

# ANOMALY DETECTION, RULE ADAPTATION AND RULE INDUCTION METHODOLOGIES IN THE CONTEXT OF AUTOMATED SPORTS VIDEO ANNOTATION

A. Khan

Submitted for the Degree of  
Doctor of Philosophy  
from the  
University of Surrey



Centre for Vision, Speech and Signal Processing  
Faculty of Engineering and Physical Sciences  
University of Surrey  
Guildford, Surrey GU2 7XH, U.K.

July 2013

© A. Khan 2013

ProQuest Number:27610179

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 27610179

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346

# Acknowledgements

It's been a long, hard and often a very challenging journey over the span of four and a half years which apparently though far from easy, had been thus far truly wonderful. I would like to think that the most valuable lessons have been taught to me by my peers, teachers and colleagues over the course of my PhD study that doesn't only tick the minimal required objectives for a research degree but most importantly it has truly transformed my problem analysis skills. It has indeed been a privilege to have been granted this opportunity and as such I would like to extend my sincerest gratitude to a few important people, in the absence of whom, achieving these milestones might not have been possible.

First and foremost, I would like to thank both of my supervisors Prof. Josef Kittler and Dr. David Windridge. I would like to extend my heartfelt appreciation to Josef in the first place for granting me the opportunity to do a PhD as well as indispensable guidance and support. I am especially indebted to David for his invaluable support throughout my thesis with his patience, knowledge and the flexible 'open door policy' based meetings whilst allowing me the room to work in my own way. I truly attribute the level of my PhD degree to the consistent encouragement and unprecedented efforts of both of my supervisors without whom this achievement would not have been possible. One can hardly imagine of any better or friendlier supervisors.

I wish to thank my parents for playing a vital role in making me who I am today. Please allow me to offer my sincerest gratitude to both of you, thanks for believing my personal abilities and aptitude at times when I would personally lose faith in myself. Your never-ending prayers and consistent encouragement has enabled me with the greatest of achievements in life. I would also like to thank my wife, Sundus, for believing in me at the most difficult of times. To my elder brother and my childhood mentor, thank you for always being there for me. To my elder sister, thank you for all your encouragement and prayers.

During the course of my research project, ACASVA (Adaptive Cognition for Automated Sports Video Annotation), I have come across a whole lot of people who have provided extensive help with regards to my work-package objectives. I greatly acknowledge the generous help and advice that I received from Dr. William Christmas, Dr. Teofilo de Campos, Dr. Fei Yan, Dr. Ibrahim Almajai and Nazli FarajiDavar. I would also like to gratefully acknowledge the support of EPSRC through grant EP/F069421/1 for providing me with the necessary funding to undertake this research degree.

And finally, last but not least I wish to thank my friends and colleagues at CVSSP including Dr. Affan Shaukat, Dr. Tariqullah Jan, Dr. Sanaul Haq, Dr. Buddaditya Goswami, Dr. Debaditya Goswami, Dr. Chi Ho Chan and Ashish Gupta for the various formal and informal discussions that have always proven to be useful and enjoyable. To all my cousins, aunts and uncles in Pakistan and United Arab Emirates, thank you for all the support and well-wishes.

## Summary

Automated video annotation is a topic of considerable interest in computer vision due to its applications in video search, object based video encoding and enhanced broadcast content. The domain of sport broadcasting is, in particular, the subject of current research attention due to its fixed, rule governed, content. This research work aims to develop, analyze and demonstrate novel methodologies that can be useful in the context of adaptive and automated video annotation systems. In this thesis, we present methodologies for addressing the problems of anomaly detection, rule adaptation and rule induction for court based sports such as tennis and badminton.

We first introduce an HMM induction strategy for a court-model based method that uses the court structure in the form of a *lattice* for two related modalities of singles and doubles tennis to tackle the problems of anomaly detection and rectification. We also introduce another anomaly detection methodology that is based on the disparity between the low-level vision based classifiers and the high-level contextual classifier. Another approach to address the problem of rule adaptation is also proposed that employs Convex hulling of the anomalous states.

We also investigate a number of novel hierarchical HMM generating methods for stochastic induction of game rules. These methodologies include, Cartesian product Label-based Hierarchical Bottom-up Clustering (CLHBC) that employs prior information within the label structures. A new constrained variant of the classical Chinese Restaurant Process (CRP) is also introduced that is relevant to sports games. We also propose two hybrid methodologies in this context and a comparative analysis is made against the flat Markov model. We also show that these methods are also generalizable to other rule based environments.

**Key words:** Anomaly Detection, Anomaly Rectification, Domain classification, Rule Induction, Chinese Restaurant Process (CRP), Hierarchical Hidden Markov Model (hHMM), Transfer learning

Email: Aftab.Khan@surrey.ac.uk

WWW: [http://www.surrey.ac.uk/cvssp/people/phd\\_students/aftab\\_khan/](http://www.surrey.ac.uk/cvssp/people/phd_students/aftab_khan/)

# Contents

|  |             |
|--|-------------|
| <b>List of Acronyms</b>  | <b>xiii</b> |
| <b>Mathematical Notation</b>   | <b>xv</b>   |
| <b>Declaration</b>   | <b>xix</b>  |
| <b>1 Introduction</b>  | <b>1</b>    |
| 1.1 Motivation . . . . .   | 1           |
| 1.2 Description of Problem . . . . .   | 3           |
| 1.3 Methodological Approach . . . . .  | 4           |
| 1.4 Aims . . . . .   | 5           |
| 1.5 Research Contributions . . . . .   | 6           |
| 1.5.1 Anomaly detection and rectification methodologies . . . . .  | 6           |
| Lattice-based Anomaly Detection & Rectification . . . . .  | 7           |
| Classifier Disparity based Anomaly Detection and Convex Hulling<br>of Anomaly States for Anomaly Rectification . . . . . | 7           |
| 1.5.2 Rule-Induction Methodologies . . . . .   | 7           |
| Cartesian-Product Label-Based Hierarchical Bottom-Up<br>Clustering (CLHBC) . . . . .                                     | 7           |
| Multi-Level Chinese Takeaway Process (MLCTP) . . . . .   | 8           |
| Hybrid Methods . . . . .   | 8           |
| 1.6 Thesis Structure . . . . .   | 8           |
| 1.7 Summary . . . . .  | 10          |

---

|          |   |           |
|----------|---|-----------|
| <b>2</b> | <b>Literature Review</b>  | <b>11</b> |
| 2.1      | Autonomous Vision-based Annotation Systems . . . . .                            | 11        |
| 2.1.1    | Sports Video Annotation . . . . .   | 12        |
|          | Visual Feature Extraction for Sports Video Annotation . . . . .                 | 13        |
|          | Game Evolution Tracking and Annotation . . . . .                                | 15        |
| 2.2      | Anomaly Detection . . . . .   | 16        |
| 2.2.1    | Anomaly Detection in a Multi-Level Knowledge Representation Framework . . . . . | 19        |
| 2.2.2    | Rule Adaptation: Dealing with Meaningful Novel Events . . . . .                 | 21        |
| 2.3      | Rule Induction . . . . .  | 22        |
| 2.3.1    | First-Order Rule Induction . . . . .  | 22        |
| 2.3.2    | Stochastic Rule Induction . . . . .   | 23        |
| 2.4      | Conclusions . . . . .   | 24        |
| <b>3</b> | <b>Computer Vision Systems for Deriving Experimental Datasets</b>               | <b>27</b> |
| 3.1      | Introduction . . . . .  | 27        |
| 3.2      | The Tennis Video Annotation System . . . . .                                    | 28        |
| 3.3      | Tennis and Badminton Ground Truth Annotation System . . . . .                   | 34        |
| 3.4      | Driving Intention Manual Annotation System . . . . .                            | 36        |
| 3.5      | Other Datasets . . . . .  | 36        |
| 3.6      | Summary . . . . .   | 39        |
| <b>4</b> | <b>Anomaly Detection and Rectification for Knowledge Transfer</b>               | <b>41</b> |
| 4.1      | Introduction . . . . .  | 41        |
| 4.2      | Lattice Based Anomaly Detection and Rectification . . . . .                     | 43        |
| 4.2.1    | Introduction . . . . .  | 43        |
| 4.2.2    | Anomaly Detection and Rectification in Court Game Environments . . . . .        | 44        |
| 4.2.3    | Methodology . . . . .   | 49        |
| 4.2.4    | Implementation Protocol and Experimental results . . . . .                      | 56        |
| 4.2.5    | Conclusion . . . . .  | 62        |
| 4.3      | Dirac-based Anomaly Detection and the Convex Hull . . . . .                     | 63        |
| 4.3.1    | Introduction . . . . .  | 63        |
| 4.3.2    | Weak Classifiers . . . . .  | 65        |

