている．より汎用的なシステムの実現には，制約のない実世界画像から多様な物体やシーンを認識・検索できる技術が必要となるが，既存手法は非常に膨大な計算処理を必要とするものが多く，実世界処理への応用は困難であった．

本稿では，提案する AI Goggles により，限定的な計算資源においてもリアルタイムかつ高精度に実世界シンボル化が行えることを示す．第 2 章ではこのシステムのコアアルゴリズムについて述べる．この手法を CCAsim と呼ぶ．第 3 章では画像・単語特徴，第 4 章では追加学習法，第 5 章では画像認識検索法について述べる．特に追加学習は想定しない事象が多発する実世界において必須の手法である．第 6 章では定量的評価と実世界での実験，第 7 章では本稿のまとめを述べる．

2. 提案手法

2.1 モデル

本研究の目的を達成するには，画像と単語との関連性を学習しなければならない．画像から得られる特徴量を $\boldsymbol{x}$，単語の特徴量を $\boldsymbol{w}$ とし，画像の特徴量と単語の特徴量が同時に出力

連絡先: 原田達也，東京大学　大学院情報理工学系研究科，〒113-8656　東京都文京区本郷 7-3-1 工学部 2 号館，03-5841-1650，harada@isi.imi.i.u-tokyo.ac.jp

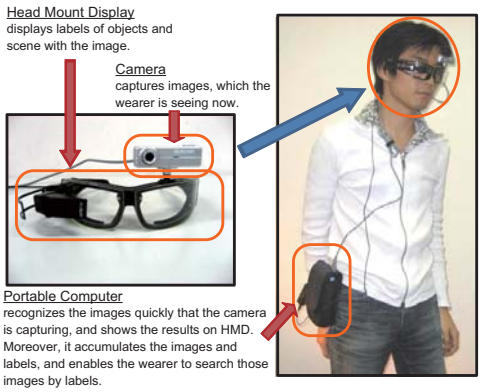

図 1: Overview of AI goggles system.

する確率を $p(\boldsymbol{x}, \boldsymbol{w})$ とする．この $p(\boldsymbol{x}, \boldsymbol{w})$ を求めることが画像と単語との関係性を学習することである．画像特徴量（$\boldsymbol{x}$）と単語特徴量（$\boldsymbol{w}$）から得られる概念を $\boldsymbol{l}$ とする．さらに，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画像特徴量と単語特徴量が同時に起きる確率 $p(\boldsymbol{x}, \boldsymbol{w})$ を次のように表現できる．

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

ここで，N_l は隠れ状態の数である．

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（CCA）を用いる．CCA 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．言い換えると，画像の特徴量 $\boldsymbol{x}$ を変換した特徴量を $\boldsymbol{s}$，この変換を $f : \boldsymbol{x} \rightarrow \boldsymbol{s}$，単語の特徴量 $\boldsymbol{w}$ を変換した特徴量を $\boldsymbol{t}$，この変換を $g : \boldsymbol{w} \rightarrow \boldsymbol{t}$ とする．変換された $\boldsymbol{s}$ と $\boldsymbol{t}$ の相関が最も高くなるように変換 f と g を求めるものが CCA である．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．実際に，確率的正準相関分析 [Bach 05] の知見より，CCA の構造は式 (1) の構造と同一となる．

今，p 個の変量を含む画像の特徴量ベクトル $\boldsymbol{x}$ =

$(x_1, \cdots, x_p)^T$ と q 個の変量を含む単語の特徴量ベクトル $\boldsymbol{w} = (w_1, \cdots, w_q)^T$ があるとする．CCA では，学習用のデータ集合 $\mathcal{D}$ =$\{(\boldsymbol{x}_1, \boldsymbol{w}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{w}_N)\}$ に対して，線形変換

$$\boldsymbol{s}_i = \boldsymbol{A}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \boldsymbol{A}^T\hat{\boldsymbol{x}}_i, \quad (2)$$

$$\boldsymbol{t}_i = \boldsymbol{B}^T(\boldsymbol{w}_i - \bar{\boldsymbol{w}}) = \boldsymbol{B}^T\hat{\boldsymbol{w}}_i, \quad (3)$$

