様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１５　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１２０６１

研究課題名（和文）大規模時系列データに対するパターン認識のためのグラフ信号処理基盤

研究課題名（英文）Pattern recognition using graph signal processing for large-scale time-sequence data

研究代表者

篠田　浩一（Koichi, Shinoda）

東京工業大学・情報理工学院・教授

研究者番号：１０３４３０９７

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：RGB-Dカメラ映像を入力として「投げる」「蹴る」などの人間の動作(ジェスチャー)を認識する動作認識において、グラフ信号処理を用いた手法を開発した。この方法では、人体の骨格をグラフとみなし、その時系列を入力とする。各フレームにおいてスペクトルグラフウェーブレット変換を用いて特徴量を抽出し、それらに対し、時系列方向に階層的なプーリングを行う。様々な角度で撮影された動作の認識において従来方法を上回る性能を得た。

研究成果の概要（英文）：We have developed an action recognition method from RGB-D camera inputs. This method uses a time sequence of human skeleton as an input. Every frame it extracts features by using spectral graph wavelet transform. Then the features are pooled in a hierarchical way in the time axis. This method achieved the state-of-the-art in multi-view action recognition.

研究分野：　統計的パターン認識

キーワード：　動作認識　グラフ信号処理　深度カメラ

１．研究開始当初の背景

インターネット上の大量の音声・映像などの時系列データから有用な情報を抽出するための高性能かつ高速なパターン認識技術が求められている。時系列データでは、互いに異なる時間スケールをもつ複数の事象が存在し、それらが相互に関連している。我々はこれまで様々なフレーム間隔で抽出した複数の特徴量を用いる手法を開発し、例えば国際競争型映像検索ワークショップ TRECVID の意味インデクシング部門で 2011,2012 年に世界一位を獲得するなど、実績をあげてきた。しかしながら、この手法を含む従来の短時間フレーム単位での処理は、異なる時間スケールをもつ事象の発見や、それらの間の関係性(構造)を抽出するためには、必ずしも最適ではない。

一方、近年、グラフ信号処理手法の開発が進展している。従来の信号処理では、ユークリッド空間上の格子(グリッド)点に配置された信号を対象としてきたが、信号の表現としてはこれに限らない。グラフ信号処理では、まずデータの構造を反映したグラフを構築し、そのグラフの頂点で観測される信号を対象とする。グラフフーリエ変換、スペクトルグラフウェーブレット変換(SGWT)などの手法が開発され、人口動態や大脳 fMRI 画像の解析に用いられている。

最近、我々は人間のジェスチャー(身振り)認識において、関節を頂点とした骨格グラフに対し SGWT を適用して特徴抽出を行い、より少ない計算量で従来法に迫る性能を得ている。過去にはグラフスペクトルクラスタリングを話者認識に適用した例もある。しかし、それらの応用では、グラフはフレーム単位のデータに対する特徴量空間に張られるものであり、時間軸方向には変化しない。情報抽出のための時系列パターン認識の性能向上のためには、様々な時間スケールにおけるダイナミックな特徴を発見し、それを活用することが重要な課題であるが、現行技術はまだその解決には不十分である。

２．研究の目的

時系列パターン認識における新たな中間表現として、特徴量空間を時間軸方向に拡張した特徴量時空間におけるグラフ中間表現を開発する。そして、そこから有用な情報を抽出するグラフ信号処理手法を開発する。高速・高性能な認識とパターン構造解析の方法論を確立する。

当初は以下の３つの応用、

1) 身体の部位毎の運動の時間スケールの違いを考慮したジェスチャー認識

2) 単語グラフ表現を中間表現として用いた音声認識・検索

3) 音声と動画の相関を考慮した映像グラフを用いたマルチメディアイベント検出

について研究を行う予定であった。しかしながら、1)で予想以上に優れた効果をあげることができた。一方、2),3)については深層学習手法の進展が早く、それに比べた提案手法の優位性を示せなかった。以下では 1)のジェスチャー認識を中心に述べる。最終年度には深層学習と組み合わせる手法も実装・評価した。

３．研究の方法

「投げる」「蹴る」などの人間の動作(ジェスチャー)を認識する手法を開発する。近年、Microsoft Kinect などの RGB-D カメラが安価に入手できるようになった。このカメラは深度情報を用いることで背景から分離された対象物体の座標を獲得することができる。また、人間を対象とする場合には、人体の骨格座標を精度よく推定することができる。しかしながら、隠ぺいなどが原因で、骨格座標の獲得に失敗することもある。そこで、本研究では、まず、この RGB-D カメラから得られる骨格座標の時系列から、動作の種類を判別することを目的とする。そして、骨格座標のエラーに対して頑健な手法を構築する。

従来は、骨格の３次元座標の時系列信号に対し、スパースコーディングや線形判別など、画像認識で一般的に用いられる手法を応用した研究が主流であった。この種の手法は、識別モデルを用いており、人間の骨格の生成モデル（どの関節とどの関節がつながっているか等）を陽には用いていない。そのため、未知の変動、例えば、「見え」(撮像角度)の違いに対して頑健ではなかった。そこで、我々は、まず、骨格座標の「見え」の正規化を行った上で、グラフ信号処理の手法を用いて、「見え」に対して不変であり、かつ、関節間の関係をよりよく表現する特徴を抽出する。判別においては、少量のデータ量でも頑健に動作するサポートベクターマシンを用いた。以下、順番に説明をする。

まず、撮像角度の違う骨格画像において、骨格座標の身体の各部分の関係を用いて身体の向きを推定し、一つの座標系に変換する。ここでは、まず頭と胴体の位置座標を用いて垂直方向を推定し、胸と右手の位置座標を用いて水平面における身体の向きを推定する。各々の動作の映像におけるすべての画像フレームを用いて統計的に推定することで頑健な推定が可能である。

次に、骨格をグラフと見なしてグラフ信号処理を適用する。ここで、骨格を構成する関節をノードとし、実際に骨で連結されている関節間にエッジを置く、エッジの重みには関節間の距離の Radial Basis Function (RBF) の出力値を用いる。ここでは、各関節の３次元位置を入力として、スペクトルグラフウェ

