マルチアングルカメラに対応した人の動作認識
A method of human action recognition for multi angle camera

宇都宮 拓之† 藤田 茂‡
Hiroyuki Utsunomiya Shigeru Fujita

1 はじめに

日常行動の映像から人の活動を記録・解析することで，ユーザの生活支援や問題発見など，様々な情報提示が行えると期待出来る．例えば，デスクワークをする人物の場合，仕事をするために席に座る，PC 作業をする，コーヒーを飲むなどの行動を取るが，映像中で頻発するうたた寝や，寝るなどの行為が見られた場合，ユーザの日常生活に対して，健康状態から見た警告を発する事が出来ると考えられる．

日常生活をおくる上で，人を支援するためにセンサを用いて様々な行動を認識する研究 [1] が既に行われている．しかし，これらセンサを用いた行動認識は，特別な装置を着ける必要があるなど，ユーザにとって心的負担が掛かるため，何も装着を行わない状態で情報観測が出来る方が望ましいと言える．

そこで本研究では，室内を撮影できるカメラを用いて，まず映像中から人物の動作特徴点の抽出を行い，その得られた連続する特徴点を離散化した単位動作記号に変換し，その単位動作記号がどの行動パターンに当てはまるのかの推定を行う事によって，行動認識をする手法を提案する．

2 関連研究

観測する映像から行動認識を行うには，どのカメラに人が写るか，カメラのどこに人が写るかという位置情報に基づいて行動認識を行う手法や，人の姿勢の変化や肌色の検出などの動きの情報に基づいて行動認識を行う方法 [2] がある．しかし，これら手法では人物位置を固定しなければならないという問題や，写る面積や角度，特定の動作しか検出が出来ないなどの問題がある．

また映像から行動認識を行う際には，複数パターンを網羅した学習データを事前に用意するよりも，用意したデータからラベル付けを自動で行う教師なし学習を用いて学習を行う方が運用上望ましいと言える．

†千葉工業大学大学院 情報科学研究科，Graduate School of Information and Computer Science, Chiba Institute of Technology
‡千葉工業大学 情報科学部，Faculty of Computer and Information Science, Chiba Institute of Technology

3 提案手法

本章では提案手法について述べる．提案する手法は大きく分けて，学習部と認識部の 2 つの段階で構成される．

1 つ目の学習部は，まず，Human Action Categorizing Method[3] を用いて，映像中の人物動作が発生する時間帯を抽出し，時空間特徴量シーケンスから Probabilistic Latent Semantic Analysis (pLSA) を用いて，単位動作の発生確率シーケンスへと変換し，単移動作の教師なし学習を行う．

2 つ目の認識部は学習部と同様に，与えられた映像から特徴点の抽出と PCA によるデータ圧縮を行い，K-means で学習されたラベルと得られた特徴データを使いラベルを得る．得られたラベルは Word であり，映像フレームが Document である．この Document と Word を使い，逐次近似 EM アルゴリズムで動作の記号列を得る．

3.1 特徴量抽出

本節では，Human Action Categorizing Method による時空間特徴の抽出について述べる．

まず映像の画素毎に，変化の激しさを返すレスポンス関数を以下のように与える．

$$R(x,y) = (I(x,y) * g(x,y;\sigma) * h_{ev}(t;\tau,\omega))^2 + (I(x,y) * g(x,y;\sigma) * h_{od}(t;\tau,\omega))^2 \quad (1)$$

ここで $I(x,y)$ は，映像中でのグレースケール画像であり，$g(x,y;\sigma)$ は，I を平滑化するためのガウシアンフィルタである．また，$h_{ev}(t;\tau,\omega)$ と $h_{od}(t;\tau,\omega)$ は一次元 Gabor フィルタであり，それぞれ以下のように与える．

$$h_{ev}(t;\tau,\omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2} \quad (2)$$

$$h_{od}(t;\tau,\omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2} \quad (3)$$

1 次元 Gabor フィルタは，映像の時間軸 t に沿って畳み込まれる．σ と ω はフィルタのスケールを表す．ω は [3] において，$\omega = 4/\tau$ が良いとされている．

このレスポンス関数は，映像の複雑な動きが発生した領域を検出することができる．例えば図.1 の人物が右から左に移動するシーンにおいて手法を適用すると，図.2 の様に特徴点を抽出する事が出来る．