で写された二つの新変量群間の相関行列のトレースの絶対値が最大となるような係数行列 $\boldsymbol{A}$ および $\boldsymbol{B}$ を求める．

サンプルから得られる共分散行列を $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{xx} & \boldsymbol{C}_{xw} \\ \boldsymbol{C}_{wx} & \boldsymbol{C}_{ww} \end{pmatrix}$ とすると，最適な係数行列は固有値問題

$$\boldsymbol{C}_{xw}\boldsymbol{C}_{ww}^{-1}\boldsymbol{C}_{wx}\boldsymbol{A} = \boldsymbol{C}_{xx}\boldsymbol{A}\boldsymbol{\Lambda}^2 \ (\boldsymbol{A}^T\boldsymbol{C}_{xx}\boldsymbol{A} = \boldsymbol{I}_d), \quad (4)$$

$$\boldsymbol{C}_{wx}\boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xw}\boldsymbol{B} = \boldsymbol{C}_{ww}\boldsymbol{B}\boldsymbol{\Lambda}^2 \ (\boldsymbol{B}^T\boldsymbol{C}_{ww}\boldsymbol{B} = \boldsymbol{I}_d), \quad (5)$$

の解として求められる．ここで $\boldsymbol{\Lambda}^2$ は固有値を対角要素とする対角行列であり，d は正準成分および $\boldsymbol{s}, \boldsymbol{t}$ の次元で，係数行列として固有値の大きいものから順に d 個取られる．これより $\boldsymbol{s}_i = \boldsymbol{A}^T\hat{\boldsymbol{x}}_i$ が変換 f，$\boldsymbol{t}_i = \boldsymbol{B}^T\hat{\boldsymbol{w}}_i$ が変換 g に相当する．CCA は行列の固有値問題を一度計算するだけでよく，非常に高速に計算が可能である．

$\boldsymbol{s}, \boldsymbol{t}$ と潜在変数 $\boldsymbol{z}$ との関係は，確率的 CCA から事後期待値を計算することにより次のように求められる．

$$E(\boldsymbol{z}|\boldsymbol{x}) = \boldsymbol{G}_1^T\boldsymbol{s}, \quad (6)$$

$$E(\boldsymbol{z}|\boldsymbol{w}) = \boldsymbol{G}_2^T\boldsymbol{t}, \quad (7)$$

ここで $\boldsymbol{G}_1, \boldsymbol{G}_2$ は $\boldsymbol{G}_1\boldsymbol{G}_2^T = \boldsymbol{\Lambda}^2$ を満たす任意の行列である．

2.2 確率密度関数の選択

学習のデータ集合 $\mathcal{D}$ から CCA を利用し，概念 $\boldsymbol{l}_i$ へのマッピング f と g が得られているとする．このマッピング f, g を用いて移された学習のデータ集合を $\mathcal{D}_l = \{(\boldsymbol{s}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{s}_N, \boldsymbol{t}_N)\}$ とする．またここでは，$\boldsymbol{l}$ として $\boldsymbol{s}$ を利用する．$\boldsymbol{s}$ を利用すると式（1）は次のようになる．

$$p(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_i)p(\boldsymbol{w}|\boldsymbol{s}_i)P(\boldsymbol{s}_i). \quad (8)$$

$\boldsymbol{s}_i$ の起きる確率は全て同じであると考えると，$P(\boldsymbol{s}_i) = 1/N$ となるので，次のようにできる．

$$p(\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_i)p(\boldsymbol{w}|\boldsymbol{s}_i). \quad (9)$$

また，$p(\boldsymbol{x}|\boldsymbol{s}_i)$ は概念 $\boldsymbol{s}_i$ から画像の特徴量 $\boldsymbol{x}$ が出現する確率であるが，この確率の計算には，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特徴で各画像を比較することが可能になる．そのため，$p(\boldsymbol{x}|\boldsymbol{s}_i)$ には概念空間での $\boldsymbol{s}_i$ を中心としたガウス分布を利用する．$\boldsymbol{x}$ を f で正準空間にマッピングした点を $\boldsymbol{s}$ とすると，$p(\boldsymbol{x}|\boldsymbol{s}_i)$ は以下のようになる．

$$p(\boldsymbol{x}|\boldsymbol{s}_i) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} e^{\left(-\frac{1}{2}(\boldsymbol{s}-\boldsymbol{s}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{s}_i)\right)}, \quad (10)$$

ここで $\boldsymbol{\Sigma}$ は分散である．ここでは，$\boldsymbol{\Sigma} = \beta I_d$ とする．バンド幅 β を用いて，事後確率密度分布の滑らかさを設計できる．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，ある概念 $\boldsymbol{s}_i$ から単語列の特徴量 が生成される確率を表す．ここでは，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．代表的な例として，MBRM [Feng 04] などが挙げられる．これらは共に，単語列の生起確率を，単語列に含まれる各単語の生起確率をベースに計算するものである．本論文では，これらを参考に次のモデルを用いることにする．

$$p(\boldsymbol{w}|\boldsymbol{s}_i) = \prod_{w \in \boldsymbol{w}} p_W(w|\boldsymbol{s}_i), \quad (11)$$

ここで，p_W は単語の生起確率である．さらに，

$$p_W(w|\boldsymbol{s}_i) = \mu\delta_{w,\boldsymbol{s}_i} + (1-\mu)\frac{N_w}{N_W}, \quad (12)$$

ここで，N_W は全学習データのラベル（アノテーション）総数，N_w は全学習データにおける単語 w の出現回数，$\delta_{w,\boldsymbol{s}_i}$ は単語 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．なお，μ は 0 から 1 までの実数値をとるパラメータであり，1 に近づける程各サンプルについたアノテーションを重視し，逆に 0 に近づけるほど全体の出現頻度を重視することになる．本研究では，$\mu = 0.99$ に固定する．

画像認識・検索で CCA を利用することは珍しくない．例えば，栗田らの研究 [栗田 92] では，後述する Color-HLAC と CCA を用いている点で，我々の手法に非常に近い．しかし，CCA をそのまま利用することは，画像と単語から得られる潜在変数の分布構造を無視し，線形の関係でとらえているために，情報の欠落が大きく，豊富な単語のアノテーションには向いていない．我々のアプローチは，従来の CCA の計算速度を維持しつつ，正準空間での分布構造を利用することで表現能力を高める効果を狙ったものである．

3. 画像と単語の特徴量

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）[Otsu 88] を色画像に拡張した Color-HLAC [栗田 92] を利用する．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．本研究では，元画像と上記前処理を行った加工画像の両方から Color-HLAC 特徴を抽出し，画像特徴として用いる．

各画像には一つ以上の単語が付与されている．これから記号列の特徴量に変換し，この変換された特徴量を $\boldsymbol{w}$ とする．本研究では，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．例えば，候補となる単語が”空”，”海”，”山”，”飛行機”，”雲”の 5 つで，画像に割りあたっている単語が，”空”，”飛行機”，”雲”であれば，このときの特徴量 は (1,0,0,1,1) となる．

4. 追加学習

提案手法は CCA を基盤とした手法であり，CCA を逐次的に拡張することで追加学習が可能となる．しかし従来のパーセプトロンを基にした追加学習のアルゴリズムでは，追加した学習データの認識性能は高くなるが従来学習したデータの認識性能が次第に低下する傾向にある．CCA のアルゴリズムは，式

(4)，式 (5) の固有値の計算がコストの高い部分であるが，これは特徴量の次元数で決定され学習サンプル数に依存しないため，現在利用されている汎用的な計算機のパワーを考慮するとコストは比較的低いと言える．しかしながら前段の共分散行列の計算コストは，学習サンプルが増えるにつれて膨大となるために，この部分がボトルネックとなる．そこで，CCA の固有値計算部分はそのままにし，共分散行列の計算のみを逐次的に行うことにする．これにより学習サンプルが増加しても一定の計算コストで済み安定した結果を得ることが可能となる．