---

|  |           |
|--|-----------|
| Ball event recognition . . . . .   | 65        |
| Action recognition . . . . .   | 67        |
| Bounce position uncertainty . . . . .  | 69        |
| Combining evidence . . . . .   | 70        |
| 4.3.3 Context Classification . . . . .   | 70        |
| 4.3.4 Experiments . . . . .  | 71        |
| 4.3.5 Data association/rule updating . . . . .   | 74        |
| 4.3.6 Conclusions . . . . .  | 75        |
| 4.4 Discussion and Summary . . . . .   | 76        |
| <b>5 Rule Induction</b>  | <b>79</b> |
| 5.1 Introduction . . . . .   | 80        |
| 5.2 Cartesian Product Label-Based Hierarchical Bottom-up Clustering . . .  | 83        |
| 5.2.1 Introduction . . . . .   | 83        |
| 5.2.2 Methodology . . . . .  | 84        |
| Topology selection criterion . . . . .   | 88        |
| Modeling hierarchical transitions . . . . .  | 88        |
| Worked example . . . . .   | 90        |
| 5.3 Multi Level Chinese Takeaway Process . . . . .   | 91        |
| 5.3.1 Introduction and Motivation . . . . .  | 91        |
| 5.3.2 Methodology . . . . .  | 93        |
| State Generation Phase . . . . .   | 93        |
| Topological State Transition Matrix Generation Phase . . . . .   | 94        |
| Hierarchical State Transition Matrix Injection Phase . . . . .   | 96        |
| Worked Example . . . . .   | 97        |
| 5.3.3 Induction Protocol . . . . .   | 102       |
| Jensen-Shannon Divergence . . . . .  | 103       |
| 5.4 Hybrid Models . . . . .  | 104       |
| 5.4.1 Multi-Level Chinese Takeaway Process with Recursive Baum-<br>Welch Estimated Hidden State Transitions (MLCTP-BW) . . . .   | 104       |
| 5.4.2 Multi-Level Chinese Takeaway Process with Cartesian Product<br>Label-Based Hierarchical Bottom-up Clustering Computed State<br>Transitions (MLCTP-CLHBC) . . . . . | 107       |
| 5.5 Experimental Results and Discussions . . . . .   | 108       |
| 5.6 Summary . . . . .  | 112       |

---

|          |                                |            |
|----------|--------------------------------|------------|
| <b>6</b> | <b>Summary and Future Work</b> | <b>117</b> |
| 6.1      | Thesis Summary . . . . .       | 117        |
| 6.2      | Future Work . . . . .          | 119        |
|          | <b>Bibliography</b>            | <b>123</b> |



# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | Example of an Anomaly Detection and Rectification system . . . . .  | 3  |
| 3.1 | A detailed diagram of the tennis video analysis system . . . . .  | 29 |
| 3.2 | A simplified diagram of the tennis video analysis system. . . . .   | 30 |
| 3.3 | An illustrative example of the composition of a tennis video. The length of each shot is proportional to the width of the corresponding block in the figure. . . . .                              | 30 |
| 3.4 | Shot types. (a) crowd. (b) close-up. (c) play. (d) commercial. . . . .  | 31 |
| 3.5 | The GUI of the tennis video annotation system. . . . .  | 32 |
| 3.6 | True tennis rule model as defined in [86, 88]. . . . .  | 33 |
| 3.7 | Tennis and Badminton Annotation Toolbox . . . . .   | 35 |
| 3.8 | Driver's Annotation System . . . . .  | 37 |
| 4.1 | Highlighted playing areas for singles and doubles tennis including the out areas . . . . .  | 45 |
| 4.2 | Tennis court lattice with a Tennis court mosaic and projection images on the left and two views (2-D and 3-D) of the constructed court lattice showing various levels of court box sizes. . . . . | 47 |
| 4.3 | Singles and Doubles Tennis court box transition at the smallest (i.e. indivisible) units . . . . .  | 51 |
| 4.4 | Singles to Doubles Tennis extension factor . . . . .  | 56 |
| 4.5 | Comparative lattice box activites for singles tennis (left) and derived doubles tennis (right) using the expansion factor shown in Figure 4.4 . .   | 57 |
| 4.6 | Gray-scale histogram of Singles ( $M^s$ ) transition counts over the lattice (ordered by box size and count-number, respectively) . . . . .   | 59 |
| 4.7 | Gray-scale histogram of Doubles ( $M^d$ ) transition counts over the lattice (ordered by box size and count-number, respectively) . . . . .   | 60 |

---

|      |  |     |
|------|--|-----|
| 4.8  | Mean prediction error of simulated game area transformation for a given number of complete singles/doubles serve sequences (x-axis) . . . . .  | 61  |
| 4.9  | RMS residual over $b_B$ for $b_A = \text{play area}$ . . . . .   | 62  |
| 4.10 | Two examples of the final ball tracking results with ball event detection. Yellow dots: detected ball positions. Black dots: interpolated ball positions. Red squares: detected ball events. In the left example, there is one false positive and one false negative in ball event detection. In the right example, there are a few false negatives. . . . . | 66  |
| 4.11 | Sample images and detected players performing each primitive action of tennis. . . . .   | 68  |
| 4.12 | Doubles tennis ball tranjectory in a complete play-shot with a highlighted anomaly. . . . .  | 71  |
| 4.13 | Singles (validation dataset) - Confusion matrix of recognized events, K. We use the same labels as [5]. . . . .  | 72  |
| 4.14 | Doubles (test dataset) - Confusion matrix of recognized events, K . . . .  | 73  |
| 4.15 | Number of event sequences of the validation set that contain errors with varied confidence thresholds . . . . .  | 74  |
| 4.16 | Detected anomaly triggering events of the test set for Tennis Doubles .  | 75  |
| 5.1  | Cartesian Product Label-Based Hierarchical Bottom-up Clustering . . .  | 85  |
| 5.2  | Three Level Cartesian Product Label-Based Hierarchical Bottom-up Clustering with Transition Matrices generated at each level (colored so as to indicate heredity) . . . . .  | 92  |
| 5.3  | Multi-Level Chinese Takeaway Process; Example topology with $H = 3$ and $\mathcal{G} = 7$ (i.e. $\mathcal{O} = 1, 2, 3, \dots, 7$ ) . . . . .  | 98  |
| 5.4  | Block Diagram for Multi-Level Chinese Takeaway Process . . . . .   | 102 |
| 5.5  | Multi-Level Chinese Takeaway Process with Baum-Welch Hidden State Transition Estimation . . . . .  | 105 |
| 5.6  | Multi-Level Chinese Takeaway Process - Cartesian Product Label-Based Hierarchical Bottom-up Clustering . . . . .   | 108 |
| 5.7  | Badminton Dataset - Comparative mean prediction with standard deviation . . . . .  | 112 |
| 5.8  | Tennis Dataset (Video Annotation System) - Comparative individual event prediction accuracies for all of the five methods. . . . .   | 113 |
| 5.9  | Badminton Dataset - Confusion matrices for all of the five methods . . .   | 114 |

# List of Tables

|     |  |     |
|-----|--|-----|
| 3.1 | Summary of tennis events from [89] . . . . .                         | 34  |
| 3.2 | Summary of Badminton events . . . . .                                | 36  |
| 3.3 | Summary of Driving Events used in [154] . . . . .                    | 37  |
| 3.4 | Summary of Website Events used in [21] . . . . .                     | 38  |
| 3.5 | Summary of Website Events used in [72] . . . . .                     | 38  |
| 5.1 | Summary of Badminton and Tennis events extracted from [86] . . . . . | 86  |
| 5.2 | Sports datasets with source information . . . . .                    | 109 |
| 5.3 | Sports datasets with the number of samples per event label . . . . . | 109 |
| 5.4 | Datasets Description . . . . .                                       | 110 |
| 5.5 | Mean Accuracy . . . . .  | 110 |

