

Automatic Extraction of Silhouettes from Video Sequences: Literature Research and Design Decision

Marco Nef *

Semester Thesis in Photogrammetry
ETH Zürich, Switzerland, April 2001

Abstract

The task of my semester thesis in the minor subject photogrammetry is to implement an algorithm that can extract human silhouettes from video sequences. As a first step I had a look through several papers about this topic that are available on the Internet and indexed by the search engine Google ¹.

This paper gives an overview over different possible methods to solve the problem. It also describes the design of the solution I am going to implement.

1 Introduction

For the project *REBOMO* (Realistic Body Modeling from Video Sequences) a person is filmed by three CCD cameras from three different points of view at the same time. This person is allowed to move around under absolutely no restrictions. The environment where the person is filmed may be any given scene. There are no special requirements for the background. It should be possible to e.g. just go to a park and film people walking there.

Using photogrammetric methods it is possible to extract 3D information of the human body and also its movements. Actual methods for this process require the user to define the silhouette of the body in all frames of the movie. The task of this semester thesis is to develop and implement an algorithm that solves

this problem automatically or at least provides some help to the user.

A property of the algorithm to develop is that it will mainly be used for postprocessing. That means that there is (almost) no limitation to processing time. Much more important than the fast execution is the quality of the resulting silhouette.

The next chapter gives a short overview over some different approaches to the problem. On one hand some of them are quite robust. On the other hand the project time for a semester thesis is very limited. This leads to a design as described in chapter 3. In the last chapter there is a short discussion of the chosen approach.

2 Existing approaches

When looking around on the Internet what other people have already done to extract silhouettes, one thing becomes immediately obvious: It is very hard to design a robust method solving the problem.

The following sections give an overview over several approaches. Most of them are not useful for my task at all, others give interesting ideas. Anyway, it is not possible to describe them in a deep manner. For more details you are asked to have a look at the referenced papers which can all be found on the Internet.

2.1 Solutions with special demands

There are many proposals for a solution of the problem to extract a human silhouette out of a video stream. But most of them require a very

* marco%shima.ch – <http://www.shima.ch/papers>

¹ <http://www.google.com> (indexes PDF documents)

special scene: They work with *chroma-keying* (e.g. blue-screening) like [1]. This method requires the person stand in front of a background consisting of a uniform-colored wall or screen. While processing the (almost ²) unicolor background is extracted from the video frames which is very easy done. This type of system restricts the user not to have the color of the background anywhere on his clothing.

Some systems require the actor to wear electronic sensors. They use methods such as electromagnetic or electric field sensing [2, 3, 4, 5].

Another approach is to light the scene with infrared light [6]. The benefit of this method is that it works like chroma-keying. The disadvantages are the fact that a special equipment (infrared lamps and cameras) is required, and the difficulty of registering chromatic and infrared pictures.

There is an actual project for a collaborative, immersive virtual environment [7] at ETH Zürich, Switzerland, where a more complex technology for the picture aquirement is used. In order to get a standardized illumination they periodically change between scanning the local environment under special light conditions and projecting the virtual scene in 3D ³.

2.2 General solutions

Some methods do not require anything special to the environment where the pictures are taken. Rather they try to analyse the content of pictures and to categorize found objects. A candidate of this method is [8]. They track people in crowded and/or unknown environments using multi-modal integration in real-time by combining stereo, color, and *face detection* modules into a single system. This method is also used to recognize heads [9].

Another proposal is Pfinder [10] which uses a multi-class statistical model of color and shape to segment a person from a background scene.

²There always is some noise.

³ At the moment I am writing a second semester thesis in the *blue-c* project. It is about expanding a geometry-based scene-graph by point-based objects. More information: <http://www.shima.ch/papers>

There are also two Fourier-based approaches [11, 12] that use *Fourier descriptors*. With these descriptors they can evaluate whether a detected object is human or just a chair. This method seems to be very robust, but is implementationally very complex.

