氏名	NGUYEN THUY NGOC			
学位の種類	博士(情報科学)			
学位記番号	博情第 350 号			
学位授与年月日	平成 28 年 9 月 23 日			
論文題目	Feature Extraction and Human-Human Interaction Recognition for Video Surveillance			
論文審査委員	主査	吉高淳夫	北陸先端科学技術大学院大学	准教授
		赤木正人	北陸先端科学技術大学院大学	教授
		金井秀明	北陸先端科学技術大学院大学	准教授
		田中宏和	北陸先端科学技術大学院大学	准教授
		平川正人	島根大学	教授

論文の内容の要旨

Human interaction recognition has been widely studied because it has great scientific importance and many potential practical applications. However, this problem is very challenging especially in realistic environments where background is dynamic and has varying lighting conditions. This dissertation addresses human activity recognition, especially human-human interactions in realistic video material, such as movies, surveillance videos. For classification problem, most existing methods rely on either spatio-temporal local features (i.e. SIFT) or human poses, or human joints to model human interactions. As a result, they are not fully unsupervised processes because they require either hand-designed features or human detection results.

Motivated by the recent success of deep learning networks, we investigate a three-layer convolutional network which uses the Independent Subspace Analysis (ISA) algorithm to learn hierarchical invariant features. The ISA algorithm is a generalization of the Independent Component Analysis (ICA), which is very well-known in natural image statistics. Compared to the ICA algorithm, the most notable advantage of the ISA is that it can learn features which are invariant to phase while being selective to orientation and frequency. However, the ISA algorithm becomes slow when applying it on video data. In order to solve this computational problem, we combine the idea of convolutional neural network with the ISA algorithm. Specifically, instead of training the ISA algorithm directly on raw video data, we first train it on small video blocks extracted by our procedure. The obtained features are then convolved with larger video blocks. The outputs of this convolution step are fed into the next layer, which is implemented by another ISA algorithm. This organization enables the three-layer convolutional ISA network to learn hierarchical invariant features. Furthermore, we introduce a pooling layer to reduce the contributions of features learned in lower layers while still achieving translation invariant. Using the invariant features learned by the three-layer convolutional ISA network, we build a bag-of-features representation for videos. Finally, we apply Support

Vector Machine (SVM) to classify human interactions. For temporal localization, we slide temporal detection windows with different durations over a continuous video sequence with a stride of 10 frames. For each temporal window, our convolutional ISA network extracts hierarchical invariant features on a dense grid. After scoring the temporal detection windows, a non-maximum suppression is applied to enforce that none of the retained windows are overlapping.

In all two cases, we conduced thorough experiments on realistic videos from challenging benchmarks used by activity recognition community. We show that our three-layer convolutional ISA network is effective to represent complex activities such as human interactions in realistic environments. Besides, we believe that our temporal localization method is the first work which reports experimental results on the continuous video sequences of human interactions. Although temporal localization results are insufficient for real applications, it is a first step for further research in localization of human interactions.

Keywords: temporal localization, classification, independent subspace analysis, human-human interactions, convolutional network, pooling

論文審査の結果の要旨

安全安心な社会環境の維持を支える技術の 1 つとして，公共の場等に設置したカメラによる監視システムが挙げられる. 近年の監視カメラ設置台数の増加などの理由により人手による監視には限界があり，計算機による自動認識技術の確立が待たれるところである. 監視対象としては人，自動車などが挙げられるが，Nguyen Thuy Ngoc 君の研究は人を対象とし，個人の行動認識と比較してより複雑であり，かつ自動認識の需要が高いと考えられる人物間の相互動作の認識に取り組んだものである. 監視カメラにおける検出対象となる相互動作の例としては，一方が他者を突き飛ばす，殴るなどが挙げられる.

人物間の相互動作認識に関する既存研究はいくつか挙げられる. 既存手法の多くは画像特徴として単純な背景中に数人の人物が含まれるような状況を検出対象のデータセットとしたものであり，画像中の人物やその動作の検出が比較的容易であるが，現実的な環境とは必ずしも言えないものである. さらに，既存研究は特定のデータセットに対して比較的高い認識精度が得られるよう，想定する特定のデータセットに対して認識アルゴリズムが特化されたものであり，別の画像的特徴を持ったデータセットに同アルゴリズムを適用した場合には認識精度が大きく低下する点が課題であった. さらには，既存研究で用いられるデータセットは 1 動作ごとに分割された映像の集合であり，現実の監視カメラ映像処理では動作ごとに分割されていない連続動画であることを考えると，実用化のために不可欠な動作区間検出の問題を扱っていなかった.

それに対して，本研究では，独立部分空間分析法を用いた 3 層畳み込みニューラルネットワークにより学習した人物間相互動作に関する特徴量に対してサポートベクタマシンにより各動

作の判別を行う手法を提案した．提案手法は既存手法と比して，異なる画像特徴を持つ複数のデータセットに対して総合的に良好な識別精度を実現した．さらに，人物間の相互動作が生じている映像区間を検出する手法として，複数時間長の窓をシフトさせ，その窓内で識別対象とする動作が検出されるか否かを判定することによる，相互動作生起区間同定法を提案した．

以上、本論文は、様々な条件下における映像中の人物間の相互動作の検出，識別について従来法と比較してデータセットの特性に依存しないより安定した手法，並びに相互動作生起区間の検出手法を確立したものであり，学術的に貢献するところが大きい．よって博士（情報科学）の学位論文として十分価値あるものと認めた．