ここで，すでに t 個の学習サンプルが得られているとする．学習サンプルの平均，相関行列，共分散行列をそれぞれ $\boldsymbol{m}$，$\boldsymbol{R}$，$\boldsymbol{C}$ とする．新たな学習サンプル $\{\boldsymbol{x}_{t+1}, \boldsymbol{w}_{t+1}\}$ が得られると画像に関する上記の変数を次式により更新する．

$$\begin{aligned} \boldsymbol{m}_x &= \frac{t-l}{t+1}\boldsymbol{m}_x + \frac{1+l}{t+1}\boldsymbol{x}_{t+1}, \\ \boldsymbol{R}_{xx} &= \frac{t-l}{t+1}\boldsymbol{R}_{xx} + \frac{1+l}{t+1}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^T, \\ \boldsymbol{C}_{xx} &= \boldsymbol{R}_{xx} - \boldsymbol{m}_x\boldsymbol{m}_x^T, \end{aligned} \tag{13}$$

ここで l は新規学習サンプルに対する重みを表す．上記の更新は，単語，画像と単語の共分散行列 $\boldsymbol{C}_{ww}$，$\boldsymbol{C}_{xw}$ についても同様に行う．

新規の学習サンプルに付与されていない単語が出現した場合は，平均と相関行列の計算を下記のように行う．これは単語特徴が一単語について一つの次元を割り当てているためである．

$$\begin{aligned} \boldsymbol{m}_w &= \frac{t-l}{t+1}\begin{pmatrix} \boldsymbol{m}_w \\ 0 \end{pmatrix} + \frac{1+l}{t+1}\boldsymbol{w}_{t+1}, \\ \boldsymbol{R}_{ww} &= \frac{t-l}{t+1}\begin{pmatrix} \boldsymbol{R}_{ww} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{pmatrix} + \frac{1+l}{t+1}\boldsymbol{w}_{t+1}\boldsymbol{w}_{t+1}^T, \\ \boldsymbol{R}_{xw} &= \frac{t-l}{t+1}\begin{pmatrix} \boldsymbol{R}_{xw} & \boldsymbol{0} \end{pmatrix} + \frac{1+l}{t+1}\boldsymbol{x}_{t+1}\boldsymbol{w}_{t+1}^T. \end{aligned} \tag{14}$$

これと式 (4)，式 (5) を用いることで新規の射影行列 $\boldsymbol{A}$，$\boldsymbol{B}$ を獲得し，この射影行列を用いて全てのサンプルを正準空間に再度マッピングする．

5. 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．

$$h = \sum_{i=1}^{N} p(\boldsymbol{x}_{new}|\boldsymbol{s}_i)p(\boldsymbol{w}|\boldsymbol{s}_i). \tag{15}$$

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．

$$g = \frac{\sum_{j=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_j)p(\boldsymbol{w}_{new}|\boldsymbol{s}_j)}{\sum_{j=1}^{N} p(\boldsymbol{x}|\boldsymbol{s}_j)}. \tag{16}$$

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．

6. 実験

6.1 Corel5K を用いた定量的比較

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 [Duygulu 02] 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（それぞれ MR と MP），リトリーバルの評価手法に関しては，全ての単語に対する Mean Average Precision（MAP）と Recall が 0 より大きい単語のみの MAP（MAP-RP）を用いて評価する．

表 1: Experimental results

	MR	MP	MAP	MAP-RP
MBRM [Feng 04]	0.25	0.24	0.30	0.35
SML [Carneiro 07]	0.29	0.23	0.31	0.49
DCMRM [Liu 07]	0.28	0.23	-	-
JEC [Makadia 08]	0.32	0.27	0.33	0.52
CCAsim [Nakayama 08]	0.32	0.25	0.32	0.58