A method that is often used in Computer Vision is to track the *optical flow* in a video sequence. One paper describing this method is [13]. They recognize humans by analyzing the curl of the optical flow which is the rotation of vectors in a vector field. Like that it is possible to extract rotating extremities of the human body like arms and legs. Optical flow based methods in general are introduced in [14] and used in [15, 16].

[14] introduces a second approach: *Snakes*. A snake is an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it towards features such as lines and edges.

2.3 Refinement

All the proposals mentionend above lack in some aspects. A problem is e.g. the recognition of shades. Because shades inherently move around with the person they belong to they are usually recognized as a part of the person's body. If the extracted person is to be used in a virtual environment, there might be different illumination and shading. That is why one should try to categorize and shade body parts.

One system that solves this problem is Ghost as a part of the W^4 real-time system for detecting and tracking people [17, 18, 19]. This system is able to name body parts of a given silhouette.

The curl-based system [13] can be used to identify arms and legs. If body parts are identified in the picture it is possible to cut the most problematic shades around the feet.

3 Design decision

3.1 MediPicture algorithm

The algorithm to be implemented uses background extraction. To do that first of all one has to generate a representation of the background that can be extracted. This representation consists of three images created by an algorithm as described in [18] and shall be called *MediPicture*. It models minimum and maximum intensity values of all the pixels, and maximum intensity differences between consecutive frames observed during the training period.

It has to be tested how a *training period* is optimally configured. Under the assumption that the person on the images moves over a noticeable distance while the video stream is taken it may be a good idea to inspect a certain amount of frames equally distributed over the whole time. Like that the background should be visible at all pixels for a larger period.

Another possibility to get something like a MediPicture is to shoot the scene as a preparation process. With this method it is not possible to get a complete MediPicture but only an approximation. Like that the background extraction has to happen in a different, less accurate way. This will not be implemented.

Anyway, if the film is taken over a longer period (day, night, sun, clouds) the MediPicture could be inaccurate. In this case the whole period has to be divided into subperiods which are analyzed separately with different MediPictures.

After having calculated the MediPicture, the background is subtracted of all frames resulting in a movie that contains the moving objects (may be several objects) and some noise. The next step is to binaryze the frames, that means to make them black and white. The background is represented in white, the objects are black. Because the background extraction results in a classification of background elements and the foreground objects, these two steps can be unified. After that filters can be applied to cut small black particles which represent noise.

The last step is to find edges in the black

and white frames and to represent them as a polygon. The polygon shall be minimized in the way that points laying on lines are not part of the result.

3.2 Data representation

The extracted silhouette can be stored in and reloaded from a silhouette file. The file contains a list of coordinates (x, y) which represent feature points of a polygon approximating the exact silhouette. The coordinates are integer values defining the locations of the points relative to the left top corner of the pictures.

3.3 User interface

A user interface shall be provided in which the user can inspect and correct the calculated silhouettes for all frames. The environment used for the implementation is the *GTK*⁴ library for creating graphical user interfaces in the *GNOME*⁵ desktop environment which is provided with most actual *LINUX* distributions⁶. *GTK* is built on top of *GDK*⁷ which is basically a wrapper around the low-level functions for accessing the underlying windowing functions (*XLib* in the case of *XWindows*). *GTK* is essentially an object oriented application programming interface (API).

The following functionality shall be provided in the user interface:

- Loading of video streams
- Calculating a MediPicture
- Calculating the silhouette, using the *MediPicture algorithm*, including background extraction and filtering
- Visualization of silhouettes
- Storing of silhouettes

Scaling of a silhouette may be a useful tool to get a silhouette that contains some of the

⁴ GIMP Toolkit, licensed using the LGPL license

⁵ GNU Network Object Model Environment

⁶ e.g. RedHat, SuSE.

⁷ GIMP Drawing Kit

neighbor pixels. Usually in raster graphics these pixels contain some color information that would be lost otherwise.

