CHAPTER 1 INTRODUCTION

1.1 Motivation

With the improvement in the living standard and the medical treatment, the average longevity of the Taiwanese extends progressively, so does the ratio of elderly population (aged 65 and older). According to the population estimation by the Cabinet's Council for Economic Planning and Development, Taiwan already reached the so-called “aging society” in the year 1994 with 7 percent of the elderly and is expected to enter into the “aged society” in 2018 with 14.25 percent. And the ratio is estimated to be 24.28 percent in 2035 that means our country will reach the “super-aged society” by then.

Dwellings are main spaces for elder life, but they are the places where many contingencies occur. According to the statistics by the Tokyo Fire Department, up to 90 percent of elders encounter contingencies, such as falling down, in the places related to dwellings. About 10 of 24 persons who encounter contingencies at home are elders. Therefore, home care for elders is an issue that is very important and worthy of studying.

We consider that a computer vision system can assist the caring of elders at home in certain degree, and it seems the most important to detect the occurrence of dangerous actions as soon as possible. Therefore, we will construct a surveillance system that can be used to detect the occurrence of dangerous actions for elder’s home care.

1.2 Problem Definition

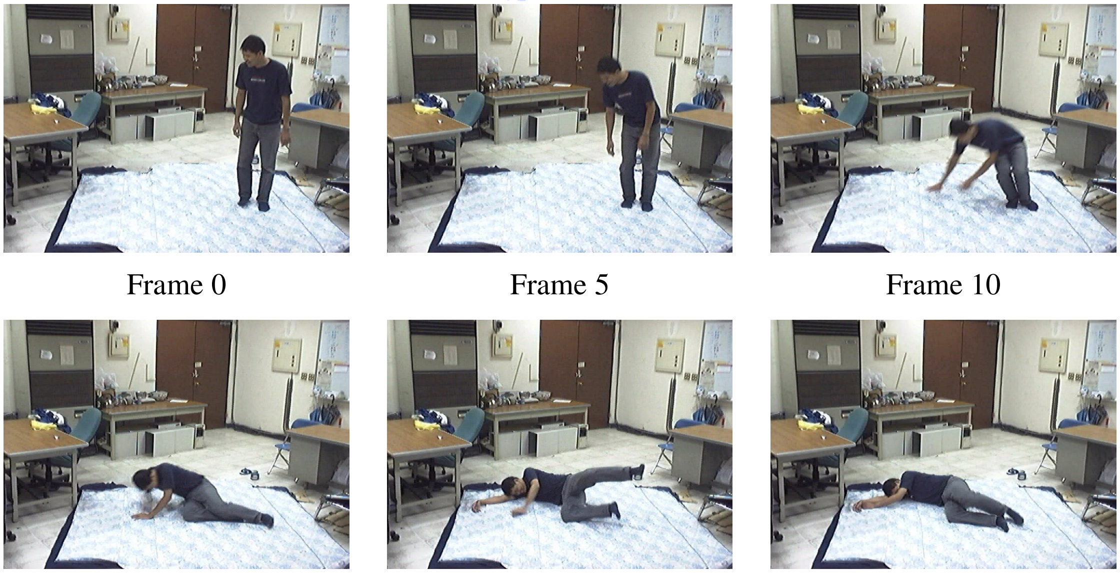
In order to construct the surveillance system for elder’s home care, the problems that we have to solve are as follows.

1.2.1 Detection of single persons

In order to recognize human actions, we need to search for humans in each image of the image sequence. That is, we need to segment human regions from each image. Human regions are generally one type of foreground regions. To obtain human regions, we need to detect foreground regions in each image, and then determine whether the foreground regions are human regions. Each human region could contain a single person or multiple persons. If a human region contains multiple persons, it obviously means that more than one person are in the surveillance environment. We assume that they will care for each other. Therefore, only single person regions are considered in our system. To avoid the problem of human tracking, we can assume that there is only one person in the surveillance environment.

1.2.2 Recognition of human actions

In this thesis, our work for human action recognition is emphasizes the detection of dangerous human actions, such as falling down. In Fig 1.1, an example of dangerous actions, falling down, is shown.



Frame 0

Frame 5

Frame 10

Frame 15

Frame 20

Frame 25

Fig. 1.1 An example of dangerous actions, falling down.

1.3 Survey of Related Works

1.3.1 Background Modeling

Stationary cameras are typically used for outdoor and indoor surveillance. Because of the stationary cameras, moving objects can be detected by comparing each frame with a representation of the background scene. This process is called background subtraction and the representation of the background scene is called the background model. A common background modeling method consists of a background model and a background subtraction process.