# List of Acronyms

| Acronym/Abbreviation   | Meaning   |
|------------------------|---|
| <b>DIRAC</b>           | Detection and Identification of Rare Audio-visual Cues          |
| <b>HMM</b>             | Hidden Markov Model   |
| <b>CRP</b>             | Chinese Restaurant Process                                      |
| <b>SB Construction</b> | Stick-Breaking Construction                                     |
| <b>HDP</b>             | Hierarchical Dirichlet Process                                  |
| <b>CLHBC</b>           | Cartesian product Label-based Hierarchical Bottom-up Clustering |
| <b>MLCTP</b>           | Multi-Level Chinese Takeaway Process                            |
| <b>BW</b>              | Baum-Welch  |
| <b>ILP</b>             | Inductive Logic Programming                                     |
| <b>SVM</b>             | Support Vector Machine  |
| <b>ACASVA</b>          | Adaptive Cognition for Automated Sports Video Annotation        |
| <b>RMS</b>             | Root Mean Square  |
| <b>HOG</b>             | Histogram of Oriented Gradients                                 |
| <b>KLDA</b>            | Kernel Linear Discriminant Analysis                             |
| <b>BoW</b>             | Bag of Words  |
| <b>JS Divergence</b>   | Jensen-Shannon Divergence                                       |
| <b>KL Divergence</b>   | Kullback-Leibler divergence                                     |
| <b>MDL</b>             | Minimum Description Length                                      |

# Mathematical Notation

## Chapter 4: Anomaly Detection and Rectification Methodologies for Knowledge Transfer in the Context of Automated Sports Video Annotation

| Symbol  | Meaning   |
|---|---|
| $H$   | Horizontal screen lines, $H = \{(h_1, h_2, \dots, h_{n_h})\}$                                   |
| $V$   | Vertical screen lines, $V = \{(v_1, v_2, \dots, v_{n_v})\}$                                     |
| $b$   | 4-tuple box designation, $b \in \{(h_\alpha, v_\alpha, h_\beta, v_\beta)\}$                     |
| $b_o$   | Old play area definition  |
| $b_n$   | New Old play area definition  |
| $\mathcal{P}$   | Set of play areas   |
| $M^s$   | Singles Tennis' transition matrix   |
| $M^d$   | Doubles Tennis' transition matrix   |
| $T$   | Matrix transform, $T(M^s, b_o, b_n) = M^d$  |
| $D$   | Distance measure  |
| $A(x)$  | Activity measure over all of the observed transitions <i>into</i> and <i>out of</i> the box $x$ |
| $E[A(b_1 b_2)]$   | The expectation of the coarse activity measure $A$ in box $b_1$ due to activity in box $b_2$    |
| $M_{add}, M_{sub}$  | <i>Lattice interaction</i> matrices   |
| $\gamma$  | Parameter representing the appropriate proportion of activity to transfer                       |
| $c1, c2$  | Distances to the inner and outer tram-lines from the center respectively                        |
| $\Lambda = \{\lambda_1, \dots, \lambda_k, \dots, \lambda_K\}$ | Continuous-density left-to-right first-order HMMs   |
| $t$   | Time  |
| $\mathbf{o}_t$  | Observation at time $t$   |
| $\mathbf{x}_t, \dot{\mathbf{x}}_t, \ddot{\mathbf{x}}_t$       | Ball position, velocity and acceleration at time $t$ respectively                               |
| $\theta$  | State transition probability distribution matrix  |
| $\eta$  | Observation probability distribution  |
| $\pi$   | Initial state distribution  |
| $S = (s_1, s_2, \dots, s_N)$                                  | $N$ number of hidden states   |
| $G_j$   | Gaussian mixture components   |
| $\mathbf{O}_t$  | Ball information observation sequence   |
| $p(e)$  | Measurement error function  |

## Chapter 5: Rule Induction in the Context of Automated Sports Video Annotation

| Symbol   | Meaning   |
|--|---|
| <b>CLHBC</b>   |   |
| $\mathcal{E}$  | Sequence of events  |
| $\Omega_i$   | Label components indexed by $i$   |
| $L_t^{\{k\}}$  | Event label at time $t$ composed with the omitted set of labels $\{k\}$   |
| $Q_n^h$  | CLHBC-defined hidden state number $n$ at level $h$  |
| $P_f(Q_t Q_{t-1})$                                       | Transition likelihood between states $Q_{t-1}$ and $Q_t$ for a standard ‘Flat’ Markovian model                              |
| $f, g$   | Indicator functions   |
| $S_n$  | Observed sequence at level $n$  |
| $Q_n^h$  | Observed state number at level $h$ of the hierarchy   |
| $G$  | Total number of leaf nodes in CLHBC model   |
| $\bigwedge^C(\mathcal{E}_X \mathcal{E}_{X-1})$           | Augmented likelihood for CLHBC generated state transition between labels $\mathcal{E}_{X-1}$ and $\mathcal{E}_X$            |
| <b>MLCTP</b>   |   |
| $\alpha, \gamma$   | MLCTP’s concentration parameters  |
| $\mathcal{G}$  | MLCTP’s truncation parameter  |
| $H$  | MLCTP-defined number of levels  |
| $o_c$  | Number of people at takeaway $c$ in MLCTP model   |
| $\mathcal{C}^h$  | Number of states at level $h$ in MLCTP  |
| ${}_x\zeta_y^h$  | MLCTP generated state number $y$ , with parent state number $x$ and at level $h$  |
| $\pi$  | Stick-breaking construction weights for MLCTP transition probabilities  |
| ${}_x\delta_y^h$   | Self transition probability for state ${}_x\zeta_y^h$   |
| ${}_x\psi_y^h$   | Remaining transition probability for state ${}_x\zeta_y^h$ i.e. $1 - {}_x\delta_y^h$  |
| $\mathcal{G}$  | Total number of leaf nodes in MLCTP model   |
| $s$  | Number of $\mathcal{G}$ -defined selected Topologies  |
| $p$  | Number of random permutations for observation label association   |
| $\tau_s^p$   | $s$ -th selected topology in it’s $p$ -th permutation   |
| $Z$  | Total number of MLCTP-generated topological transition matrices   |
| $\bigwedge^U(\mathcal{E}_\sigma \mathcal{E}_{\sigma-1})$ | Augmented likelihood for MLCTP generated state transition between events, $\mathcal{E}_{\sigma-1}$ and $\mathcal{E}_\sigma$ |
| $Y(J_{tr}, J_Z)$   | Jensen-Shannon divergence between matrices $J_{tr}$ and $J_Z$   |
| $KL(M_1    M_2)$   | KL-divergence between two vectors, $M_1$ and $M_2$  |
| $K$  | Average distribution of two sources   |
| <b>Hybrid Model</b>                                      |   |
| $\lambda_h$  | HMM parameters for level $h$  |
| $a_{ij}^h$   | Transition probability of a state transiting from state $i$ to $j$ for level $h$  |
| $e_i^h(\cdot)$   | Emission probability for each level $h$ i.e. the probability of state $i$ at level $h$ emitting a symbol at level $h + 1$   |

---

|               |   |
|---------------|---|
| $\eta^h(i)$   | Initial distribution of states for each level $h$ defined by MLCTP  |
| $A_{ij}^h$    | Estimated state transition probabilities from state $i$ to $j$ for level $h$  |
| $F(t, i)$     | Probability of the model emitting symbols when in state $i$ upto time $t$ , obtained using the Forward algorithm                              |
| $B(t + 1, j)$ | Probability of the model emitting the remaining sequence if the model is in state $j$ at time $t + 1$ , computed using the Backward algorithm |
| $\{Q_t^h\}$   | Sequence of observations at level $h$   |

# Declaration

Research carried out in this thesis has been published or might be published in the following conference/journal proceedings:

- 7 Aftab Khan, David Windridge and Josef Kittler, **Multi-Level Chinese Takeaway Process and Label-Based Processes for Rule Induction in the Context of Automated Sports Video Annotation**, [Under Review], *IEEE Transactions on Cybernetics*, 2013.
- 6 David Windridge, T. E. deCampos, Fei Yan, William Christmas, Josef Kittler, A. Khan, **Rule Induction for Adaptive Sport Video Characterization Using MLN Clause Templates**, [Under Review], *IEEE Transactions on Multimedia*, 2013.
- 5 T. E. de Campos, Aftab Khan, Fei Yan, Nazli FarajiDavar, David Windridge, Josef Kittler, and William Christmas, **A framework for automatic sports video annotation with anomaly detection and transfer learning**, In *Proceedings of Machine Learning and Cognitive Science, EUCogIII conference, Palma de Mallorca, Spain*, 2013.
- 4 Ibrahim Almajai, Fei Yan, Teofilo de Campos, Aftab Khan, William Christmas, David Windridge and Josef Kittler, **Anomaly Detection and Knowledge Transfer in Automatic Sports Video Annotation**, In *Studies in Computational Intelligence*, isbn = {978-364224033-1}, 2012.
- 3 Aftab Khan, David Windridge, Teofilo de Campos, Josef Kittler and William Christmas, **Lattice-based Anomaly Rectification for Sport Video Annotation**, In *Proceedings of International Conference on Pattern Recognition (ICPR), Istanbul, Turkey*, 2010.
- 2 Ibrahim Almajai, Fei Yan, Teofilo de Campos, Aftab Khan, William Christmas, David Windridge and Josef Kittler, **Anomaly Detection and Knowledge Transfer in Automatic Sports Video Annotation**, In *Proceedings of DIRAC Workshop, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Barcelona, Spain*, 2010.
- 1 Ibrahim Almajai, Josef Kittler, Teofilo de Campos, William Christmas, Fei Yan, David Windridge and Aftab Khan, **Ball Event Recognition using HMM for Automatic Tennis Annotation**, In *Proceedings of International Conference on Image Processing (ICIP), Hong Kong*, 2010.



