論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Graph Spectral Approach to Human Action Recognition from Depth Maps
著者(和文)	KerolaTommi Matias
Author(English)	Tommi Matias Kerola
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第10383号, 授与年月日:2016年12月31日, 学位の種別:課程博士, 審査員:篠田 浩一,徳永 健伸,小池 英樹,藤井 敦,下坂 正倫
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第10383号, Conferred date:2016/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	Computer Science 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 (Philosophy) Doctor of
学生氏名： Student's Name	KEROLA Tommi Matias	指導教員（主）： Academic Advisor(main)	SHINODA Koichi
		指導教員（副）： Academic Advisor(sub)	

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

Chapter 1 briefly introduces the task of human action recognition and motivates why depth cameras and graph signal processing are suitable tools for solving the problem. Recent work is also discussed, together with their limitations. Finally, an overview of the proposed action recognition system is given, together with a summary of the contributions of this work. The contributions are: (1) a framework for human action recognition using graph signal processing, (2) a simple but efficient rotation cancellation method based on Gram-Schmidt orthonormalization, (3) a view-invariant graph-based action representation that significantly outperforms previous methods in cross-view action recognition, and (4) an efficient algorithm for calculating the spectral graph wavelet transform that takes advantage of the explicit sparsity structure of our graph.

Chapter 2 introduces the task of human action recognition and limitations of this study. A general framework for human action recognition is also presented, with an overview of terminology and standard techniques used for solving the task. Applications of human action recognition to real-world problems is also discussed. Finally, a survey of related research is presented.

Chapter 3 introduces the theory of graph signal processing, including terminology and standard techniques that will be used in the following chapters for creating our proposed action recognition system. Specifically, the graph Laplacian matrix, the graph Fourier transform, and the spectral graph wavelet transform (SGWT) are discussed. Further, the validity of the presented graph signal processing techniques is motivated. Finally, previous research in graph signal processing is surveyed.

Chapter 4 discusses about how to represent human actions as graphs, where two different view-invariant candidates for graph construction are considered. The first candidate is based on tracked skeleton joints, while the second variant is based on spatio-temporal keypoints. Graphs based on skeleton joints capture the spatial pose of the human body, which is suitable for representing actions that are defined by larger general limb movements, where the semantic knowledge of body part positions is vital for recognition. Spatio-temporal keypoints, on the other hand, capture complementary detailed information directly from the point cloud. Each keypoint describes the spatio-temporal shape of a point cloud, and is thus able to capture fine intrinsic detail, while also being robust against noisy skeleton estimates, which can be caused by complex poses.

Chapter 5 presents our feature descriptor called Spectral Graph Sequences (SGS) as well as our proposed action recognition system. Our system consists of five parts. First, we design an augmented graph by connecting together a sequence of graphs using temporal edges. Each graph in the sequence describes the point cloud in a single frame. Second, spectral graph wavelet coefficients are calculated using the SGWT. The coefficients capture second order gradient information about the graph signal along both temporal and local edge directions. Third, in order to cope with varying action sequence length, we leverage a temporal pyramid pooling scheme. The pooling operator aggregates information about the wavelet coefficients, while the pyramid structure allows us to capture the temporal order of the graph signal propagation. Fourth, we reduce the dimensionality of the feature vector using PCA and apply a standard SVM for classification. Finally, using late fusion of SVM decision functions, we can also combine the complementary effects of several graph types. In addition, the end of the chapter presents some analysis of the interpretation and effects of the proposed feature descriptor.

Chapter 6 presents experimental evaluation of the proposed system on five publicly available datasets:

MSRAction3D, MSRActionPairs3D, UCF-Kinect, N- UCLA Multiview Action3D and Online RGBD Action. In addition to the experiments, the end of the chapter provides some analysis of the rotation cancellation scheme, keypoint locations, and parameters of the method. MSRAction3D is a standard benchmark dataset for 3D action recognition, which has remained a challenging dataset due to high inter-class similarities between actions. For testing the ability of our method to capture temporal directionality of actions, we turn to the MSRActionPairs3D dataset, which consists of pairs of actions that differ only in the direction that the action is performed. UCF-Kinect is a dataset that contains actions suitable for interactive movements used in games. The N-UCLA Multiview Action3D dataset is quite different from the previous three, as it was captured with three different camera angles, which drastically changes the appearance of the actions, requiring the usage of features that are invariant across different views. Finally, the Online RGBD Action dataset aims to evaluate human-object interaction, and contains several action types that differ in the type of object interacted with.
Experiments on these datasets using both skeleton-based and keypoint-based graphs show the efficiency of our method. In particular, skeleton-based graphs work well for interactive actions due to the semantic labeling of the skeleton joints to body parts, whereas keypoint-based graphs show their strength in capturing complementary information to the skeleton joints, as it provides a feature that captures the spatio-temporal shape of the depth map point cloud.

Chapter 7 summarizes the dissertation and lists the contributions of the thesis. Unsolved problems and directions for future work are also discussed. In conclusion, skeleton-based graphs worked well for the recognition of both frontal- and cross-view actions due to them capturing second-order information together with a semantic labeling of the skeleton joints. Keypoint-based graphs, on the other hand, were suitable for cross-view action recognition and human-object interaction. Combining both methods through late fusion showed that they are complementary, leading to a significant improvement in recognition accuracy for cross-view action recognition.

Future work should focus on the analysis of the effects of the SGWT kernels, other interest point types, as well as the recognition of more complex behavior, such as human-human interaction.