Expert Systems With Applications 45 (2016) 131–141

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system

Hanguen Kim [a], Sangwon Lee [a], Youngjae Kim [a], Serin Lee [b], Dongsung Lee [c], Jinsun Ju [c], Hyun Myung [d,*]

[a] Urban Robotics Laboratory (URL), Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro (373–1 Guseong-dong), Yuseong-gu, Daejeon 305-701, Republic of Korea
[b] Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore
[c] Image & Video Research Group, Samsung S1 Cooperation, 168 S1 Building, Soonhwa-dong, Joong-gu, Seoul 100-773, Republic of Korea
[d] Director of Urban Robotics Lab. and a professor in the robotics program and Deptartment of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro (373–1 Guseong-dong), Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

*Keywords:*
Video-surveillance system
Human joint estimation
Behavior recognition
Human-computer interaction (HCI)

ABSTRACT

Recent advances in 3D depth sensors have created many opportunities for security, surveillance, and entertainment. The 3D depth sensors provide more powerful monitoring systems for dangerous situations irrespective of lighting conditions in buildings or production facilities. To robustly recognize emergency actions or hazardous situations of workers at a production facility, we present human joint estimation and behavior recognition algorithms that solely use depth information in this paper. To estimate human joints on a low cost computing platform, we propose a human joint estimation algorithm that integrates a geodesic graph and a support vector machine (SVM). The human feature points are extracted within a range of geodesic distance from a geodesic graph. The geodesic graph is used for optimizing the estimation result. The SVM-based human joint estimator uses randomly selected human features to reduce computation. Body parts that typically involve many motions are then estimated by the geodesic distance value. The proposed algorithm can work for any human without calibration, and thus the system can be used with any subject immediately even with a low cost computing platform. In the case of the behavior recognition algorithm, the algorithm should have a simple behavior registration process, and it also should be robust to environmental changes. To meet these goals, we propose a template matching-based behavior recognition algorithm. Our method creates a behavior template set that consists of weighted human joint data with scale and rotation invariant properties. A single behavior template consists of the joint information that is estimated per frame. Additionally, we propose adaptive template rejection and a sliding window filter to prevent misrecognition between similar behaviors. The human joint estimation and behavior recognition algorithms are evaluated individually through several experiments and the performance is proven through a comparison with other algorithms. The experimental results show that our method performs well and is applicable in real environments.



## 1. Introduction

In recent years, 3D depth information-based human behavior recognition with human joints has become an important topic in the areas of human computer interaction (Aggarwal & Xia, 2014). Recent advances in 3D depth sensors such as Microsoft Kinect have meanwhile created many opportunities for security, surveillance, and entertainment (Zhang, 2012). Among these applications, the video-surveillance area has been extensively studied. To maintain the security of both people and infrastructure, new technologies are contributing to the realization of more powerful systems that detect dangerous situations (Castro, Delgado, Medina, & Ruiz-Lozano, 2011). In order to detect dangerous situations irrespective of lighting conditions in buildings or production facilities, researchers have extensively attempted to use depth information from 3D sensors. Kinect is a low-cost 3D depth sensor based on structured light technology but is limited to indoor use (Freedman, Shpunt, Machline, Arieli et al., 2008). A structured light sensor infers the depth at any location by

* Corresponding author. Tel.: +82 42 350 3630; fax: +82 42 350 3610.
*E-mail addresses:* sskhk05@kaist.ac.kr (H. Kim), lsw618@gmail.com (S. Lee), david-kim@kaist.ac.kr (Y. Kim), serin.lee@gmail.com (S. Lee), dslee.lee@samsung.com (D. Lee), jinsun.ju@samsung.com (J. Ju), hmyung@kaist.ac.kr (H. Myung).

http://dx.doi.org/10.1016/j.eswa.2015.09.035
0957-4174/

projecting a known infrared light pattern onto a scene and evaluating the distortion of the projected pattern. In spite of these weaknesses, there are potential avenues for improving the current methods of human behavior recognition for surveillance systems (Escalera, 2012; Ren, Yuan, Meng, & Zhang, 2013; Schwarz, Mkhitaryan, Mateus, & Navab, 2012).

