

ON HARNESSING VIDEO CONTENT

by

GAURAV HARIT

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of

Doctor of Philosophy

to the

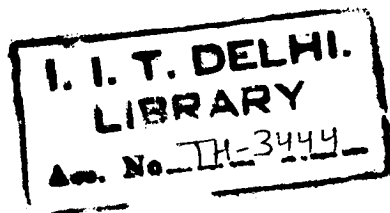


INDIAN INSTITUTE OF TECHNOLOGY, DELHI

INDIA

JANUARY 2007

. Video Content



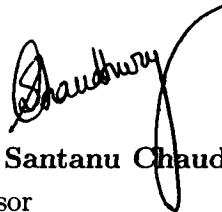
TH
681.846.7
MAR-0



Certificate

This is to certify that the thesis titled “**On Harnessing Video Content**” being submitted by Gaurav Harit to the Department of Electrical Engineering, Indian Institute of Technology, Delhi, for the award of the degree of Doctor of Philosophy, is a record of bona-fide research work carried out by him under my guidance and supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.



Prof. Santanu Chaudhury
Professor
Department of Electrical Engineering
Indian Institute of Technology
New Delhi - 110 016

Acknowledgments

I am grateful..

- To my supervisor, Prof Santanu Chaudhury, for his guidance, encouragement, impetus, highly positive attitude, and for his deep insight into the problems and the ideas which have formed the core of this thesis. I owe him immense gratitude for allowing me to work at my own pace, for carefully reviewing and giving valuable feedback for all my papers so that they got accepted promptly, for supporting me financially, for funding all my conference travels, and for giving me full freedom in the Multimedia lab. His superior mentoring skills and efficient thinking have played a great role in pushing me through when I would reach a stagnation and keeping me actively involved. My acknowledgment in words will always stand pale to what I have inculcated while working under him.
- To my SRC members, Prof S. D. Joshi, Prof Prem Kalra, and Prof Subrat Kar, for their useful suggestions while monitoring the progress of my research.
- To Prof. J. K. Chatterjee, Head, Electrical Engineering Department, Prof A. N. Jha, Chairman DRC, and Dr. S Prakriya, for their prompt help whenever I approached them with a problem.
- To Mrs. Usha Bhola, technician in-charge of Multimedia Lab for arranging all what I needed, and for being a counsellor. To Mr. Fateh Singh, former technician in-charge of our lab for his care during the initial years of my PhD. To Sushilaji, helper in our lab, and to the mess staff of Kumaon hostel, where I resided for 5 years, for their cheerful greetings, good wishes, and every possible help.
- To Dr. Hiranmay Ghosh, for his useful suggestions on how to improve my work related to deploying video ontologies. To Geetika, for sharing her research experiences which counted a lot.
- To Venu Kesam, for making me most comfortable in his little sweet home, when I was having accommodation problems.
- To my friends, since my school days, my college times, and here at IIT Delhi,

who have inspired me to believe in myself. They will surely recognize themselves.

- To my relatives, who have always wanted that I'd be with them during all the holidays, but had to contend with just few hours.
- To my father Prof K. C. Harit, my mother Smt. Usha Harit, and my younger sister Pratibha, for stabilizing my mood swings, for boosting my tempo, and most of all for their prayers which have got me sail across the formidable ocean.

New Delhi

January 8, 2007


Gaurav Harit

Abstract

Video content is rich in information. Meaning from the video content can be extracted at different levels of abstraction. In this thesis, we develop techniques for processing and analyzing video at sub-object level, object level, and cognitive level. We extract spatio-temporal regions which depict homogeneity in color. A video shot segmented into tracks of homogeneous color regions is then processed in a grouping framework to identify the subjects. We then formulate a measure of prominence for the subjects and interpret the overall scene into general categories by taking into account the perceptual attributes of subjects as well as the context.

We also develop formalisms for representing video content knowledge in a formal ontological framework and deploying the knowledge for automatic semantic annotation using Bayesian network as a tool. The aggregate of techniques presented in this thesis is an attempt to harness video content by enabling effective conceptual access to the extracted semantic meta-data which characterizes video information.

Our contributions are briefly outlined as follows:

1. We have developed a novel clustering methodology, termed as Decoupled Semantics Clustering Tree (DSCT), which uses decoupled clustering in different feature sub-spaces. We have done extensive experiments that show that the proposed clustering methodology overcomes problems like model selection, cluster suitability, smoothing, etc in traditional clustering methods.
2. We have developed a perceptual grouping algorithm which makes use of a specified spatio-temporal grouping model and identifies the perceptual clusters or subjects in the scene. Gestalt principles have been used in the formulation of perceptual associations and a spatio-temporal coherence model, and thus our

grouping scheme is able to identify general classes of subjects which come out as perceptually distinct with strong associations between parts within the grouping. We have also shown how domain knowledge can be integrated with the grouping process to benefit detection of subjects, as well as to perform grouping with recognition.

3. We have developed a novel concept of perceptual prominence for subjects in the scene. Prominence is closely linked to how one cognitively interprets a scene. Our formalism for scene interpretation makes use of perceptual prominence and *mise-en-scène* features.
4. For representing video content knowledge in an ontology, we have proposed new constructs to the Web Ontology Language (OWL). We call our extensions as Multimedia-OWL (M-OWL). We propose a probabilistic reasoning framework to infer concepts in presence of uncertain relations between concepts and observable video features.
5. We develop technique for automatic creation of semantic-metadata using a Bayesian network which encapsulates the semantic content knowledge. Video annotations are exploited to provide conceptual access – browsing and querying in a video collection.