# Chapter 1

## Introduction

### 1.1 Motivation

There has been a significant growth in multimedia data production over the past decade. This data exists in various forms such as broadcast content (including television news and sports), personal content (e.g. uploaded mobile phone footage and social media videos), recorded interviews or meetings and footage from surveillance cameras etc. Due to easy availability of high quality digital hand-held devices such as cameras, mobile phones and camcorders, there has been a significant expansion in digital video production at a domestic consumer and industrial level.

Most of this data is intended for general viewing and hence basic labelling (Date, Time and Title etc.) is attached to it. However in many cases it would be useful to attach additional labels to retrieve information in a more flexible and systematic fashion (e.g. a tennis sports video can potentially be labelled with match-events description). Such meta-data will assist in finding material within the multimedia footage via browsing, querying or searching.

For easy retrieval of information from a very large quantity of archived footage, it would be very useful to have them annotated *automatically* i.e. to create a system that could understand the content of the video (manual annotation being too unwieldy). Sports videos have a high demand for automatic annotation as there is considerable interest

in browsing key events (such as goals in football etc.). These annotations may also be used to extract match statistics and performance analysis of teams.

Recent advances in computer vision and machine learning as well as the exponential growth in the processing capacity and memory of computer technology have created the conditions where it becomes pertinent to investigate the possibility of designing such systems. Sports footage provides a useful test-ground given its fixed, rule-governed content.

Sports videos also consist of rich multimedia content, as well as contextual details. Key temporal event information is critical in understanding sports videos. Furthermore, such an intelligent system can be made generic by extracting rule structures associated with the input footage. This involves dealing with the problems of anomaly detection and rectification i.e. when the input domain changes and the existing knowledge base becomes redundant for the new scenario e.g. in the context of domain classification when switching from a “Tennis environment” to “Badminton environment”. Additionally the system must also be capable of inferring high-level arbitrary rule structures for eventual meaningful annotations using rule induction methodologies.

Anomaly detection has received a large commercial interest due to its generic applicability in a vast range of sectors such as in surveillance where anomalies are referred to as a behaviour deviating from normality. Anomaly detection systems are also of high interest in various other applications such as fault detection in engineering systems. A broad overview of anomaly detection in the literature is presented in Chapter 2.

Rule induction, being a major subcategory of machine learning, also has received substantial interest due to its generic nature. Rule induction methodologies allow systems to extract formal rules using the input data with applications such as in making credit decisions for loan companies and in automatic classification of celestial objects. In the context of an expert system capable of decision making, rule induction techniques have helped in preventing breakdowns in electrical transformers via inferring faults from symptoms and suggesting corrective actions.

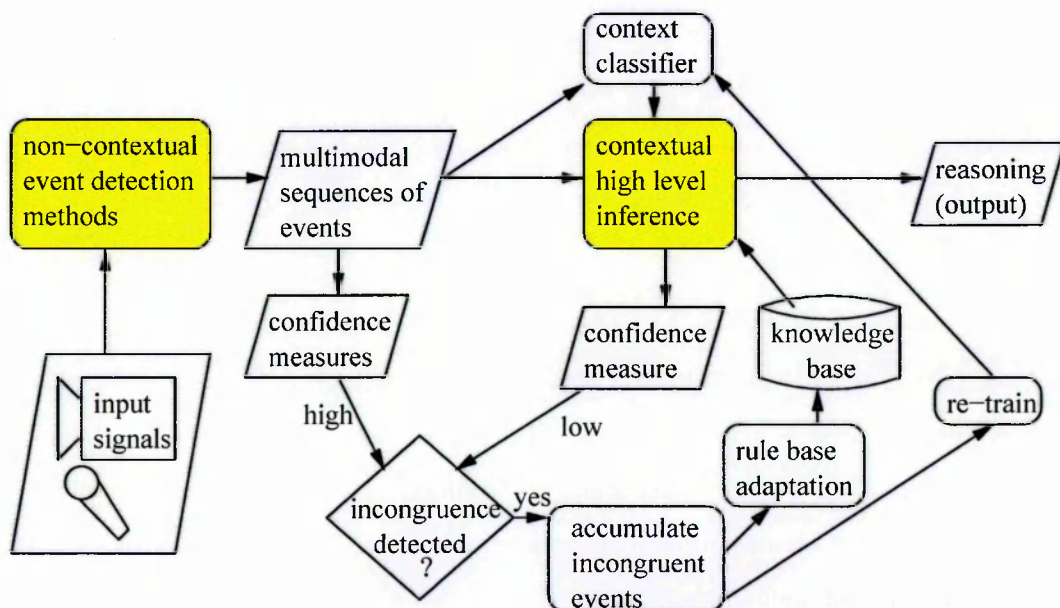
Automated video annotation also has various other relevant applications, apart from its usage in video search engines. It can significantly improve the performance of object-

based video encoding, enhance broadcast video streaming, assist analysis of player tactics in a sports video and serve as a basis for computer aided coaching.

## 1.2 Description of Problem

Autonomous sports video annotation systems have been recently developed which are generally domain specific (e.g. [35, 130, 196] for Soccer, formula 1 and snooker respectively). There is a real need for making such systems adaptive so that these systems can annotate contextually similar but visually different games. This entails significantly difficult problems such as either incorporating prior domain knowledge or knowing what parts of the knowledge base may be shared among different domains.

The first problem to be addressed in the realization of such an adaptive system with knowledge transfer capabilities is the notion of Anomaly detection i.e. detecting incongruence when a new domain is introduced to the existing knowledge base. Anomalies can be defined as those descriptive annotations which are incongruent according to the knowledge base but which are different from mere errors or outliers [101, 102] (see Figure 1.1).



**Figure 1.1:** Example of an Anomaly Detection and Rectification system

Anomaly rectification is a related problem of dealing with the detected anomalies i.e., either incorporating them as additional rules in the knowledge base or discarding them. New methodologies are required in the context of an adaptive system able to switch the knowledge base according to the new learning environment.

Another problem addressed in this thesis is of modelling the observed video sequence in terms of a rule structure i.e. the problem of rule induction. In an adaptive annotation system, it is crucial to have such a contextual level that captures observations in the form of concrete rules. Once there is such a model, the problems of anomaly detection and rectification can be more robustly tackled i.e. the system will be able to switch the rule base according to the observations. Such a system will also be able to transfer learning from one domain to another reducing the need for re-training.

### 1.3 Methodological Approach

An automated video annotation system must be able to determine at what point the existing learned model ceases to apply, and secondly, what aspects of the existing model can be brought to bear on the newly-defined learning domain. *Anomalies* must thus be distinguished from mere *outliers*, i.e. cases in which the learned model has failed to produce a clear response; it is also necessary to distinguish novel (but meaningful) input from misclassification error within the existing models. We thus propose methodologies to tackle the related problems of anomaly detection, knowledge transfer and rule induction for changing domains primarily in the context of automated sports video annotation.

Firstly, we introduce in Chapter 4, a Lattice-based method to solve the problem of anomaly detection and rectification that exploits the implicit court structure of sports games like tennis. This is achieved via a novel lattice-based Hidden Markov Model induction strategy for arbitrary court-game environments. We test the ability of the method to adapt to a change of rule structures going from tennis singles to tennis doubles (using real and simulated data).