In order to robustly recognize emergency actions or dangerous situations of workers at a production facility, the authors develop a human joint-based behavior recognition algorithm that uses depth information only. The human joint estimation algorithm can recognize human behavior easily even in complex environments such as offices and factories. Han, Shao, Xu, and Shotton (2013) briefly introduced the recent developments in Kinect sensor-based technologies including human joint estimation and behavior recognition algorithms. In the case of human joint estimation algorithms, the traditional methods can be classified into model-based and model-free algorithms, depending upon whether a priori information about the object shape is employed (Poppe, 2007). There is a vast body of research on human joint estimation and the area has been surveyed by Escalera (2012), Poppe (2007), Moeslund, Hilton, and Krüger (2006), and Shotton et al. (2013). Currently, human joint estimation methods can be classified into three categories. The first category is graph-based approaches: most graph-based approaches use a geodesic distance for the graph generation. The geodesic distance is calculated along the graph node as opposed to the Euclidean distance, which does not use graph information (Alpaydin, 2004). If the graph representation method is used, it is easy to represent the 3D information. It is also possible to reduce the noise of the 3D data by using various optimization techniques. Plagemann, Ganapathi, Koller, and Thrun (2010) detected points of interest, based on identifying the maximum geodesic distance value on a 3D point cloud mesh, and they coincide with salient points of the body that can be classified, such as the hands, feet, and head, using local shape descriptors. Visutsak and Prachumrak (2011) generated a human joint of a 3D meshed model in a Riemannian space, based on Blum's medial axis transform and geodesic distance algorithm. Schwarz et al. (2012) proposed a full-body joint estimation algorithm that robustly detects anatomical landmarks in a geodesic graph using depth information and fits a skeleton body model using constrained inverse kinematics. Another graph-based algorithm uses a skeletal graph extracted from a volumetric representation of the human body (Straka, Hauswiesner, Rüther, & Bischof, 2011). The skeletal graph is a tree that has the same topology as the human body (arms, legs, and body). The second category is machine learning-based approaches: multi-class problems such as human joint estimation can be solved effectively by using various machine learning algorithms. After the publication of Shotton et al. (2011), several studies extended this work or focused on efficient use of parallel processing. Rogez, Rihan, Orrite-Uruñuela, and Torr (2012) proposed a multi-class joint detector that uses random forests that can classify joints based on histograms of orientated gradient features. Random forests are a combination of decision tree predictors (Breiman, 2001). Hernández-Vela et al. (2012) extended the work of Shotton et al. (2011) using graph-cut optimization. The graph-cut is an energy minimization framework, and it has been widely applied in image segmentation. Buys et al. (2014) estimated human joints using a random forest algorithm from RGB-D sensor information. The proposed system adapts online to difficult unstructured scenes taken from a moving camera. It thus does not require background subtraction.

Other notable approaches are as follows: Jain, Subramanian, Das, and Mittal (2011) proposed an upper-body joint estimation algorithm using a weighted distance transform map and human joint ratio. del Rincón, Makris, Uruñuela, and Nebel (2011) introduced a framework for visual tracking of lower body parts using Kalman and particle filters. Sheasby, Warrell, Zhang, Crook, and Torr (2012) proposed a formulation for solving the problems of human segmentation and joint estimation, using a single energy function. Zhang, Soon Seah, Kwang Quah, and Sun (2013) introduced a generative sampling algorithm with a refinement step of local optimization for body joint tracking. This multi-layer search method does not rely on strong motion priors and generalizes well to general human motions. Tran and Trivedi (2012) presented upper body pose tracking using upper body extremities and a kinematic model of the upper body in 3D with multiple cameras. Toshev and Szegedy (2014) introduced a deep neural network-based human joint estimation algorithm for RGB images. The proposed algorithm can extract the human pose information irrespective of clothing style and body type. Jain, Tompson, LeCun, and Bregler (2014) also proposed a deep learning-based human joint estimation algorithm with a new human body pose dataset (FLIC-motion). The proposed algorithm uses the RGB and optical flow information as input data of the convolutional neural network. Following the work reported in Shotton et al. (2013), the results of Shotton's work represent the best performance to date, but it is currently difficult to run their algorithms on a low-cost platform. Thus, the conventional algorithms cannot be applied to embedded video-surveillance systems. Behavior recognition algorithms, meanwhile, have been extensively discussed in recent decades, and are expected to be the next generation solutions for human machine interaction (HMI) challenges. With the popularity of 3D depth sensors, many researchers have used depth information and a human joint model for behavior recognition (Celebi, Aydin, Temiz, & Arici, 2013; Lai, Konrad, & Ishwar, 2012; Megavannan, Agarwal, & Babu, 2012; Tran & Trivedi, 2012). However, with most of the proposed algorithms, it is difficult to perform the human behavior registration process and the recognition rate is strongly affected by environmental changes (Mitra & Acharya, 2007; Poppe, 2010).

Human behavior recognition methods can be classified into three categories. Machine learning-based approaches belong to the first category: Sigalas, Baltzakis, and Trahanias (2010) presented upper body part tracking and combined a multi-layer perceptron and radial basis function neural network classifiers for human behavior recognition. Biswas and Basu (2011) proposed a human behavior recognition algorithm using a support vector machine (SVM) with depth difference information. Dubey, Ni, and Moulin (2012) introduced a fall recognition system using an SVM with RGB-D information. Lai et al. (2012) proposed a close-range human behavior recognition algorithm using feature vectors from a human joint model and nearest-neighbor classification. Liu and Shao (2013) introduced an adaptive learning method with spatio-temporal features which simultaneously fuse the RGB and depth information for hand gesture recognition. They also proposed a restricted graph-based genetic programming approach to evolve discriminative spatio-temporal features for visual recognition tasks. Wu and Shao (2014) proposed an action recognition algorithm using human pose information that can be obtained by a deep neural network algorithm. By using a hidden Markov model-based hierarchical parametric model, they showed improved action recognition performance. Fan et al. (2015) proposed a three dimensional human activity recognition algorithm with spatio-temporal local texture features. They estimated the human action by using the $k$-nearest neighbor and the hidden Markov model algorithm with integrated features from a local binary pattern operator. The first category requires training time to generate the classifier. The second category includes matching-based approaches: Megavannan et al. (2012) presented human action recognition using the motion dynamics of an object from depth difference and average depth information. Wu, Konrad, and Ishwar (2013) proposed a dynamic time-warping-based user identification and gesture recognition framework from human joint data. Celebi et al. (2013) proposed a weighted dynamic time-warping method that weights joints by optimizing a discriminant ratio. These methods exhibit different performance depending on the environment in which they are used. The computation cost, however, is lower than that of other methods. The following studies belong to the last category: