氏　　名	TRAN THANG THANH		
学位の種類	博士(情報科学)		
学位記番号	博情第 308 号		
学位授与年月日	平成 26 年 9 月 24 日		
論文題目	Automatic Extraction of Discriminative Semantic Features for Action Recognition（人の動作認識のための可分な意味特徴の自動抽出に関する研究）		
論文審査委員	主査	小谷　一孔	北陸先端科学技術大学院大学　准教授
		党　　建武	同　教授
		田中　宏和	同　准教授
		Le, Bac Hoai	同　客員准教授
		阿部　　亨	東北大学　准教授

論文の内容の要旨

In our everyday lives, we are constantly confronted with human motion: we watch other people as they move, and we perceive our own movements. Motion, in terms of physics, is a change in the position of a body with respect to time. Since the human body is not simply a rigid block but rather a complex aggregation of flexibly connected limbs and body parts, human motion can have a very complex spatial-temporal structure. Deeper understanding on human actions is required in many applications, e.g., action recognition (security), animation (sport, 3D cartoon movies and virtual world), etc.

With the development of the technology like 3D specialized markers, we could capture the moving signals from marker joints and create a huge set of 3D action motion capture (MOCAP) data which is capable of accurately digitizing a motion's spatial-temporal structure for further processing on a computer. Recently, motion capture data have become publicly available on a larger scale, e.g. CMU, HDM05. The task of automatic extraction of semantic action features is gaining in importance. The underlying questions are how to measure similarity between motions or how to compare motions in a meaningful way. The main problem is that the granularity of MOCAP data is too fine for our purpose: Human actions typically exhibit global and local temporal deformation, i.e. different movement speed and timing difference. The similar types of motions may exhibit significant spatial and temporal variations. The irrelevant details (like noise) as well as spatial pose deformations may interfere with the actual semantics that we are trying to capture. The problems require the identification and extraction of logically related motions scattered within the data set. This leads us to the field of motion analysis for identifying the significant spatial and temporal features of an action.

To automatically extract the action features from 3D MOCAP data, we proposed two approaches dealing at features levels: 1) Extract of Discriminate Patterns from Skeleton Sequences approach provides a foundation in lower dimensional representation for the movement sequence analysis, retrieval, identification and synthesis; and 2) Automatic Extraction of Semantic Action Features approach which focuses on solving the high-dimensional computational problems arising from the human motion sequences. They support the follow-up stages of processing the human movement on a natural language level. As one common underlying concept, the proposed approaches contain a retrieval component for extracting the above-mentioned features.

Firstly, we extract the discriminative patterns as local features and the utilization of a statistical approach in text classification to recognize actions. In text classification, documents are presented as vectors where each component is associated to a particular word from the code book. Traditional weighting methods like Term Frequency Inverse Document Frequency (TF-IDF) are used to estimate the importance of each word in the document. In this approach, we use the beyond TF-IDF weighting method to extract discrimination patterns which obtain a set of characteristics that remain relatively constant to separate different categories. This weighting defines the importance of words in representing specific categories of documents. It not only reduces the number of feature dimension compared to the original 3D sequence of skeletons, but also reduces the viewing time of browsing, bandwidth, and computational requirement of retrieval.

Secondly, we propose the semantic annotation approach of the human motion capture data and use the Relational Feature concept to automatically extract a set of action features. For each action class, we propose a statistical method to extract the common sets. The features extracted is used to recognize the action in real-time. We extract the set of action features automatically based on the velocity feature of body joints. We consider this set as action spatial information. We combine both spatial and temporal processes to extract the action features and use them for action recognition. In our experiments, we use the 3D motion capture database HDM05 for performance validation. With few training samples, our experiment shows that the features extracted by this method achieve high accuracy in recognizing actions on testing data. Our proposed method gets high accuracy comparing to others state-of-art approaches.

Keywords

Discriminate Semantic Features; Automatic Extraction Features; Semantic Action Features; Action Recognition; Joint Velocity.

論文審査の結果の要旨

本論文は、人の動きの 3 次元モーションキャプチャデータを用いて人物動作の解析、認識を行う次の２つのアプローチを提案している。

第一のアプローチ（EDPSS）では予め学習データ系列中の特徴的なパターンを抽出し、これをプロトタイプ（学習データにおける特徴ベクトル）と見なして入力データ系列内に当てはめ、該当するプロトタイプのインデックスの系列を人物動作と見なして認識結果を与えるものである。この手法はデータ列（コンテンツ）の内容に関する事前知識がないときに有効なボトムアップ的なアプローチとみなせる。

第二のアプローチ（AESAF）は予め想定した人物動作の典型的なパターンをモデル化しておき、更に動作に意味（右足でキック、左腕のパンチ…、など）を持たせておいて、各動作の遷移に制約を与えることで複雑な人物動作を安定的に認識している。この手法は解析したい、あるいは抽出したい具体的な動作を予め指定できる場合に有効なトップダウン的なアプローチと見なせる。

これら２つのアプローチは状況によって使い分けが可能であるが、AESAF は予め人物動作に関する知識に基づいた semantic base のアプローチであるため、EDPSS よりも高い動作認識精度が期待できる。特に人物動作に意味を与えて認識できるので、単に人の動作だけでなく行動や意図までも推定可能となろう。また、これらは 3 次元モーションキャプチャデータを対象にしていることから従来の歩行やジェスチャのような 2 次元に投影した動作認識に比べてオクルージョンの影響などを受けにくく、複雑な動作も認識可能である。

以上、本論文は人の体の動きを各部位の変位としてではなく、意味ある動作としてモデル化し、 3 次元空間上の複雑な動作を安定的に認識できる新しい手法を提案しており、学術的に貢献するところが大きい。更に、提案手法は人の動作解析やセキュリティ、Digital Signage、CG 等に適用が可能であり、高い実用性も認められる。よって博士（情報科学）の学位論文として充分価値あるものと認めた。