表 1 に性能の比較を載せる．このように，提案手法は MBRM，SML，DCMRM などと比較し高い性能を示すことがわかる．JEC とはそれぞれのスコアに優劣があり判断が難しいがほぼ同等の性能と言える．

Corel5K において，SML では LinuxPC 3000 台を用い 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，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．また，JEC との比較においては，JEC は kNN ベースの手法を用いており，認識にかかる時間は提案手法と同様に学習サンプル数に比例する．しかしながら JEC はサンプル間距離の計算に画像特徴の空間を用いており，画像特徴は高次元であるためにこの計算コストが高い．本提案手法は，正準空間の高く圧縮された空間での距離計算を行うために，JEC と比較し高速に認識が可能となる．

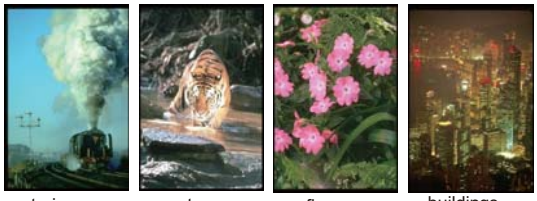

図 2: Annotation examples.

6.2 実世界での実証実験

本システムを図 1 に示す．このように，ゴーグルにカメラと HMD を搭載し，これらをタブレット型計算機（Core2Duo 1.2GHz，2GB RAM）に接続したものである．全ての処理はこのタブレット型計算機によって行われ，他のいかなる外部資源も必要としない．計算機は非常に小型であるため，容易に持ち歩くことが可能であり，装着者の負担は小さい．

実験環境の設定として，一般的な生活環境を想定した部屋をセットアップした．desk，chair，table のような家具にあた

る大きな物体や，机の上に配置した books，clock，photo などのような小物など，多様な物体を配置した．また，テーブルの上には観葉植物をおいてあるが，これらについてはその種類まで詳しく学習させる．これは，単に身の回りの物品の認識に留まらない，映像辞書のような用途として用いることを想定したものである．このように，環境中におかれたさまざまなレベルの物体のアノテーション・リトリーバルを行う．現在，用意している単語は 33 個である．

あらかじめ，ゴーグルを用いて大量の画像を撮影し，それぞれについて数語のラベルを付与し，学習サンプルとする．この時のサンプルはできるだけ多くとることが望ましい．また，本手法で画像特徴として用いる HLAC 特徴は，画像全体を積分するため，画面上で小さい物体の特徴は埋もれてしまう場合がある．このため，学習させたい物体はある程度近づいた状態で撮影する．今回は約 2500 サンプルにより学習を行った．

次に，学習した結果を用いて，実際にゴーグルを装着し，環境内の物体のリアルタイム認識を行った．視覚ログの保存は毎秒ごとに行う．視覚ログには，画像自身と，各単語の事後確率が含まれる．今回は，事後確率の閾値を 0.40 に設定し，この値を超えた単語をシステムの認識結果とした．認識結果の一例を図 3 に示す．単語の横の数値が，各単語の事後確率を表す．このように，実世界中の多様な物体について，柔軟に認識が可能であることが示された．さらに，記録した視覚ログの中から，特定の物体が映った画像を検索させる実験を行った．検索単語の尤度の大きい順に候補画像を出力する単純な処理であるため，検索は瞬間的に行われる．この結果の一例を図 4 に示す．このように，さまざまな画像が含まれる視覚ログの中から所望の画像を正確に取り出せていることが分かる．

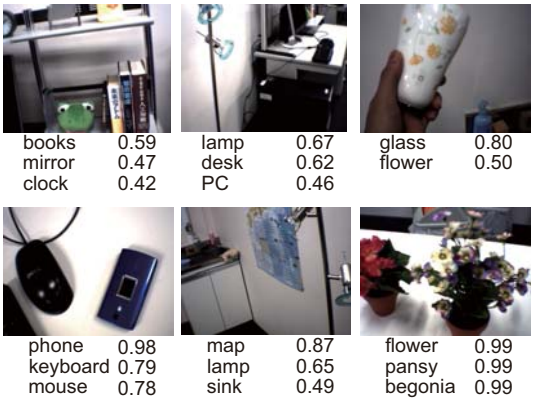

図 3: Annotation examples with goggles.