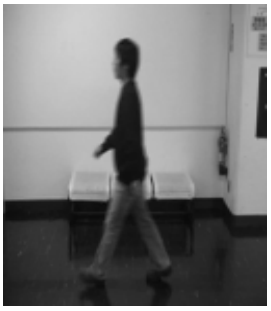

図 1: 通常画像

図 2: 特徴点抽出結果

レスポンス関数から得られた値に対して，レスポンス値が空間状で極大点を取る点の検出を行い，その画素を中心に，時間軸を含めた連続データからキューブ状に切り出す．サイズは [3] において，ω, τ の 6 倍が良いとされている．

切り出したキューブは，異なるカーネルサイズでフィルタリング処理をし，画素毎に輝度勾配列に変換して 1 本のベクトルを作成する．そして PCA を用いて各特徴ベクトルを次元圧縮し，K-means 法を用いてクラスタリングを行う．

これにより，特徴データは各クラスタに分類される．画像中の各フレームを一つの Document とし，各 Document は分類されたクラスタである複数の Word を持つ．類似した Document は似た Word を持つことになる．

3.2 pLSA による単位動作認識

pLSA は共起データの解析に用いられる手法である．動作モデルを図.3 に示す．

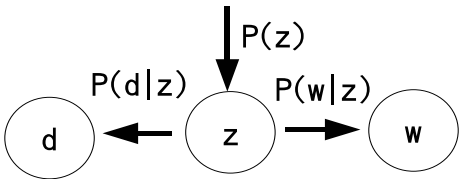

図 3: pLSA 手法の動作モデル

ここでは，映像で発生した Word の共起を基に，単位動作に対する隠れトピックの混合によって映像を表現する．

$$\begin{cases} d_j \in D & = & \{d_1, \cdots, d_N\} \\ w_i \in W & = & \{w_1, \cdots, w_M\} \\ z_k \in Z & = & \{z_1, \cdots, z_K\} \end{cases} \quad (4)$$

人物の動作を含む映像 d と，動作特徴の w，隠れトピックの z を用意する．隠れトピックは観測されてない単位動作に対応する．ここで未観測の単移動 z_k が観測されたとし，観測される $(d_j|w_i)$ のペアがに独立に生成されたとすると，その同時確率は以下の様に表現される．

$$P(w_i|d_j) = \sum_{k=1}^{K} P(d_j|z_k)P(w_i|z_k) \quad (5)$$

$p(d_i|z_k)$ は単位動作 z_k における word d_j の発生確率であり，$p(w_i|z_k)$ は単位動作 z_k における word w_i の発生確率である．

ここで，式 5 を求めるために各パラメータ $P(z), P(w|z), P(d|z)$ を式 6 の対数尺度が最大となる基準を求める．

$$L = \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) log P(w_i, d_j) \quad (6)$$

ここで要素 $n(w_i, d_j)$ は，映像 d_j における word w_i の発生個数であり，これを Expection Maximization(EM) アルゴリズムを用いて解く．

続いて，テスト映像 d_{test} があたえられたとすると，その単位動作 z_k の事後確率は以下のようになる．

$$P(z_k|d_{test}) = \frac{P(w_i|z_k)P(z_k|d_{test})}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_{test})} \quad (7)$$

しかし，$P(z_k|d_{test})$ は観測されてない．そこで同じように逐次型近似アルゴリズム [4] を用いて計算を行う．

その結果得られたものの最大確率となったものがアクションカテゴリとなる．

$$Action\ Category = arg\ max_k\ P(z_k|d_{test}) \quad (8)$$

4 評価実験

本手法の有効性を検証するために，研究室で一人の人物がデスクワークをしている状態を撮影した．撮影した映像は 30 分あり，映像中で席に座る，飲物を飲む，キーボードを打つ，本/資料を読む/めくる，振り向く，席を立つなどのデスク上での日常行動をしている．映像は 320x240 サイズである．

4.1 実験条件

実験を行った時のパラメータを，$\sigma = 13$，$\omega = 4/19$，$\tau = 19$，クラスタリング数を 800，pLSA の単位動作数を 13 として実験を行った．

4.2 実験結果

席に座る動作と，書類のページをめくる動作，眼鏡を外す, 目を擦る動作，ヘッドフォンを取る, 装着する動作などの各シーンにおける実験結果を図 4 - 7 に示す．図中の四角で囲まれた箇所は抽出された特徴点を示し，中に示されている数字は分類された動作カテゴリを表している．