We also introduce a methodology for anomaly detection based on the disparity between

the low-level vision based classifiers and the high-level contextual classifiers ([101, 102, 184]) to deal with the problem of detecting rule-incongruence involved in the transition from singles to doubles tennis videos. We then demonstrate how the detected anomalies can be used to transfer learning from one (initially known) rule-governed structure to another employing a convex hull of anomaly states for anomaly-adaptation/rule-update.

We also investigate a number of novel hierarchical HMM [47] generating methods for rule induction in the context of automated sports video annotation including a novel Cartesian Product Label-based Hierarchical Bottom-up Clustering (CLHBC) method that employs prior information contained within label structures. We also propose a new variant called, the Multi-Level Chinese Takeaway Process (MLCTP), based on the classical Chinese Restaurant Process [4] and Stick-Breaking construction [153] for rule induction. We also present two hybrid methodologies that uses MLCTP and the label structures in two different settings. We compare all of these methodologies against the flat Markov model to show that the methods proposed are generalizable to other rule based environments.

## 1.4 Aims

The research carried out in this thesis focuses on developing a framework for adaptive and automated sports video annotation via addressing the problems of anomaly detection and rectification, and rule induction. The first proposed methodology is a tennis court lattice-based method that discovers anomalies when introducing contextually dissimilar video to the system. This is demonstrated by starting with a system trained on ‘singles’ tennis matches, and then changing to a new input material in the form of a doubles tennis match i.e. one in which only the rule structures differ. We also show another anomaly detection mechanism via quantifying the disparity between different classifiers. A framework for the related problem of rule adaptation is also introduced to update the rule base for incorporating detected anomalies which is crucial for a generic and automated video annotation system.

Anomaly rectification for the lattice-based method, in the case of a tennis match, is achieved by redefining the play area. We show results for real and simulated tennis

singles and doubles games. We also present a related methodology for the anomaly adaptation problem using a convex hull of anomaly states usable in the context of tennis matches.

Sports games are inherently hierarchical in nature thus we also investigate methodologies for creating hierarchical HMMs [47] for rule induction. This is achieved initially via a method based on useful information contained within label structures. We also present novel stochastic rule induction methodology based on the classical Chinese restaurant process. A comprehensive evaluation is carried out for each of the proposed rule induction methodologies against a baseline method in the context of sports and various other domains.

These rule induction methodologies are predictively evaluated but they can also be predictively and retrospectively employed to detect any missing events in the context of an automated annotation system. Furthermore, this potentially enables the system to detect rule-based anomalies and make autonomous decisions on switching the rule base. Rule induction methodologies can also be used in the context of knowledge transfer, determining the amount of knowledge shared between various domains.

The aim of this thesis is thus to provide enabling technologies for use in automated, adaptive sports annotation systems.

## 1.5 Research Contributions

The novel contributions of this thesis in the area of automated sports video annotation and machine learning more generally are summarized as follows:

### 1.5.1 Anomaly detection and rectification methodologies

We present two new anomaly detection and anomaly rectification methodologies for automated sports video annotation in this thesis. These methodologies are briefly introduced as follows:

---

### **Lattice-based Anomaly Detection & Rectification**

Anomaly detection for changing domains is initially addressed in the context of the automated annotation of tennis singles and doubles games. Anomaly detection is employed as a means of determining, in an unsupervised manner, whether the rule base has changed in a fundamental way as differentiated from e.g. visual changes due to different international venues. This may require continuous adaptive learning to be abandoned and a new learning process initiated in the new domain. We also address a related problem of anomaly rectification i.e. the adaptation of the existing learning mechanism to the change of domain. As a concrete instantiation of this notion, a novel lattice-based HMM induction strategy for arbitrary court-game environments is proposed.

### **Classifier Disparity based Anomaly Detection and Convex Hulling of Anomaly States for Anomaly Rectification**

We apply another methodology for anomaly detection that is based on comparing the outputs of strong and weak classifiers [101, 102, 184] to address the problem of detecting the rule-incongruence involved in the transition from one domain to another. The strategy for anomaly rectification employs a convex hull of the detected anomalies which are then incorporated in the rule base appropriately.

#### **1.5.2 Rule-Induction Methodologies**

We also present four novel rule-induction methodologies in this thesis that are primarily developed in the context of automated sports video annotation. These methods are briefly introduced as follows:

#### **Cartesian-Product Label-Based Hierarchical Bottom-Up Clustering (CLHBC)**

Sports games have a specific structure built around temporal events that are based on transitions between labelled states according to structured game rules. By taking the

whole sequence of event labels into account, we thus represent rule-related information by using the Cartesian combinations of these sub-labels where they collectively constitute a *lattice* subset in which coarse-grained event labels are clustered bottom-up to form a hierarchical topology that can potentially represent abstract rule structures.

### Multi-Level Chinese Takeaway Process (MLCTP)

Court-games are inherently hierarchical in nature and we attempt to create stochastic approximations of the game rules using hierarchical Hidden Markov Models (hHMMs) for contextual game description covering various levels of abstractions, ultimately, giving rise to meaningful annotations. We propose a constrained variant of the widely used Chinese Restaurant Process (CRP) first introduced in [4] in conjunction with a Stick-Breaking construction [153] that allows us to establish rule structures that are capable of describing the sports game in a compact and efficient fashion. We refer to it as the Multi-Level Chinese Takeaway Process (MLCTP).

### Hybrid Methods

MLCTP does not intrinsically exploit labeled states and we speculate that the highest likelihood inferred rule structure given a set of hyper-parameters representing the MLCTP model can be further improved via employing the label structures. Thus, we also propose two hybrid methods that combine the unlabeled MLCTP with the labelled structure using Baum-Welch [134] hidden state transition estimation and CLHBC's label structure computation. These models effectively use the stochasticity of MLCTP whereby various hierarchical structures are produced, in conjunction with the labels containing important sequential information.

## 1.6 Thesis Structure

This thesis is structured as follows:

- **Literature Review:** Chapter 2 introduces a number of key concepts related to our research work, i.e., automated vision-based annotation systems, anomaly



detection strategies and anomaly rectification mechanisms reported in the literature. We also present various types of rule induction methodologies employed in machine learning applications.

- **Video Annotation Systems:** Chapter 3 briefly introduces all of the various datasets that we use for our experiments. These datasets include an automated computer vision-based annotation system of [89] (introduced in this chapter) and for which most of our novel algorithms are primarily designed. In addition to this system, we also introduce a purpose-built ground truth annotation system for labeling *key events* in Tennis and Badminton with information like *Serve* and *Hit* etc. Similarly, we also introduce another manual annotation system capable of labeling human *driving* for a car driven across a city with labels such as *turn left* and *signal right* etc.
- **Anomaly Detection and Rectification:** In chapter 4, we propose methodologies dealing with the problem of anomaly detection for triggering domain change when a new domain is presented to the system i.e., knowing when to switch the adaptive system to be able to annotate a different game. We attempt to do this in a sports environment in particular court games where players follow certain rules of the game with respect to a fixed court reference.
- **Rule Induction:** In Chapter 5 we investigate a number of novel hierarchical HMM generating methods for rule induction in the context of automated sports video annotation including the Multi-Level Chinese Takeaway Process (MLCTP) based on the Chinese Restaurant Process and a novel Cartesian Product Label-based Hierarchical Bottom-up Clustering method that employs prior information contained within label structures. We also present two hybrid methodologies in this chapter and make comparisons against the flat Markov model. We also show that the methods proposed are generalizable to other rule based environments.
- **Summary and Future Work:** Chapter 6 presents conclusions drawn from our research results, along with potential directions for future research in this area.

## 1.7 Summary

This chapter presented a brief introduction into the problems of anomaly detection, rule adaptation and rule induction methodologies and the ever growing need of adaptive systems. The problem of anomaly detection has motivated research in various fields such as surveillance and fault detection in systems etc. Similarly, the problem of rule induction has also received substantial interest because of its generic nature. Due to an immense growth in multimedia data production, it would be useful in many cases to attach meaningful descriptions autonomously. This thesis, addresses the aforementioned problems in the context of an adaptive and automated sports video annotation systems. Several contributions and publications resulted from this work, including: *i)* a new tennis court-lattice based anomaly detection and rectification method, *ii)* classifier disparity based anomaly detection and related anomaly rectification methodology using convex hulling of anomalous states, and *iii)* four new rule induction methodologies including the Cartesian product label-based hierarchical bottom-up clustering method and a constrained variant of the classical Chinese restaurant process that allows us to establish rule structures capable of describing sports games like Tennis and Badminton etc.