Contents

1	Introduction	1
1.1	Problem Definition	2
1.2	Our Approach to the problem	2
1.3	Chapter layout of the thesis	5
2	Video content analysis and description: A review	7
2.1	Outline of the chapter	8
2.2	Extraction of meaningful entities in video domain	9
2.2.1	Spatio-temporal segmentation	10
2.2.2	Clustering Approaches	13
2.2.3	Perceptual Grouping	16
2.2.4	Saliency evaluation of Structures	20
2.3	Scene Characterization and Interpretation	21
2.4	Representation of Video	23
2.4.1	Content structuring	23
2.4.2	Approaches towards structuring film videos: Motivation from Film Grammar	26
2.4.3	Representation of Video Content	29
2.5	Representation of Video Content Knowledge	34
2.5.1	Knowledge Representation of Concepts	34
2.5.2	Knowledge Representation of Events	38
2.5.3	Bridging the Semantic Gap	40
2.6	Accessing video content – Browsing and Querying	44
2.6.1	Presentation of Video Content for Browsing	44

2.6.2	Retrieval of video content	45
2.7	The thesis in perspective: Motivation and contributions	47
3	Clustering in Video data: Dealing with Heterogeneous Semantics of Features	51
3.1	The clustering problem	52
3.1.1	<i>k</i> -Means Clustering	52
3.1.2	Gaussian Mixture (GMM)	54
3.1.3	Mean shift clustering algorithm	56
3.2	The clustering problem in video	58
3.3	Decoupled Clustering of Data in a heterogeneous feature space	60
3.3.1	Partitioning in Feature Space	61
3.3.2	Designing the DSCT for Video Data	62
3.4	Clustering Methodology for each level of DSCT	67
3.4.1	Color Model for the Stack	67
3.4.2	Spatial Clustering Requirements and Clustering Algorithm	73
3.4.3	Meta-clustering for Temporal Coherence	77
3.5	Results	81
3.5.1	Traditional clustering: Model selection and cluster suitability	81
3.5.2	Traditional clustering: Changing Spread of Data Distribution	88
3.5.3	Results for DSCT clustering applied to Video	91
3.6	Conclusions	93
4	Perceptual grouping in spatio-temporal domain	99
4.1	Motivation for the proposed perceptual grouping scheme	100
4.2	Perceptual Grouping in Spatio-Temporal Domain	101
4.2.1	Motion Similarity Association	102
4.2.2	Adjacency Association	104
4.2.3	Cluster Bias Association	105
4.2.4	Self Bias Association	105
4.2.5	Configuration Stability Association	106

4.3	The Spatio Temporal Grouping Model: Computing the saliency for clusters	107
4.4	The Perceptual grouping Algorithm for identifying clusters	110
4.4.1	The grouping problem and the Perceptual Grouping Algorithm (PGA)	110
4.4.2	Complexity Analysis for PGA	114
4.4.3	Comparison of Perceptual Grouping with other forms of grouping	115
4.5	Results for identifying Perceptual Clusters using the Perceptual Grouping Algorithm (PGA)	116
4.6	Conclusions	128
5	Perceptual prominence and scene interpretation	131
5.1	Motivation for the proposed concept of Perceptual Prominence	132
5.2	Perceptual Prominence	133
5.2.1	Computation of prominence attributes and virtual evidences	135
5.2.2	Results of computing the Perceptual Prominence for some Video Scenes	139
5.2.3	Scene categorization using perceptual prominence	139
5.2.4	Results for scene categorization	145
5.3	Perceptual Grouping with Learned Object Models	147
5.3.1	Object Model as a pictorial structure	147
5.3.2	Learning the object model	148
5.3.3	Formulation of Appearance Parameters for object parts	149
5.4	Spatio-temporal grouping model incorporating object model knowledge	150
5.4.1	Some Results demonstrating use of Object Models with Perceptual Grouping	153
5.5	Conclusions	155
6	Encoding video ontology: Multimedia Extensions to Web Ontology Language (OWL)	157
6.1	Motivation for using Ontologies to represent content-knowledge	158
6.2	Requirements of a Multimedia Ontology	160

6.3	Extensions to OWL: Additional language constructs	161
6.3.1	Concepts and Media Observables	161
6.3.2	M-OWL Relations	162
6.3.3	Syntax for specifying the Spatio-temporal relations	168
6.3.4	Comparison with MPEG-7 based spatio-temporal event descriptions	172
6.3.5	Uncertainty Specification	173
6.4	Inferencing framework	174
6.5	Making the Observation Graph	175
6.6	An Illustrative Example: Ontology defining <i>Tajmahal</i>	179
6.7	Comparison with Existing Work	182
6.8	Implementation	186
6.9	Conclusions	186
7	Conceptual Access to Video Content	187
7.1	Semantic Annotation of Video Content	188
7.1.1	Construction of Spatio-temporal entity region hierarchy (<i>erh</i>)	190
7.2	Semantic Annotation algorithm	191
7.3	Annotation Structure for Video	191
7.4	Browsing and Querying Video content: SMIL presentation	193
7.5	Results and discussions	197
7.6	Conclusions	207
8	Conclusions	209
8.1	Summary of the work done in the thesis	209
8.2	Contributions	212
8.3	Further Research Directions	214