Human Body Part Classification and Activity Recognition for Real-Time Systems

Burak Ozer
Verificon Corp., USA

Tiehan Lv
Princeton University, USA

Wayne Wolf
Princeton University, USA

INTRODUCTION

Recent advances in camera and storage systems along with increased algorithmic and computational power of 3D graphics hardware are main factors driving the increased popularity of multicamera applications. Since prices continue to drop on components, cost-effective systems can be developed for a wide range of applications such as teleimmersion, humanoid robots systems, automated video surveillance, and interactive video games.

The increased importance of such applications that require fast, cheap, small, and highly accurate smart cameras necessitates research efforts to provide efficient solutions to the problem of real-time detection of persons and classification of their activities. A great effort has been devoted to three-dimensional human modeling and motion estimation by using multicamera systems in order to overcome the problems due to the occlusion and motion ambiguities related to projection into the image plane. However, introduced computational complexity is the main obstacle for many practical applications. Different applications require different levels of modeling-related performance metrics, that is, accuracy, speed, and robustness, hence different 3D techniques. Therefore, the relationship between the activity recognition algorithms and the architectures required to perform these tasks in real time is a fundamental design issue.

BACKGROUND

Human detection and motion recognition involve interpretation of elementary visual features such as shape, color, texture, and motion. A geometric model is an approximation of the shape and of the deformations of the object. This model can be two-dimensional (modeling the contours of the projections of the object in the images) or three-dimensional (modeling the surfaces of the object). Two-dimensional human shape models are generally made of curves, segments, sticks, snakes, and so forth, where 3D models are either systems of rigid bodies (e.g., spheres, superquadrics, etc.) or deformable surfaces (e.g., mesh). Template matching, chain coding, Fourier transform, boundary signatures, moments, and polygonal approximations are popular shape-matching algorithms. The required properties of a shape description scheme are invariance to translation, scale, rotation, luminance, and robustness to partial occlusion. Color and texture are mostly used to detect skin areas in human-detection tasks. The articulations may be modeled by joints or by the motion of control points (e.g., B-splines). Major approaches for analyzing spatial and temporal patterns include dynamic time warping, neural networks, hidden Markov models (HMMs), time-frequency analysis, template matching, and principal component analysis. The choice between a 2D or 3D model depends on the application, for example, the needed precision, number of cameras, or type of motion to be recognized.

Several researchers work with 2D features to recognize human movement (Gavrila, 1999). Goddard (1994) uses model-based recognition techniques, namely, stick figures. Other researchers who use 2D models are Comaniciu, Ramesh, and Meer (2000), Isard and MacCormick (2001), and Papageorgiu and Poggio (1999). Most of the work in this area is based on the segmentation of different body parts. Wren, Clarkson, and Pentland (2000) proposed a system, Pfinder, to track people by using blobs that represent different body parts. W4 is another real-time human-tracking system (Haritaoglu, Harwood, and Davis, 1998), where the background information should be collected before the system can track foreground objects.

Three-dimensional human and activity recognition systems can be classified in terms of the visual analysis of multiple cameras, that is, the projection of 3D models onto 2D images versus 3D visual reconstruction from stereo images. Aggarwal and Cai (1999), Gavrila (1999), and Moeslund & Granum (2001) present overviews of various methods used for articulated and elastic nonrigid motion detection. More background information on gesture recognition can be found in Wu and Huang (1999).

One of the early works on tracking articulated objects is proposed by O'Rourke and Badler (1980). The authors used a 3D model of a person made of overlapping spheres. Kakadiaris and Metaxas (1998) proposed a method to generate the 3D model of an articulated object from different views. The authors used extended Kalman filter for motion prediction. Gavrila and Davis (1996) used superquadrics to model the human body and dynamic time warping to recognize the human motion. Luck, Debrunner, Hoff, He, and Small (2002) obtained 3D models of the moving human body by extracting the silhouettes from multiple cameras. In this work, 3D information is used for HMM-based activity recognition in real time for different applications without requiring any specific pose or user interaction and using distributed processing. This method is explained in more detail in the next section with an emphasis on issues related to the relationship between the algorithms and design architecture.

Most of the previous work for human detection depends highly on the segmentation results, and mostly motion is used as the cue for segmentation. Most of the activity-recognition techniques rely on the successful feature extraction, and proposed approaches are suitable for a specific application type. There is a need for more robust techniques that can use model-based information feedback and connect low-level features to high-level semantics in real time.

MAIN THRUST OF THE ARTICLE

As previously stated, 3D human- and activity-recognition systems can be classified in terms of the visual analysis of multiple cameras, that is, the projection of 3D models onto 2D images versus 3D visual reconstruction from stereo images. Depending on the selected technique, different steps must be implemented. Visual reconstruction for virtual reality requires high accuracy while real-time activity recognition and trajectory estimation require high-speed techniques. Various aspects and challenges of 3D visual-analysis algorithms include instruction statistics, branch behavior, and memory access behavior of different program parts, for example, stereo matching, disparity map generation, reconstruction, projection, 2D or 3D human body part detection, 2D or 3D tracking, 2D or 3D activity recognition, and so forth.

Algorithmic and Hardware Issues

Algorithmic Issues

In general, we can classify 3D human detection and activity recognition methods into two categories (Cheung, Kanade, JBouguet, & Holler, 2000): off-line methods, where the algorithms focus on detailed model reconstruction, for example, wire-frame generation, and real-time methods with a global 3D human model reconstruction (Delamarre & Faugeras, 2001). The major challenge in many 3D applications is to compute dense range data at high frame rates, since participants cannot easily communicate if the processing cycle or network latencies are long. The works by Kakadiaris and Metaxas (1995) and Mulligan, Isler, and Daniilidis (2001) are examples for non-real-time methods. Most of the real-time methods use a generic 3D human model and fit the projected model to the projected silhouette features (Luck et al., 2002). The speed of the systems depends highly on the voxel resolution.

The algorithmic pipelines perform a wide range of disparate operations:

- pixel-by-pixel operations,
- pixel-region operations,
- mixed operations, and
- nonpixel operations.

Traditionally, the algorithms start with operations that are clearly signal oriented and should move steadily away from the signal representation until the data are very far removed from a traditional signal representation. In general, the volume of data goes down as image processing progresses.

Hardware Issues

Real-time implementation of image- and video-processing algorithms necessitates data- and instruction-level parallelism techniques to achieve the best performance for several application types. Besides the algorithm development, hardware design is one of the most important issues for a real-time system. Watlington and Bove (1997) proposed a data-flow model for parallel media processing. Davis, Borovikov, Cutler, and Horprasert (1999) developed a multiperspective video system. Fritts, Wolf, and Liu (1999) evaluated the characteristics of multimedia applications for media processors. Researchers also pay attention to multiprocessor architecture. Simultaneous multithreading is proposed by Tullsen, Eggers, and Levy (1995). An IMAGINE processor is being developed at Stanford University, which has an explicit, programmable