# Chapter 2

## Literature Review

The following chapter reviews the state-of-the-art in autonomous vision-based annotation systems, and the key concepts and techniques utilized by this thesis, including other methodologies employed primarily for anomaly detection and knowledge transfer. Rule induction methodologies are also reviewed in this chapter.

### 2.1 Autonomous Vision-based Annotation Systems

Automated vision-based annotation systems have received much attention within the literature and has often been referred to as “video concept detection” [119], “high-level feature extraction” [155], or “video semantic analysis” [158]. The main goal of such a system is to assign related concepts in the form of meta-labels to video clips or video frames. Machine learning techniques are typically employed in this context as follows:

- 1 Pre-processing is performed for data preparation
- 2 Videos are segmented into short units referred to as shots
- 3 Low-level features are extracted for every relevant shot using object/agent detection and tracking techniques
- 4 Domain rules are used as priors to learn relationships between the detected low-level features and related concepts

Note, most video annotation systems use only a subset of these, typically the first three only.

In the context of an automated video annotation system, event detection and action recognition are considered most frequently [51, 113, 131, 177]. A wide variety of methods have been proposed in this context such as person detection and tracking [84, 169], articulated body tracking [67] that can be crucial for classifying actions in a static images environment [33, 43, 193].

Bag of Words (BoW) based approaches are some of the most commonly used techniques employed in this context such as in [182], a BoW-based approach is proposed that builds a global vectorial representation of a whole video sequence implicitly employing visual context.

In terms of application domains, surveillance [46], entertainment movies [104] and TV shows [128] have all been explored. Additionally, sign language recognition has also been tackled in [27, 145, 161, 186].

In the following sub-sections, we review some of the concepts related to automated sports video annotation systems by exploring techniques related to low-level visual feature extraction, and game evolution tracking and annotation.

### **2.1.1 Sports Video Annotation**

Sports videos have received much attention within the video annotation literature [37, 190] and various sports have been explored. Soccer is probably the most extensively researched sporting domain because it offers a wide range of challenging research problems such as tracking multiple players with high levels of occlusion [37, 40, 55, 57, 107, 138, 139, 176, 183, 188]. Cricket [24, 69] and snooker [34, 136] have also attracted significant attention for research problems involving gesture recognition of the umpires and video summarization.

Tennis games have also been processed for shot classification [37, 78, 79], within-shot event detection [31, 137], players' stroke type classification [122, 129], analysis of player tactics [199] and scene retrieval [30, 112, 171].

---

## Visual Feature Extraction for Sports Video Annotation

**Object Detection and Tracking:** Various sports have different objects of attention, such as the cricket ball in cricket, the tennis ball in tennis and the shuttlecock in badminton etc. To design an artificial system capable of modeling these games, it is necessary to track them with respect to specific reference points to obtain those key events which relate to the rule structures of the game. For example, in a game of tennis, it is necessary to know where the ball bounces within the court area to establish the notion of *in* and *out* in order to annotate the play shot with the information about the point structure of the game.

Tracking objects involves two main tasks; finding and then following the object of interest in a video sequence. For this purpose, object appearance and dynamics are generally considered in designing an object tracking technique. Track After Detection (TAD) and Track Before Detection (TBD) are the two main sub-categories related to object tracking approaches. In the first approach, object candidates are initially extracted which are then used for tracking via data association (for measuring origin uncertainty) and estimation (for dealing with measurement inaccuracy). TAD approaches are suitable for small and fast moving objects like tennis balls [38]. Other examples can be found in the defense sector where a point in a radar signal is tracked [10, 156]. TBD approach is suitable for tracking large and slow moving objects such as people tracking [68]. This is achieved by making an initial object position hypotheses which are then evaluated using the image unlike a typical TAD approach where images are discarded after object candidates are extracted.

Object tracking, thus, has been and remains a key area of research for automated sports video annotation systems achieved via employing a number of different techniques. For example, Pingali *et al.* [1] focus on real time tracking of a tennis ball using multiple cameras. Techniques similar to stereo matching are then performed to produce a three dimensional virtual view. In [58], an object tracking algorithm is proposed which is based on object contour prediction. As the object continues its motion in subsequent frames, an update to contour prediction is made in case of occlusion/dis-occlusion. This method is computationally less expensive compared to other region-based methods

where, generally, the video object is initially defined by the user and video segmentation is then performed with tools like the watershed transformation [13, 152, 180] to establish temporal correspondence between the extracted regions. This enables object tracking in subsequent frames [15, 52, 148].

Similarly, several tennis ball tracking algorithms are reported in the literature such as in the hawk-eye system [117], where the authors adopt a Track after Detection (TAD) approach. Model based tracking is performed by detecting court lines and using camera calibration, a 3D ball position is determined. Ball flight is then predicted for making match-related decisions assisting match referees.

**Agent Tracking and Description** In a sports environment, agent tracking is one of the most important problems reported in the literature. For example, in terms of court-based sports, agents are those entities which may be considered independently active within the play area, and which act upon objects to cause them move about the court area. Agents tend not to follow strict motion trajectories and hence tracking them in terms of extrapolated motion vectors may be sub-optimal. (Agents will, in fact, tend to act in terms of the game rules, rather than simplified physical rules).

Marszalek *et. al.* in [105] proposed an action recognition system using an SVM (Support Vector Machine) [178] based classifier where a bag-of-features (BoF) framework is implemented to perform the action and scene recognition tasks in the context of natural video. A BoF approach represents images as orderless collections of local features [123], originated from the BoW representation of words for textual information retrieval. There are generally three main tasks in such an approach; (i) Building vocabulary of visual features (words), (ii) Assigning extracted features to terms in the vocabulary using nearest neighbor [111] or related methods, and (iii) Generating term vectors via recording counts of each term and creating a normalized histogram.

**Reference Area** Almost all sports played have a definite area of play which may be partitioned into different squares, circles and rectangles. For example, in the game of tennis, the court structure with different lines and boxes can be considered as the key court reference points. As rules of the game are highly dependent on these low-level

---

structures of the game-play area it is important to have an accurate model available. These low-level visual features can be detected by e.g., Hough transform [38], corner detectors [59, 157] and edge detectors [22] etc. For example, in [90], corner points are initially detected in a tennis court using the SUSAN corner detector [157] which are used to match features between successive frames to calculate projective transformation parameters (i.e. homography). After the most likely correspondences between corner pixels in successive frames is established, the RANSAC algorithm [50] is applied to yield an accurate inter-field homography. Using extracted homography matrices, images are warped back to the reference coordinate frame. This is important in a court-game environment to establish a reference court area which is crucial for allocating correct match points.

### Game Evolution Tracking and Annotation

Automated sports video annotation requires not only a low-level feature extraction framework but also a high-level contextual annotation system. For this purpose, a system with multiple levels of abstraction is required e.g. [89], where this is referred to as game evolution tracking.

Various techniques have been developed for the purpose of creating a decision-making system based on graphical models; one of the most important model widely used is the Hidden Markov Model (HMM) [133]. HMMs belong to a subset of the much broader set of frameworks collectively known as Bayesian Networks (BNs) [99]. They can be defined as doubly stochastic processes where only one of the processes is observable; the underlying process, of the system states, cannot be observed but only inferred through existing observations. Hidden Markov Models can effectively model temporal sequences of data (e.g. stock market [61], audio/video signals [96], and patient's Electrocardiography (ECG) [91] etc.). Many variants of the classical HMM have been proposed (see e.g. [17, 18, 48, 53, 97, 150]). HMMs can provide a useful link between stochastic models and their graphical semantics [116].

In a sports related setting, various kinds of human reasoning processes can be applied to understand the dynamics of sports environment. This involves decision making in

terms of classifying between competing scenarios, important in the case of missing or erroneous data.

A machine vision-based system for Formula 1 racing sport videos has been proposed in [110] though the annotation performed in this sport is only at the camera shots level. Similarly, soccer has also been explored where HMM-based methods are employed [181, 188, 191]. Video sequences are segmented into various types of shots, e.g. in [188], authors have defined various types of shots (such as *play*, *break*, *close-up*, *global views* and *zoom-in* etc) for segmentation using *dominant color ratio* (e.g. grass pixels vs. non-grass pixels) and *motion intensity* (i.e. average magnitude of the effective motion vectors in a frame) as features.

In a similar context, Assfalg *et. al.* [8] have explored videos of sport related news; where shots of contrasting nature are present e.g. a sequence of anchor-persons and players playing a particular sport. Match highlight detection has also been addressed [9].

