Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Curriculum in Ingegneria Informatica, Gestionale e dell'Automazione

---

Human Behaviour Understanding using Top-View RGB-D Data

Ph.D. Dissertation of:
Daniele Liciotti

Advisor:
Prof. Emanuele Frontoni

Curriculum Supervisor:
Prof. Francesco Piazza

XVI edition - new series

Abstract

The capability of automatically detecting people and understanding their behaviours is an important functionality of intelligent video systems. The interest in behaviour understanding has effectively increased in recent years, motivated by a societal needs.

This thesis is focused on the development of algorithms and solutions for different environments exploiting top-view RGB-D data. In particular, the addressed topics refer to Human Behaviour Understanding (HBU) in different research areas.

The first goal is to implement people detection algorithms in order to monitor the people activities. To this aim, a thorough study of the state of the art has been conducted to identify the advantages and weakness. An initial approach, proposed in this thesis, is based on Computer Vision (CV) techniques, it regards the extraction the head of each person using depth data. Another approach is based on deep learning and is proposed to simplify the heads detection implementation in chaotic environments and in the presence of people with different heights. These solutions are validated with a specific dataset.

The second goal is to extract several feature from subject and to identify possible interactions that they have with the surrounding environment.

Finally, in order to demonstrate the actual contribution of algorithms for understanding the human behaviour in different environments, several use cases have been realized and tested.

Declaration

I, Daniele Liciotti, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that this work was done wholly while in candidature for a research degree at the Università Politecnica delle Marche and that this thesis has not previously been submitted for a degree or any other qualification at this University or any other institution. I also confirm that where I have consulted the published work of others, this is always clearly attributed and that where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work and I have acknowledged all main sources of help. I confirm that where the thesis is based on work done by myself jointly with others. Parts of this work have been published as:

- D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, and V. Placidi. Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network. In *International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage*, pages 146–157. Springer, Cham, 2014.

- D. Liciotti, E. Frontoni, A. Mancini, and P. Zingasretti. Pervasive system for consumer behaviour analysis in retail environments. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, volume 2. 2017.

- D. Liciotti, E. Frontoni, P. Zingaretti, N. Bellotto, and T. Duckett. Hmm-based activity recognition with a ceiling rgb-d camera. In *ICPRAM (International Conference on Pattern Recognition Applications and Methods)*, 2017.

- D. Liciotti, G. Massi, E. Frontoni, A. Mancini, and P. Zingaretti. Human activity analysis for in-home fall risk assessment. In *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pages 284–289. IEEE, 2015.

- D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti. Person re-identification dataset with rgb-d camera in a top-view configuration. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 1–11. Springer, 2016.

- M. Sturari, D. Liciotti, R. Pierdicca, E. Frontoni, A. Mancini, M. Contigiani, and P. Zingaretti. Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. *Pattern Recognition Letters*, 2016.

Contents

Chapter 1.

Introduction

This Thesis addresses the subject of HBU using top-view RGB-D data. The objective of the Thesis is described in this Chapter, together with the definition of the research problem, the main contributions, and the Thesis organization.

1.1. Research problem

In recent years, a lot of researchers have focused the attention on automatic analysis of human behaviour because of its important potential applications and its intrinsic scientific challenges. In several technological fields the awareness is emerging that a system can provide better and more suitable services to people only if it can understand much more about users' preferences, personality, social relationships etc., as well as about what people are doing, the activities they have been concerned in the past, their life-styles and routines, etc.

CV and deep learning techniques are currently the most interesting solutions to analyse the human behaviour. In particular, if these are used in combination with RGB-D data that provide high availability, reliability and affordability.

Detection and tracking algorithms allow to generate motion descriptions of subjects which are used to identify actions or interactions. Consequently, it is possible to associate to a certain sequence of actions a particular behaviour. In this view, investigating technological solutions aimed at improving the environments and adapting them to the specific user requirements, can be very useful.