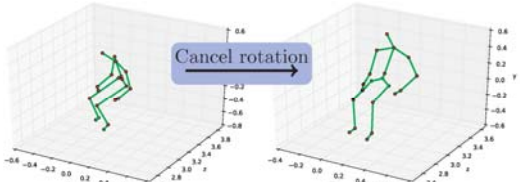

図１ 身体の向きの正規化

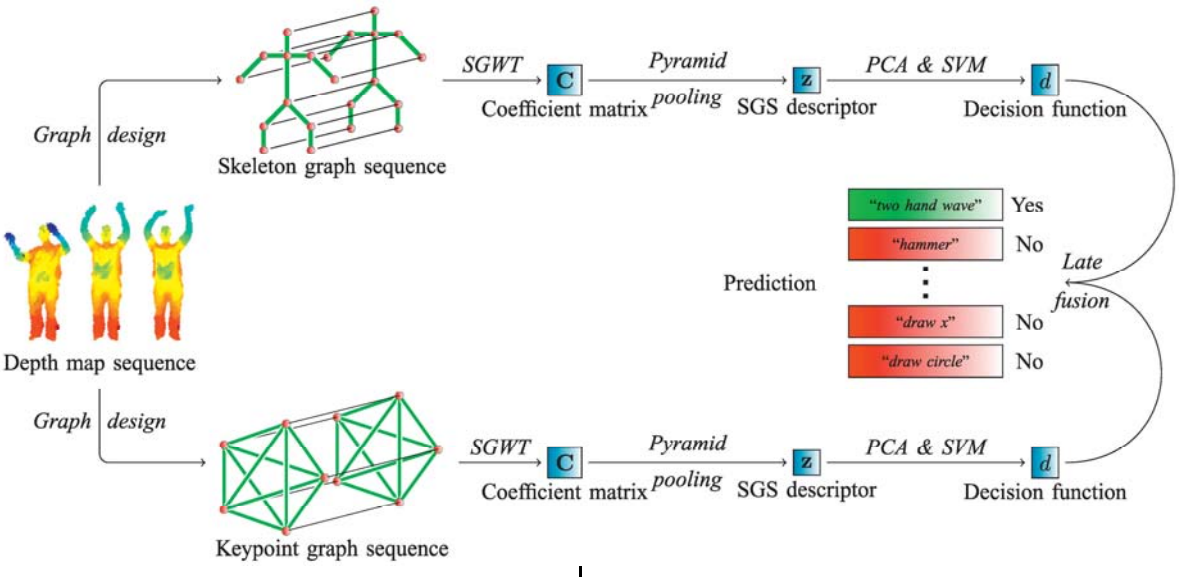

図 2 グラフ信号処理を用いた動作認識手法

ーブレット変換(Spectral Graph Wavelet Transform; SGWT)を適用することで、そこからグラフ特徴量を抽出する。SGWT は、一般のユークリッド空間におけるウェーブレット分析を、グラフ空間に適用するために改変したものであり、グラフ信号処理における代表的な手法である。

そして、各フレームの特徴量を時間軸方向にプーリングする。ここでは平均プーリングを用いた。その際、時間軸方向に階層的なプーリングを用いることにより、時系列の情報も表現する特徴量を得る。そして得られた特徴量をサポートベクターマシンで判別する。

また、RGB-D カメラから骨格を得る際に、隠ぺいなどが原因で、誤りが生じることがある。その影響を軽減するために RGB-D 画像(ポイントクラウド)のクラスタリングを行い、クラスタ中心を結んだグラフを用いて上述と同様の処理を行う。そして、最後にサポートベクターマシンの出力するスコアを足し合わせることで、骨格データを用いた識別との融合を行う(Late fusion)。

さらに、最終年度では骨格グラフの特徴量を入力としたディープラーニング(深層学習)手法を開発した。ここでは画像などで用いられる一般的な構造の畳み込みネットワークの代わりに、骨格の特徴に依存した構造の畳み込みニューラルネットワークを用いる。人体における関節の接続関係が、ニューラルネットワークにおけるノード間の接続関係に反映されるように設計されている。

４．研究成果

いくつかの公開データセットを用いて評価を行い良好な結果を得た。

まず、撮像角度を固定したデータベース MSRAction3D データセットでは、91%の識別率であった。識別的な特徴量を用いない手法の中では最高の認識性能であった。識別的な特徴量と併用することで、更なる性能向上が見込める。

次に、「見え」の角度の異なるデータを含む N-UCLA Multi-view Action3D データセットを用いて評価を行った。従来手法の認識率が 75.8%のところ、90.8%とはるかに上回る性能を達成した。

また、人間の骨格構造を反映したディープラーニング手法も従来方法とほぼ同等の性能を得ることができた(2018 年 6 月現在投稿中、http://arxiv.org/abs/1805.11790)。

本成果は、人間のジェスチャーの認識に対してグラフ信号処理を用いた有効な特徴量を提案するものである。また、骨格情報のみならず、様々なグラフに対して適用が可能である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

Tommi Kerola, Nakamasa Inoue, Koichi Shinoda, “Cross-View Human Action Recognition from Depth Maps Using Spectral Graph Sequences”, Elsevier Journal of Computer Vision and Image Understanding (CVIU), vol. 154, pp. 108-126, Jan. 1, 2017. (査読有)

〔学会発表〕（計　1　件）

Tommi Kerola, Nakamasa Inoue, Koichi Shinoda, “Graph Regularized Implicit Pose for 3D Human Action Recognition”, Proc. APSIPA, pp. 155-159, Dec. 12, 2016. (査読有)

〔図書〕（計　1　件）

篠田浩一, 「音声認識(機械学習プロフェッ

ショナルシリーズ)」,講談社, Dec. 9, 2017, 165.

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等
http://www.ks.cs.titech.ac.jp/japanese/index.html

６．研究組織
(1)研究代表者
篠田　浩一（SHINODA, Koichi）
東京工業大学・情報理工学院・教授
研究者番号：10343097

(2)研究分担者
井上 中順（INOUE, Nakamasa）
東京工業大学・情報理工学院・助教
研究者番号： 10733397