
Content-Based Representation of Sign Language Video Sequences

Nariman Habili

B.Sc., B.Eng. (Flinders)

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy
in the
School of Electrical and Electronic Engineering
The University of Adelaide
Australia

September, 2002

Contents

Abstract	xv
Statement of Originality	xvii
Acknowledgments	xix
Dedication	xxi
Publications	xxiii
List of Principal Symbols	xxv
List of Abbreviations	xxvii
1 Introduction	1
1.1 Content Based Representation: An Overview	2
1.2 Sign Language Video Communication	3
1.3 Research Objectives	3
1.4 Contributions of the Thesis	4
1.5 Outline of the Thesis	6
2 Segmentation, Video Coding, and Sign Language: An Overview	9
2.1 Segmentation	10
2.1.1 Approaches to Still Image Segmentation	14
2.1.2 The Use of Motion in Video Segmentation	17
2.2 Video Coding	19
2.2.1 Block-Based Video Coding	21

2.2.2	Content-Based Video Coding	22
2.3	Sign Language	26
2.3.1	Characteristics of Sign Language	27
2.3.2	Sign Language Video	28
2.4	Summary	29
3	Color	31
3.1	Light and Color	32
3.2	The Human Visual System	33
3.2.1	Anatomy of the Eye	33
3.2.2	Color Perception	34
3.2.3	The Opponent-Color Model of Chromatic Vision	36
3.3	The Trichromatic Theory of Color Mixture	37
3.4	The Dichromatic Reflection Model	38
3.5	Color Spaces	40
3.5.1	CIE XYZ Color Space	40
3.5.2	YUV Color Space	41
3.5.3	YIQ Color Space	43
3.5.4	YCbCr Color Space	43
3.6	Summary	44
4	Motion	47
4.1	Camera Models	48
4.2	Motion Models	51
4.2.1	Three-Dimensional Motion	51
4.2.2	Two-Dimensional Motion	53
4.3	Scene Model	56
4.4	2D Motion Versus Apparent Motion	56
4.5	Summary	58
5	Skin-Color Segmentation	61
5.1	Introduction	62

5.2	Previous Research	64
5.3	Generation of the Skin-Color Model	70
5.3.1	Manual Segmentation of Training Images	71
5.3.2	The Skin-Color Model	72
5.4	Generation of the Skin Detection Mask	76
5.4.1	Median Filtering	77
5.4.2	Pixel Classification	78
5.4.3	Derivation of the Segmentation Threshold	78
5.5	Simulation Results and Discussions	88
5.5.1	Performance Evaluation	89
5.5.2	Still Images	90
5.5.3	Video Sequences	92
5.6	Summary	94
6	Statistical Change Detection	101
6.1	Introduction	102
6.2	Previous Research	105
6.3	Change Detection Based on the F Test	108
6.3.1	The F Test	109
6.3.2	Estimation of the Background Sample Variance	111
6.4	Simulation Results and Discussions	118
6.4.1	Synthetic Frames	118
6.4.2	Real Frames	118
6.5	Summary	122
7	Segmentation and Tracking	127
7.1	FHSM Generation	128
7.2	Face Detection and Tracking	131
7.2.1	Face Detection	134
7.2.2	Face Tracking	140
7.3	Simulation Results and Discussions	141
7.3.1	FHSM Generation	143

7.3.2	Face Detection and Tracking	149
7.4	Summary	150
8	Conclusions and Future Work	153
8.1	Conclusions	154
8.1.1	Skin-Color Segmentation	154
8.1.2	Statistical Change Detection	155
8.1.3	FHSM Generation, Face Detection, and Tracking	155
8.2	Future Work	156
A	The Common Intermediate Format	159
B	Description of the Video Sequences	163
C	Additional Simulation Results	165
	Bibliography	175

Abstract

Sign language is a visual language used by deaf or hearing-impaired people to communicate. For distant communication, deaf people commonly use the text telephone, which is at least 10 times slower than sign language. Moreover, sign language is the first language of many pre-lingually deaf individuals, and its speed is comparable to that of normal speech. Video communication would allow deaf individuals to communicate remotely via sign language, providing them the equivalent of the telephone for individuals of normal hearing. Therefore, video communication would be a boon to the deaf community.

Block-based video coding strategies, the cornerstone of the H.261 and H.263 coding standards for video conferencing, are unsuitable for the transmission of sign language video over affordable low bit-rate channels. This is mainly due to the presence of rapid hand and arm motion in sign language video, as well as the necessity of smooth motion perception. Accordingly, sign language video will require content-based coding strategies to achieve the image quality and frame rate necessary for accurate perception. Using content-based coding, video sequences are typically segmented into different objects which may be independently coded and transmitted. More resources are allocated to the perceptually important objects, which in the case of sign language, are the face and hands.

In this thesis, a methodology is devised for the segmentation of the face and hands in sign language video sequences. As well as an improved coding performance, the content-based representation of video data would allow other functionalities, such as improved error-robustness and scalability. The proposed algorithm employs color and motion cues to segment the face and hands. First, a color segmentation algorithm is devised to locate skin-color regions in each frame. Second, we note that sign language is characterized by the motion of the hands and the face. Based on this observation, the proposed face and hand segmentation

methodology employs motion information to locate the moving skin-color regions in each frame. To this end, a statistical change detection method is proposed based on the F test and block-based motion estimation. In addition to the face and hand segmentation methodology, a face detection and temporal tracking method is also presented. This has applications in lip-reading, where more coding resources are allocated to the face.

The performance of the skin-color segmentation algorithm is demonstrated by simulations carried out on both still images and video sequences. The proposed change detection method is tested on four video sequences. The simulation results demonstrate the effectiveness of the proposed face and hand segmentation methodology, and the face detection and tracking method.