図 4: Retrieval examples with goggles.

追加学習については，未知画像を発見したときに，PC を通じてユーザが複数の単語を教示することで学習を行った．未知画像を複数の角度から約 10 秒間眺めることで多くの物体について学習が安定した．追加学習にかかる時間は現在の実験セットアップで 1 フレームあたり約 0.1 秒であり，通常の CCA の計算が 3 秒近くかかるため十分高速と言える．また，40 単語初期状態に学習してから，追加学習で 106 単語まで教示し，認識させる実験を行い，定量的な評価は行っていないものの破綻することなく適切な結果を得ることを確認している．

7. まとめ

本研究では，ゴーグル型のシステムの提案と実現を行い，装着者が見たものをリアルタイムでアノテーションし，この結果を用いて見た映像を言葉でリトリーバルを行うことで，自らの視覚記憶の言葉による検索が可能であることを示した．このシステムで用いられているコアアルゴリズムは，高次局所自己相関特徴と確率的正準相関分析の拡張，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・リトリーバル手法である．また，標準的なデータセットを用いて，本手法が精度の面で既存の最良手法と同等の性能を示し，認識速度では上回ることを示した．さらに，未知画像を追加学習する機能を提案・実現し，破綻することなく動作することを確認した．

参考文献

[Bach 05] Bach, F. R. and Jordan, M. I.: A probabilistic interpretation of canonical correlation analysis, Technical Report 688, Department of Statistics, University of California, Berkeley (2005)

[Carneiro 07] Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval, *IEEE Trans. PAMI*, Vol. 29, No. 3, pp. 394–410 (2007)

[Duygulu 02] Duygulu, P., Barnard, K., and Freitas, D. F. N.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, in *Proc. IEEE ECCV* (2002)

[Feng 04] Feng, S., Manmatha, R., and Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation, in *Proc. IEEE CVPR*, Vol. 2, pp. 1002–1009 (2004)

[Liu 07] Liu, J., Wang, B., Li, M., Li, Z., Ma, W. Y., Lu, H., and Ma, S.: Dual cross-media relevance model for image annotation, in *Proc. ACM Multimedia*, pp. 605–614 (2007)

[Makadia 08] Makadia, A., Pavlovic, V., and Kumar, S.: A new baseline for image annotation, in *Proc. ECCV*, pp. 316–329 (2008)

[Nakayama 08] Nakayama, H., Harada, T., Kuniyoshi, Y., and Otsu, N.: High-performance image annotation and retrieval for weakly labeled images using latent space learning, in *Proc. PCM*, pp. 601–610 (2008)

[Nakayama 09] Nakayama, H., Harada, T., and Kuniyoshi, Y.: AI Goggles: Real-time Description and Retrieval in the Real World with Online Learning, in *Proc. CRV* (2009)

[Otsu 88] Otsu, N. and Kurita, T.: A new scheme for practical, flexible and intelligent vision systems, in *Proc. IAPR Workshop on Computer Vision* (1988)

[Torralba 03] Torralba, A., P.Murphy, K., Freeman, W. T., and Rubin, M. A.: Context-based vision system for place and object recognition, in *Proc. IEEE ICCV* (2003)

[栗田 92] 栗田多喜夫, 加藤俊一, 福田郁美, 板倉あゆみ：印象語による絵画データベースの検索, 情報処理学会論文誌, Vol. 33, No. 11, pp. 1373–1383 (1992)

[原田 08] 原田達也, 中山英樹, 國吉康夫：超高速汎用的画像認識検索手法の開発と実世界応用, 第 4 回デジタルコンテンツシンポジウム予稿集 (2008)