The problem remains largely open due to several serious challenges, such as occlusions, change of appearance, complex and dynamic background. To counter these challenges, several studies adopt the top-view configuration because it eases the task and makes simple to extract different trajectory features. This setup also introduces robustness, due to the lack of occlusions among individuals.

Different domains are analysed in this Thesis, such as those of video surveillance, Intelligent Retail Environments (IRE) and Activities of Daily Living (ADLs).

1.2. Objectives and contributions

The objective of this Thesis is to understand the human behaviour in different real scenarios using CV techniques applied on RGB-D data in top-view configuration. To this aim, a thorough study of the literature will be presented, identifying advantages, challenges and issues related to the use of this particular configuration. Furthermore, in order to support this research, several use cases will be presented. In particular, one of these was conducted during the five months of Ph.D. visiting period at Lincoln Centre for Autonomous Systems (LCAS) in the School of Computer Science at the University of Lincoln (UK).

1.3. Structure of the Thesis

The Thesis is organized in five Chapters, which describe and detail the different approaches and applications for human behaviour understanding using top-view RGB-D data. The Thesis has the following structure.

Chapter 2 reviews the state-of-the-art about the two main topics addressed: human behaviour analysis and RGB-D data from top-view.

In Chapter 3 are proposed some algorithms for people detection using RGB-D data in top-view configuration.

Chapter 4 describes different use cases, in particular are analysed applications on: video surveillance and analytics, intelligent retail environment, and ADLs.

Finally, conclusions and discussion are drawn in Chapter 5 where, after clarifying the contribution of this work, some future research directions are identified. Furthermore, this Chapter besides arguing over the possibilities that the proposed applications opens up in different topics, summaries also the challenges, the open issues and the limitations that require further investigations.

Chapter 5.

Conclusions and future works

In this Thesis, different algorithms and applications based on RGB-D data in top-view configuration for HBU have been proposed. In particular, the implemented algorithms can be used for people detection and tracking and, for human interactions understanding. This Chapter starts presenting a discussion summarizing the main results achieved in the Thesis. Then, the main contribution is highlighted. Finally, the open issues and future research directions are presented.

5.1. Discussion

After the introduction, a review of the literature on the two main topics addressed in this Thesis, i.e. HBU and RGB-D data form top-view has been provided. Chapter 2 included also an overview of main public available datasets.

Two approaches for people detection are proposed in Chapter 3. In particular, the objective of these approaches is to find the heads of people present in the depth image.

Image processing techniques, such as water filling and multi level segmentation, provide the shape of heads by evaluating depth data local minima. In this way, the number of false positives increases when the subjects gesticulate with their hands. In fact, these could be confused for heads. The introduction of particular constraints on the shape could reduce this number, but the risk is that the algorithm may become too specific for each setup.

Approaches based on semantic segmentation techniques allow a neural network to distinguish a particular class, which in this case corresponds to the class "head". Five different types of CNNs were tested and achieved significantly better results than image processing approaches.

In Chapter 4 are presented some applications for different use cases. Previous algorithms for people detection have been used to monitor their movements within a particular area. In fact, three macro-research fields were analysed.

In the context of video surveillance, it was necessary to extract different characteristics of the subject in order to re-identify the latter when it reappears

a second time. A dedicated dataset has been built to solve this problem and anthropometric and colour based features have been extracted. CMC curves evidence the robustness and good ability of descriptors to recognize the various subjects.

Another application addressed has been developed in an IRE. Several RGB-D sensors have been installed in a real store in order to monitor the behaviour of consumers in front of different shelves. Moreover, through depth data, the system detects all type of interaction that the costumers has with the shelf. Four different CNNs have been developed to understand the type of interaction, i.e. whether or not the customer has taken a product from the shelf. This system is useful because it is able to extract a series of indicators that describe the performance of store up to the single shelf in real time.