The most common feature used in background modeling is pixel intensity (color). For a completely static background scene, the pixel intensity is reasonably modeled with a normal distribution. The normal distribution model can adapt to slow changes, e.g., gradual illumination changes, in the background scene by recursively updating the model with a simple adaptive filter. The basic model was used by Wren et al. [1]. Karmann and Brandt [2] used Kalman filtering to adapt changes. In general, background scenes could not be completely static, e.g., outdoor scenes with moving trees and bushes. The pixel of the background scene could have different intensity (color) values over time. Therefore, a single normal distribution model for the pixel intensity does not hold. Instead, a generalization based on a mixture of normal distributions was used to model such variations. Grimson et al. [3] used a mixture of K normal distributions (typically, K is a small number from three to five) to model the pixel intensity. The mixture is weighted by the frequency with which each of the normal distributions explains the background. Friedman and Russell [4] used a mixture of three normal distributions to model the pixel value for traffic surveillance applications. The pixel intensity was modeled as a weighted mixture of three normal distributions corresponding to road, shadow, and vehicle distribution. Using an incremental version of the EM algorithm, adaptation of the normal distribution mixture models can be achieved. Toyama et al. [5] used linear prediction using the Wiener filter to predict pixel intensity given a recent history of values. The prediction coefficients are recomputed each frame from the sample covariance to achieve adaptation. Linear prediction using the Kalman filter was used by Karmann and Brandt [2]. Elgammal et al. [6] used a nonparametric technique to model the pixel intensity. Adaptation of this model can be achieved by adding new pixel values and ignoring older pixel values.

1.3.2 Human Action Recognition

A paper by Gavrila [7] is an excellent survey on the visual analysis of human movement. The scope of this survey is limited to work on whole-body or hand motion. The emphasis is on discussing the various methodologies; they are grouped into three categories:

1. 2-D approaches without explicit shape models

   The 2-D approaches without explicit shape models are to bypass a pose recovery step altogether and to describe human movement in terms of simple low-level, 2-D features from a region of interest.

2. 2-D approaches with explicit shape models

   The 2-D approaches with explicit shape models take essentially a model- and view-based approach to segment, track, and label body parts using explicit a priori knowledge of how the human body (or hand) appears in 2-D.

3. 3-D approaches

   The 3-D approaches aim to recover 3-D articulated pose over time.

More recently, a comprehensive survey of computer vision-based human motion capture literature has been presented by Moeslund and Granum [8]. This survey focuses on a general overview based on human motion capture system functionalities that are initialization, tracking, pose estimation, and recognition. In this survey, the recognition is classified into two categories:

1. Static recognition

Static recognition is concerned with spatial data, one frame at a time. The approaches usually compare pre-stored information with the current image.

2. Dynamic recognition

These approaches use temporal characteristics in the recognition task. Relatively simple activities, such as walking, are typically used as test scenarios.

Our approach to human action recognition is similar to the 2-D approaches without explicit shape models surveyed by Gavrila. Since our approach combines the static and dynamic recognition approaches, it differs from all the approaches surveyed by Moeslund and Granum.

Unlike our work, which emphasizes the ability to detect dangerous human actions, past works focused on recognition of normal human actions. Fujiyoshi and Lipton [9] classified human motion into “running” or “walking” using the frequency analysis of internal human motion features. Skeletonization was used to extract the motion features. Yamato et al. [10] used Hidden Markov Models (HMMs) for human action recognition in time sequential images. Template matching techniques were used by Polana and Nelson [11] to recognize human activities. Bobick and Davis [12] used Motion Energy Images (MEIs) for recognition. A surveillance system, called W^4, was constructed by Haritaoglu et al. [13] for detecting and tracking humans and monitoring their activities. The activity recognition part of this system is based on analysis of the projected histograms of detected human silhouettes. In this system, human postures were classified into one of four main postures (standing, sitting, crawling/bending, and lying) and one of three view-based appearances (front/back, left-side, and right-side). Human activities were monitored by tracking the posture changes over time. HMM and stochastic parsing were used by Ivanov and Bobick [14] to recognize generic activities. First, generic activities were detected as a stream of low-level action primitives represented using HMM. Then, these activities were recognized by parsing the stream of primitive representations using a context-free grammar. Bobick and Davis [15] matched temporal templates against stored instances of views of known actions to recognize human activities. More recently, Variable-Length Markov Models (VLMMs) was used by Galata et al. [16] for modeling human behavior. VLMMs were used because of their more powerful encoding of temporal dependencies.