Tennis video annotation has also been performed by summarizing video contents using multiple cues [172]. This is different to the tennis video annotation system of [89] (introduced in Chapter 3), where videos are annotated with contextual meta-labels using low-level visual observations (e.g. players) and cues from different types of features (e.g. tennis ball trajectory).

In this thesis, we aim at providing enabling technologies for building an adaptive and autonomous multimedia annotation system capable of detecting domain change and rule learning.

## 2.2 Anomaly Detection

Anomaly detection refers to the problem of discovering samples with unexpected behavior in the data [25]. These samples are often referred to via different terms including *anomalies*, *outliers*, *discordant observations*, *exceptions*, *aberrations*, *surprises*, *peculiarities* or *contaminants* etc. depending upon the application domains. However,



---

anomalies and outliers are the most commonly used terms in the context of anomaly detection.

Anomaly detection has received considerable interest in the literature due to its wide range of potential applications. These applications are spread across a broad range of domains, spanning from medical applications to fault detection systems including financial services, medical diagnostics, behavior analysis, surveillance, and defense etc.

Due to the high relevance of anomaly detection technology in various application sectors, a large number of anomaly detection methods have been developed some of which are reviewed in this section. This has also resulted in a number of survey papers ([3, 65, 101, 102] etc.) in the literature that attempt to classify various types of anomaly detection methods.

Edgeworth in [39], presents discordant (i.e. incongruous) observations as those which present the appearance of differing in respect of their law of frequency relative to other combined observations. This led to the theory of robust estimation [66] via identifying outliers. Robust estimation generally refers to an estimation technique which is insensitive to small departures from the idealized assumptions which have been used to optimize the estimator [132] e.g. median is a more robust estimator of central value than the mean. Related to the problem of robust estimation methodology is the identification of outliers employed for anomaly detection in [11, 147].

An anomaly is classically defined as an outlier (representing *abnormal* behavior) with respect to some known *normal* distribution. Such anomalies are classified in various surveys of [3, 65, 101, 102] in the following fashion:

- Statistical [11, 64, 106, 142]
- Nearest neighbor [85]
- Classification [28, 29, 70, 71, 118, 143, 162, 179]
- Clustering [62, 63]

A more comprehensive and recent survey of Chandola *et. al.* in [25] has enhanced the aforementioned categorization by the following two additional classes of methodologies:

- Information theoretic [6]
- Spectral [197]

These approaches use different criteria to define incongruence (e.g. [6] considers it as minority detection by measuring a cost function that expresses *atypicalness* of clusters against the simplicity of the clustering) however, they essentially relate to the same notion of anomaly detection as defined above.

While defining various criteria for delineating anomalous data from the normal data, it is crucial to first define normality. The process of learning normality is driven by the available training data by initially representing just the normal data or both the normal and samples of anomalous data. Distribution functions are modeled by statistical approaches and a boundary of normal behavior is delineated by the classification methodologies. Learning normality can be achieved using a normal dataset with positive training data [135, 151, 165, 166, 167] or with negative training data i.e. the anomalous dataset [146, 163].

In practical scenarios, a normal training data contains samples of error that can be mistaken for anomalies. In [41], a learning method is proposed for detecting anomalies within a dataset that contains a large number of normal elements and relatively few anomalies. This is achieved using maximum entropy to estimate a probability distribution over the data and thereafter a statistical test is applied to detect anomalies. In [42], a comparative analysis has been made which favors learning a positive instances detector rather than learning a negative instances detector.

Xiang *et. al.* in [187], proposed an approach for online normal behavior recognition and anomaly detection in the context of surveillance videos. This is achieved by using a runtime accumulative anomaly measure to detect abnormal behavior based on an online Likelihood Ratio Test (LRT).

Furthermore, adding to the types of anomalies, normal data can also contain *contextual anomalies* defined as anomalies that are consistent within a specific context but otherwise fall into the category of *abnormality* (also referred to as *conditional anomalies* [159]).

---

Each data instance, in dealing with contextual anomalies, is defined using two sets of attributes [25, 149]: (i) Contextual attributes (i.e. determining the context for that data instance) and (ii) Behavioral attributes (i.e. defining the non-contextual features of a data point).

For example, in the context of a credit card fraud detection domain, *time of purchase* can be a contextual attribute while a *weekly shopping bill* of a person can be a behavioral attribute. Anomalies can be triggered when the shopping bill exceeds a certain amount (behavior) during a particular time (context).

Another example of contextual anomaly could be that of an ordered sequence of observations, where any single observation in the sequence may appear normal, but as a group, or jointly with its neighbors, the observation is an outlier [12, 74, 75, 92, 159]. Anomalies in sequences of symbolic data have been studied in [32] and spatial outliers in [164]. A Markov chain model has been applied to the problem of contextual anomaly detection in [194].

Anomaly detection in a multi-sensor system is a more complicated situation where anomalies can exist due to corrupted data, faulty sensors and interesting events such as intrusion [26, 36, 44, 160, 198]. For this purpose, a more sophisticated reasoning framework is required [127]; as such, these complicated anomalous situations cannot be dealt with by simple point anomaly detection.

### 2.2.1 Anomaly Detection in a Multi-Level Knowledge Representation Framework

None of the literature cited above addresses the problem of anomaly detection in a complex multilevel knowledge representation system such as a machine perception framework. In a single level system, the idea of anomaly is relatively straightforward where comparative analysis is required between the data against the reference normal data. In a system with a multilevel representation of knowledge (such as e.g. [14, 89]), each phenomenon will have more than one reference (normal model) depending on the number of levels of knowledge representation. Anomalies associated with disagreement among level-based interpretations of observations results in a completely new type of anomaly,

described as a *compound anomaly* in [80]. This disparity between various levels of knowledge representation is referred to as incongruence. There has been a very limited amount of work carried out in this context, with the exception of speech recognition as mentioned in [184].

The European Union project, DIRAC, was concerned with the detection of rare events in multi-level systems i.e. incongruence detection. In [184], Weinshall *et. al.* compare the outputs of non-contextual (i.e. “specific-level”) and contextual (i.e. “interpretation level”) classifiers. An incongruence (i.e. anomaly) flag is triggered by the disparity between the output of these classifiers. The approach follows efforts in out-of-vocabulary word detection [19]. When the weak classifier (in this case, the phoneme detector) delivers a phoneme hypotheses with confidence, and when the strong classifier (i.e. contextual classifier) rejects the sequence of detected phonemes due to the absence of the word they correspond to in the system vocabulary, a disagreement occurs between the two classifiers suggesting that an out-of-vocabulary word has been encountered instead of a noisy speech segment (which would have produced a low weak classifier output i.e. low confidence phoneme hypothesis).

The disparity between a contextual classifier with a low confidence output and a non-contextual classifier with high confidence output is measured to detect new subcategories of objects in [124, 184, 202]. One of the anomaly detection mechanisms presented in Chapter 4 is based on this type of anomalies.

**Anomaly detection in a multi-modal system** Anomaly detection in a multi-modal system is performed by measuring inconsistency between the outputs of various data channels. In [7], incongruence is detected in multi-modal information arising from a wearable audio-visual device when, for example, audio detection of a voice occurs in spite of the person in the field of view not moving their lips.

It also deals with inconsistent gender classification results when a male person speaks with a high-pitched voice leading to contradictions in the different modalities. This is achieved by using a similar notion of incongruence detection as above. Authors in [7] have also constructed a new hardware platform (containing a stereo panoramic vision

sensors and hearing aids) with the goal of assisting people with disabilities or a high cognitive load to deal with novel events.

The strong classifier, in this scenario, was trained for classifying sequences of phonemes and the weak classifier was trained to classify a particular set of words from the observation data. The posterior probabilities arising from these two classifiers (two domains) are then compared using techniques based on Kullback-Leiber (KL) divergence [121] highlighting discrepancy (i.e. incongruence) in the classifier outputs.

### 2.2.2 Rule Adaptation: Dealing with Meaningful Novel Events

Often, anomaly detection is motivated by the need for a system capable of adapting to new environments. In such a setting, anomaly may be manifest due to environmental changes causing data drift. Such discordant change can, when detected, be accommodated by habituation processes such as those exercised by humans [170], as discussed by Crook *et. al.* in [29].

Model updating and acquisition in the context of tracking in computer vision [200] exploits similar ideas of adapting to changing situations as in [201]. More recently though, in the context of dealing with anomalous events, [125] presents an approach for learning from incongruence where incongruence is used to indicate where to improve the model of the universe by incorporating the detected novel concepts. This is demonstrated in an experiment with human audio-visual detection by combining the incongruent data model with existing models to remove the incongruence.