The last application scenario, where top-view RGB-D data for people detection algorithms has been used, was home environments. Two different types of problems have been addressed: ADLs and fall detection. The first was HMM approached. 3D points of head and hands were used as input of model (observation) in order to predict the activity of the user (state). The model was validated with a dataset and five types distinct of activity were considered. Instead, fall detection problem was solved by monitoring the person's height value.

5.2. Thesis contributions

The main contributions of this Thesis can be summarized as follows:

- design and implementation of two novel algorithms for people detection from top-view configuration with RGB-D data using image processing approaches. In particular, a performance improvement of water filling algorithm is proposed in terms of computational complexity. Furthermore, a new algorithm, called multi level segmentation, has been developed. It carries out several segmentations on different levels of height in order to find all the heads of people.

  This work has been published in [61, 65, 63];

- development of semantic segmentation CNNs for heads detection, in particular, U-Net, SegNet, FractalNet, and ResNet are used in this work. By introducing changes on different layers of these nets, the performances are significantly improved;

- proposal and validation of new descriptors for Re-id task in top-view configuration. Descriptors are composed of anthropometric and colour-based features.

This work has been published in [66];

- design and implementation of several CNNs for user-shelf interaction recognition. Through a manually annotated dataset made up of images representing interactions between user and shelf, four different types of CNNs have been trained.

- creation of four public available datasets:
  - TVPR Dataset
  - TVHeads Dataset
  - RADiAL Dataset
  - User-Shelf Interactions Dataset

  Some of these datasets have been published in [66, 64];

In this Thesis, the potential of top-view configuration for detection and tracking applications in several sub-domains has been demonstrate, to outline key limitations and to indicate areas of technology where solutions for remaining challenges may be found. The success of RGB-D cameras can be closely linked to their affordability and to the additional depth information coupled with visual images that this approach provides. These cameras have already been successfully applied in the several field to identify people and to analyse behaviours and interactions. The choice of the RGB-D camera in a top view configuration is due to its greater suitability compared with a front view configuration, usually adopted for gesture recognition or even for video gaming. The top-view configuration reduces the problem of occlusions and has the advantage of being more privacy preserving, because a person's face is not recorded by the camera. Starting from this, further investigation could be devoted to explore approaches more accurate and effective such as Convolutional Neural Networks or U-Net [99].

5.3. Open issues and future works

This section analyses the open issues in the proposed algorithms and applications, identifying some future research directions.

Novel algorithms for head detection are based on deep learning approach. The CNNs used in this Thesis are the best techniques of semantic segmentation. Indeed, this approach is better than traditional image processing techniques because it is not based on geometric or colour constraints, but rather allows to identify the heads of individuals based on previous learning.

Further investigation on Re-id task and video surveillance field, will be devoted to the study of more sophisticated features. The CMC curves have suggested that for the different distance metric approaches the depth descriptor has strong discriminative power. The integration of more features in the model seems to improve the identity discrimination. This aspect is of great importance, in order to perform a classification model. Future works would include the use of other types of RGB-D sensors, such as time of flight (TOF) ones. The system can additionally be integrated as a source of high semantic level information in a networked ambient intelligence scenario, to provide cues for different problems, such as detecting abnormal speed and dimension outliers, that can alert of a possible uncontrolled circumstance.

Future projects about IRE are directed towards a detailed study of person Re-id using top-view RGB-D data from several cameras, a task necessary to assign a single and robust identifier to each buyer. Among several other information (i.e. audio recognition system, carts tracking), this will allow us to better describe the client behaviour inside the shop and not only in front of a single shelf.

Future efforts in the field of assistive technology are expected on the integration of video and audio systems. In this way, the identification of abnormal events, such as the strange activity of the user, an intrusion or a object breakage, can can be detected using audio microphones. Further investigation will be devoted to extend HMM approach to select human joints that provide the most informative spatio-temporal relations for ADLs classification. The long term goal in this field is to develop a mobile robot that searches for the best location to observe and successfully recognise ADLs in domestic environments.