4 Discussion and future

There are mainly two problems that are not solved:

- Shades are not identified
- Objects occluding each other are not identified

In order to get more robust results methods as described under 2.3 may be implemented. Because of the time limitation of a semester thesis this cannot be part of my thesis.

References

- [1] T. Darrell, P. Maes, B. Blumberg, A. Pentland. A novel environment for situated vision and behavior. 1994. IEEE Workshop for Visual Behaviors.
- [2] J. Strickon, J. Paradiso. Tracking hands above large interactive surfaces with a low-cost scanning laser rangefinder.
- [3] J. Paradiso. Electronic music interfaces. *IEEE Spectrum* 34, pages 18–30, 1997.
- [4] J. Rekimoto, N. Matsushita. Perceptual surfaces: towards a human and object sensitive interactive display.
- [5] H. Ishii, B. Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. *Conference on Human Factors in Computing Systems*, pages 234–241, 1997.
- [6] James W. Davis, Aaron F. Bobick. A Robust Human-Silhouette Extraction Technique for Interactive Virtual Environments. MIT Media Lab, Cambridge MA 02139, USA.
- [7] Prof. Dr. Markus Gross. blue-c project web site. <http://blue-c.ethz.ch>.
- [8] T. Darrell, G. Gordon, M. Harville, J. Woodfill. Integrated Person Tracking Using Stereo, Color and Pattern Detection. 1998. Interval Research Corp., 1801C Page Mill Road, Palo Alto CA 94304, USA.
- [9] Marco La Cascia, Stan Sclaroff. Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22/4, April 2000.
- [10] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland. Pfinder: Real-Time Tracking of the Human Body. *In Proc. of the SPIE Conference on Integration Issues in Large Commercial Media Delivery Systems*, October, 1995.
- [11] Hannu Kauppinen, Tapio Seppänen, Matti Pietikäinen. An Experimental Comparison of Autoregressive and Fourier-based Descriptors in 2-d Shape Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:201–207, 1995. University of Oulu, Department of Electrical Engineering, 90570 Oulu, Finland.
- [12] Rocío Díaz de León, Luis Enrique Súcar. Human Silhouette Recognition with Fourier Descriptors. ITESM-Campus Cuernavaca, Morelos, México.
- [13] Toru Tamaki, Tsuyoshi Yamamura, Noboru Ohnishi. Extraction of Human Limb Regions and Parameter Estimation based on Curl of Optical Flow. *Proceedings of the Fourth Asian Conference on Computer Vision (ACCV2000)*, II:1008–1013 (2000,1), 2000.
- [14] Diego Garcés. Body silhouette extraction from video sequences. 1999. Computer Graphics Laboratory, EPF Lausanne, Switzerland.

- [15] Nicola D'Appuzzo. Motion Capture by Least Squares Matching Tracking Algorithm. *AVATARS2000, 30.11.-1.12.2000, Lausanne, Switzerland*, 2000. Institute of Geodesy and Photogrammetry, ETH Zürich, Switzerland.
- [16] Ralf Plänklers, Pascal Fua. Tracking and Modeling People in Video Sequences. Computer Graphics Lab, EPF Lausanne, Switzerland.
- [17] Ismail Haritaoglu, David Harwood, Larray S. Davis. *Ghost: A Human Body Part Labeling System Using Silhouettes. 14th International Conference on Pattern Recognition, August 16-20, 1998, Brisbane, Australia.* Computer Vision Laboratory, University of Maryland, USA.
- [18] Ismail Haritaoglu, David Harwood, Larray S. Davis. *W⁴: Who? When? Where? What? A Real Time System for Detecting and Tracking People. 3. International Conference on Face and Gesture Recognition, April 14-16, 1998, Nara, Japan.* Computer Vision Laboratory, University of Maryland, USA.
- [19] Ismail Haritaoglu, David Harwood, Larray S. Davis. *W⁴S: A Real-Time System for Detecting and Tracking People in 2 $\frac{1}{2}$ d.* Computer Vision Laboratory, University of Maryland, USA.