博士（情報科学）　卢　　国　梁

学位論文題名

# Vision-Based Action Recognition for Single Users

## （コンピュータビジョンに基づく単一人物の行動認識に関する研究）

## 学位論文内容の要旨

Vision-based human activity analysis has attracted much attention in the past several decades, which is the process of understanding human action/motion patterns in given video sequences with or sometimes without prior knowledge of probably performed actions/motions. In this thesis, driven by the natural demands in single-person oriented applications, we focus on recognizing actions performed by single users, and do not explicitly consider interactions with other persons or surrounding objects, and context, e.g. the cluttered/crowded environment. Moreover, we concentrate on monocular videos excluding multi-views videos, considering simpleness and low-cost in a wide range of practical usage.

In real-world applications, especially in on-line processing, time saving is heavily concerned as well as other significant characteristics, e.g., recognition rate and invariance of camera-view. As contributions of this study, two approaches/techniques have been proposed separately to save the processing time under the premise of keeping comparable recognition accuracy.

The thesis is structured as follows. A brief introduction of this study is presented in Chapter 1, and then the detailed presentations of two proposed approaches are given respectively in Part I and Part II as the main parts of this thesis. Each part starts with an introduction of the domain (Chapter 2 and Chapter 6), followed by practical implementation (Chapter 3 and Chapter 7) and then experimental investigation (Chapter 4 and Chapter 8). In Chapter 5, especially, we present an approach of temporal segmentation and assignment of successive actions, taking advantages of selected frames by the method shown in Chapter 3. Finally, in Part III, we summarized the thesis and show the future work.

On the basis of the assumption that basic actions can be discriminated by only a few representative frames taken from an entire video, a new martingale-test based framework was proposed for selecting representative frames (called *characteristic frames*). Then we employed an exemplar-matching based recognition method to exploit the selected characteristic frames. Experimental results on the public Weizmann dataset showed that the method using characteristic frames was carried out remarkably faster than the ones using the entire frames, while keeping comparable performance in recognition rate. In addition, taking advantage of time-differentiated information brought from characteristic frame, we proposed a novel framework for temporal segmentation and assignment of successive actions in a long-term video. Through the experiments, we obtained the assignment accuracy of 80.5% at frame level on IXMAS dataset, in computation time of 1.57 *s* per video which has 1160 frames on the average, which was two orders of magnitude faster than the compared method.

Motivated by recent efforts to bypass a time-consuming step of subject location in action recognition, i.e., subject detection followed by tracking from frame to frame, we also proposed a novel algorithm that extracts action patterns directly from difference images in terms of temporal self-similarity (TSS).

We then employed *a bag-of-words (BoW)* framework to assemble these extracted patterns for action recognition. Experimental results on two public datasets of the Weizmann dataset and the KTH dataset showed that this framework is useful for reducing the processing time without degrading the performance comparable to those of the *state-of-the-art* approaches.

We have used publicly available datasets for evaluation. Accordingly, we could measure the degree of improvement objectively and fairly. On the other hand, the general performance of the proposed approaches has not fully revealed. Therefore, in the future work, it is necessary to confirm the generalization ability of the proposed approaches and to improve the performance and stability. In addition, we notice that the usage of 2D images/videos for action recognition is limited in several points, for example, largely influenced by noise, scattered environment, illuminations and many factors. We could use devices other than video cameras, e.g., low-cost depth sensors such as Kinect. As a result, there are many more challenging issues to be progressed.

In summary, our contributions are described in two folds. First, the introduced martingale-test based frame-selection framework can be executed without requiring any prior knowledge about the performed actions and the selected frames by this framework can give the basic description of performed actions. Second, the proposed method for extracting action patterns directly from difference images is efficient since it does not need a time-consuming pre-processing of finding bounding boxes of human body, and meanwhile, it inherits general properties of TSS. These two techniques would contribute to the development of the field of action recognition.