1.4 Assumptions

In order to solve the considered problems, the assumptions that have to be made are listed below.

1. The environment is an indoor environment.
2. The lighting is sufficient and stable.
3. The sensor is stationary.
4. The images are color, whose size is 640×480.

1.5 System Description

The diagram in Fig 1.2 shows our system architecture. In Background Modeling, a background model is built using statistical modeling. The background model can be updated to adapt to the changes of the background scene. Foreground pixels are detected using the background model, and then are grouped into foreground regions through some processes in Foreground Detection. In Human Extraction, whether a foreground region contains a single person is determined according to the number of its pixels and the proportion of its skin color pixels to its pixels. If a foreground region is considered as a single person, then a silhouette-based posture analysis is applied to this foreground region to estimate the posture of the single person in Posture Analysis. Features used to represent the single person posture are computed in Posture Description, and then the posture of the single person is classified into one of seven predetermined postures using the features: standing, stooping, sitting with crooked legs, squatting, kneeling, sitting with stretched legs, and lying down/prone in Posture Estimation. In Action Recognition, a posture state transition diagram constructed with the seven main postures is used to recognize human actions. The results of the silhouette-based posture analysis are used to implement state transitions. According to the state transitions, any human action can be classified into normal action or dangerous action.

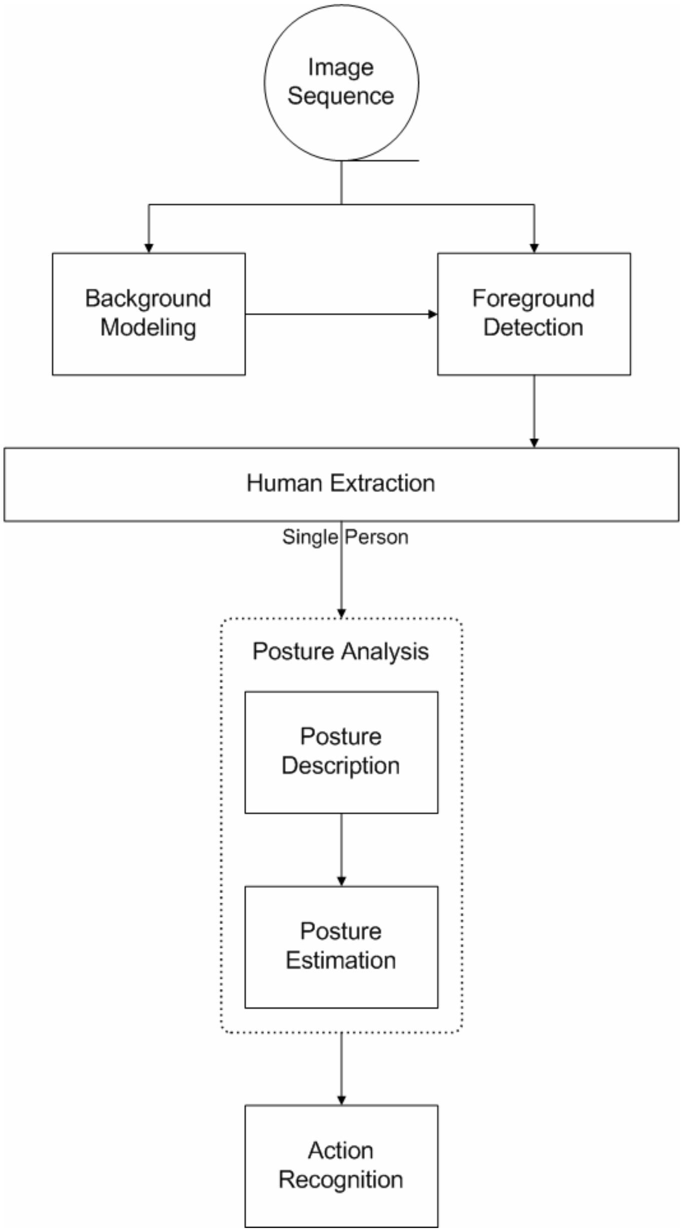


Fig. 1.2 The system architecture.

1.6 Organization of this Thesis

The rest of this thesis is organized as follows. Chapter 2 describes the method for the detection of single persons. Chapter 3 specifies the silhouette-based posture analysis used to estimate human posture. In chapter 4, the approach to human action recognition is described in detail. Some experimental results and analyses are presented in chapter 5. Chapter 6 offers some conclusions and future works.