図 4: 席に座る動作 6 → 7

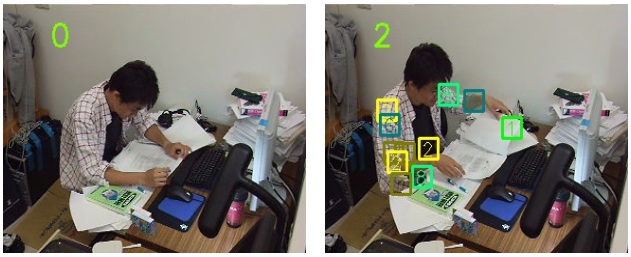

図 5: 書類のページをめくる動作 0 → 2

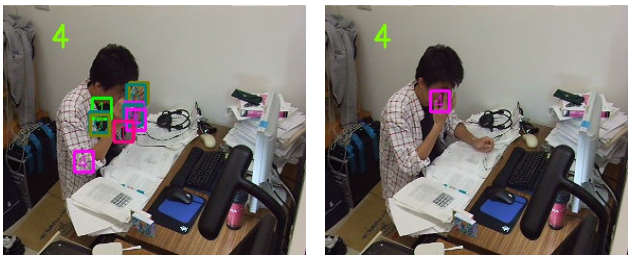

図 6: 眼鏡を外す, 目を擦る動作 4

図 7: ヘッドフォンを取る, 装着する 1 → 5

5 考察

まず，席に人が動画中に出現し座る動作の認識については概ね検出が出来ていると言える．しかし，全体的にノイズが多く目立つ．これは特徴点を得るレスポンス関数 (式 1) から極値を求める際，極値の閾値を低めに取った結果，過剰に特徴点を検出してしまった結果だと考えられる．適切な値を設定する方法については今後の検討課題である．

また誤認識が多い事については，学習データ不足であったと考えられる．今回は学習データとして 1 時間分の動画を用意したが，その映像では十分に学習知識を与えられなかったのではないかと思われる．適切な学習データ数なども今後の検討課題である．

次に眼鏡を取る行為と，目を擦る行為について．これら行動については，顔面付近に手を持っていくという動作が似ている為に，同じ動作として認識してしまったのではないかと考えられる．同様に，ヘッドフォンを装着する行為も誤って同じカテゴリと認識するケースが見られた．今後これらケースに対応する為には，手に持っている物の認識を行うなどをして対処を行う手法が考えられる．

また，予備実験での別環境において撮影したデスクは窓際で日照変化が強く，人の動作箇所でない領域を検出するなど，無視できないレベルでの誤検出が目立った．環境によっては日照変化に強い背景差分法などを用いた上で，本手法を適用する必要があると言える．

6 おわりに

本稿ではカメラ映像から人の行動認識を行う事を目的として，デスクワークを中心とした室内の撮影をしたカメラから，動画中の人物動作の行動認識を行えることを示した．手法には Human Action Categorizing Method を用いて動画中の動作特徴点を検出し，学習と判定には pLSA 手法を用いて，その動作が何であるかの判定を行った．

今後の課題として，K-means クラスタリングの適切なクラスタ数や，pLSA のアクションカテゴリ数の自動設定が挙げられる．また，未学習動作が検出された場合に対して，適切に未知動作の学習/検出が行える手法を今後検討していく予定である．

本手法はマルチアングルに対応する予定であったが実装上の問題で可能に出来なかった．この問題については，カメラの同期を行い，その特徴点は同一動作であるという情報を先に与え，クラスタリングのラベル付けと pLSA での結果推定を行うときに，両特徴点はカテゴリが近いクラスであると判定させる事で解決が出来ると考えている．

参考文献

[1] 西田 佳史, 相澤 洋志, 北村 光司, 堀 俊夫, 柿倉 正義 溝口 博. “センサルームを用いた人の日常活動の頑健な観察とその応用”. 情報処理学会研究報告. HI, ヒューマンインタフェース研究会報告. pp.37-44, no.111, vol.2003, 20031107.

[2] 入江 耕太, 若村 直弘, 梅田 和昇, “ジェスチャ認識に基づくインテリジェントルームの構築”, 日本機械学会論文集 C 編, pp.258-265, num.725, vol.73, 20070125.

[3] Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei. “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words”. International Journal of Computer Vision (IJCV). September, 2008.

[4] Dainiel Gildea and Thomas Hofmann. “Topic-based language models using em”. In In Proceedings of EUROSPEECH, pp. 2167-2170, 1999.