# 学位論文審査の要旨

主　査　　教　授　　工　藤　峰　一
副　査　　教　授　　今　井　英　幸
副　査　　教　授　　金　子　俊　一

## 学位論文題名

# Vision-Based Action Recognition for Single Users

## (コンピュータビジョンに基づく単一人物の行動認識に関する研究)

ビデオカメラを使った行動解析が近年注目されている. 特に, 独居老人の見守りや侵入者の検知と見分け, ジェスチャーを用いたコンピュータとのインタラクションなど, 需要は益々増えている. 複数台のカメラを使った研究が多いものの本研究ではコストや設置の容易さを優先して一台のカメラのみを想定している. また, 基礎的な方法論の発展を意図してある程度離れたところから観察された一人による行動の分類に限定している.

典型的な行動分類の処理は以下の通りである. 前処理によりフレーム毎に雑音を除去した後, 時間差分や輪郭線追跡などにより (人として想定される) 動く物体が写っている領域を切り出す. 続いて, その領域において局所特徴量など行動識別に有効な特徴量を求める. 最後に, 特徴量の値を識別器に渡して分類結果を得る. 識別はフレーム毎あるいは一つの行動全体を含むフレームシーケンス毎に行われる. これにより識別方法も二つに分かれる. フレーム毎に予め各行動の典型例としてとってあったフレームと比較する "テンプレートマッチング" と, シーケンスにおける各フレーム間の類似性を表す行列を作り, 行列から新たな特徴量を生成して SVM などの典型的な識別器を利用して識別するものである.

本研究はこれらの二つの基本的方法のそれぞれにおいて時間的改良を試みていおり, それらが二部形式で記述されている. 第 I 部ではテンプレートマッチングの時間削減法を述べている. ここでは, 隣接したフレームの類似性に着目して, 複数の行動を含むシーケンスから行動が切替わった瞬間のフレームだけを切出して識別に利用することを提案している. 切替り判定は一般に難く, そのために本研究ではマルチンゲールといわれる特殊な時系列を用いた異常検出手法を提案している. 結果として, 認識率の低下をほとんど招かずに識別に用いるフレーム数をほぼ 10 分の 1 まで減らすことができている.

第 II 部では, 一つの行動のみを含むビデオシーケンスの認識方法を検討している. 通常, 前処理として行われる人領域の切り出しやシルエットの切出しが計算量的に重い. そこで, 本研究では, この前処理を省略する方法を提案している. 具体的には, フレーム間の類似性を表す TSS(Temporal Self-Similarity) 行列の時間計算量がフレーム長 T に関して二乗オーダであるため, この計算を t(< T) フレームに局所化して高速化を図っている. さらに, より高速に計算できる類似度を提案している. 結果として,1 桁程度の時間削減を行うことに成功している. 副次的な効果として, 不完全な前処理による悪影響を避けることができたため認識性能も向上している.

本論文では, 第 1 章で問題背景を述べた後三部構成で成果をまとめている. 第 I 部において対象と

する認識方法を説明した後にこれまでの研究動向を第 2 章にまとめている. 第 3 章で提案手法の基礎となるマルチンゲールを使った異常検出方法を説明している. 第 4 章で, この異常検出方法をフレームシーケンスの切り出し, ならびに, 各行動の“代表的フレーム”の選出に利用している. 第 5 章では, 代表フレームを利用した連続行動の認識を行いその有効性を実験的に検証している. 第 II 部に移り, 第 6 章で, 改めて土台となるもう一つの方式の説明した後, 第 7 章で TTS を利用した一般的な認識手法を紹介し, 続いて局所 TSS を提案している. 第 8 章において実験的評価を行っている. 第 III 部 (第 9 章) では, これらの検討を総括するとともに, 今後の改良に向けた研究の方向性を議論している.

本論文による成果は以下にまとめられる.

1. 複数の動作が含まれるビデオシーケンスを動作毎に切り分けるオンライン方式としてマルチンゲールを使った方式を提案した. これにより, 動作毎に代表的なフレームを選び出すことができるようになり, 認識時間の大幅な時間削減に成功した.
2. 単一行動のみを含むフレームシーケンスを認識する方法論において, 典型的に用いられてきた計算量の重い前処理を省く方式を提案して時間的な効率化を図った. これは前処理に依存した悪影響を除去することにもつながり, 認識率の向上を果たすこともできた.

これを要するに, 著者は, 行動認識をビデオにより行う伝統的な二つの方式それぞれにに対して時間的な改善方法を提案することでそれらの適用範囲を飛躍的に向上させた. また, 識別性能向上に関する貢献も大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.