Also, in [175], a transfer learning algorithm is employed to learn the parameters of the new incongruent event from very few labeled samples. The degree of incongruence of the new event is also evaluated. This is achieved by using a recently introduced Multi-model Knowledge Transfer algorithm (Multi-KT) employing an SVM-based model adaptation setting [174] that is able to select and weight appropriately prior knowledge coming from different categories resulting in regulation of transfer learning for evaluating the degree of incongruence of the new event.

In this thesis, anomaly detection and rule adaptation mechanisms are employed in the context of sports video annotation. This results in the realization of constructing an

adaptive and autonomous annotation framework capable of detecting the input domain and consequently using the right knowledge base for annotation. In Chapter 4, new methodologies are proposed for anomaly detection and rectification for court games like tennis.

## 2.3 Rule Induction

### 2.3.1 First-Order Rule Induction

Sport rules can be modelled in the form of first order logic. This can be achieved using inductive logic programming (ILP), which is a hybrid of machine learning and logic programming [114]. PROGOL is a popular ILP system explained in [115], where rule learning can be performed using a sequential covering algorithm. Inference, using the PROGOL ILP system, of temporal rules related to agent and object interactions using a sensor input is discussed in [100, 120].

Rule induction is a “bottom-up” process that refers to the inference of a set of formal rules from a training set containing examples of specific facts [114]. Inductive inference can be considered as the inverse of deduction which refers to the process in which a logically certain conclusion is drawn from one or more general statements in relation to the facts [115].

Rule induction has received considerable interest in the literature with decision trees perhaps the most common approach. For example, Leech in [94], proposed a rule-based process control method using decision-tree induction. Samples of pellet batches (Uranium dioxide powder) are collected to determine high and low quality batches based on their generation parameters. A decision-tree algorithm is used to construct rules that are able to predict pellet quality. This resulted in the eventual increased throughput and high pellet yield as reported in [93].

Similarly, rule induction has been used in the context of making credit related decisions for loan companies. Michie [109] used inductive decision tree to predict loan decisions related to the borderline applicants. Using rule induction, the prediction accuracy increased from 50% (achieved by loan officers’ decision) to 70%.

In addition to the applications above, rule induction mechanisms are also employed for diagnosis of mechanical devices in [54], where it was established that a learned knowledge base was more accurate than the hand-crafted one. Automatic classification of celestial objects [45], preventing breakdowns in electrical transformers [140], quality monitoring of rolling emulsions [73], and improving separation of gas from oil [56] are some of the other applications of rule induction reported in the survey of Langley *et al.* [93].

### 2.3.2 Stochastic Rule Induction

Hidden Markov Models [134], as introduced earlier, are often used to represent stochastic processes and are capable of modeling temporal sequences of data. However, some domains including sports games in general are inherently hierarchical in nature, containing low-level audio-visual representations with progressively higher levels of contextual interpretations e.g., a top-down view of tennis rules may look like:

$$(set_1, set_2, set_3, \dots) > (game_1, game_2, game_3, \dots) > (serve, hit, bounce, \dots) \text{ etc.}$$

If such hierarchical data is to be modeled stochastically, a hierarchical framework like *hierarchical* Hidden Markov Models (hHMMs)[49] is required to model game transitions.

In order to design an autonomous rule induction system, the classical HMM framework is not directly applicable as such it generally requires the number of states to be fixed a priori. For this purpose, a hierarchical Dirichlet Process (HDP) may be employed to provide a prior distribution over countably infinite state spaces for HMM generalization (introduced by Teh *et al.* in [168]). This results in a non-parametric Bayesian implementation of the HMM with applications in e.g. visual scene recognition [82], and the modeling of genetic recombination [189] etc.

However, in the context of this thesis, a constrained variant based on the classical Chinese restaurant process (CRP) [4] and Stick-Breaking construction [153] is proposed for stochastic induction of game rules for various environments like sports with limited rule depth (i.e. tennis and badminton etc.). This is achieved by systematically param-

eterizing hierarchical HMMs to build a rule model that describes the observed game. Following are the analogical definitions of the two aforementioned mechanisms:

**Chinese Restaurant Process:** CRP was first introduced in 1985 by Aldous [4]. In this process, customers,  $1, 2, \dots$ , enter an empty restaurant with an unlimited number of tables with unlimited capacity. The first customer sits at the first available table. A new customer is then seated either with the previous customer or is seated at the new, unoccupied table. The concentration parameter determines how likely a customer is to sit at a new unoccupied table. A variant of CRP is employed in [16] that describes a distribution on hierarchical partitions and is applied to the problem of learning topic hierarchies.

**Stick-Breaking Construction:** SB construction was introduced by Sethuraman [153]. In this a stick of a unit length is considered. It is then broken at a certain point and a value, say  $\pi_1$ , is assigned to the stick that is just broken off. The process of breaking the stick is continued unlimited number of times to obtain  $\pi_1, \pi_2, \pi_3, \dots$  etc. A tree-structured stick-breaking process is presented in [2] where, in addition to topic modeling of text data, hierarchical clustering of images is also performed.

CRP and SB-construction are both formally defined in Chapter 5, where both of these methodologies are employed to form a new constrained variant called the Multi-Level Chinese Takeaway Process (MLCTP) [77] that is suited to an environment where rules can be established in a finite fashion.

## 2.4 Conclusions

In this chapter a literature survey of some of the key concepts and techniques related to this thesis are reviewed. We highlighted some of the annotation systems present in the literature. We also introduced concepts related to anomaly detection using various sub-categories to convey a clear and concise understanding. Anomaly detection in a multi-level knowledge representation framework was also reviewed while one such system in the context of this thesis is also introduced in Chapter 3. We also presented an overview



---

of anomaly adaptation methodologies (domain specific anomaly adaptation techniques are proposed in Chapter 4 of this thesis). We also reviewed one of the most important learning paradigms of machine learning i.e. rule induction and its two sub-categories, logical rule induction and stochastic rule induction. Four generic and stochastic rule induction methodologies are proposed in Chapter 5 of this thesis.

# Computer Vision Systems for Deriving Experimental Datasets

## 3.1 Introduction

In this chapter, we briefly introduce all the sources of datasets that we use for our experiments related to anomaly detection, rule adaptation and rule induction. These datasets are extracted from either computer vision-based systems or other similar domains. We start with the tennis video annotation system of [89], for which most of our novel algorithms are primarily designed. In order to benchmark the novel methodologies presented in this thesis, we also introduce a purpose-built ground truth annotation system capable of annotating sports videos like tennis and badminton via labeling *key events* with information like *Serve* and *Hit* etc. Similarly, we also introduce another manual annotation system capable of labeling human driving intentions using a camera-equipped car driven across a city. Additionally, we also introduce two datasets from the UCI repository including website browsing behavior data and human activity localization data.

### 3.2 The Tennis Video Annotation System

Figure 3.1 shows a detailed block diagram for the Tennis video annotation system of [89]. A simplified version of the system diagram is shown in Figure 3.2. Each block in it consists of several “modules” of the system providing specific functionality (which can be found in Figure 3.1, where each rectangular block represents one module of the system). It should also be noted that the novel contribution of this thesis to the system can be summarised in the “anomalyDetection”, “anomalyAnalysis” and “highLevel” blocks of Figure 3.2 for which we introduce novel methods to make the system capable of annotating other similar sports as well. The development and implementation of the algorithms in other blocks, as well as a memory architecture that enables the modules to communicate with each other, and a graphical interface for the system, are parts of a pre-existing work carried out in this context (see [83, 89, 122, 144, 192] for more details).

**Pre-Processing** Tennis videos from various sources used in our experiments are generally recorded with interlaced cameras, thus in the “pre-processing” block of Figure 3.2, image frames are first de-interlaced into fields. Fields are used, rather than frames, in order to alleviate the effects of temporal aliasing. This is particularly important for the ball tracker. When the tennis ball is moving fast, the ball is alternately present and absent on successive frame lines, hence the need to operate on fields rather than frames. For simplicity, we will use, the word “frame” to refer to “field” in the rest of this thesis.

After de-interlacing, the geometric distortion of camera lens is corrected. The camera position on the court is assumed to be fixed, and the global transformation between frames is assumed to be a homography [60]. As discussed in Chapter 2, the homography is found by: tracking corners through the sequence; applying RANSAC to the corners to find a robust estimate of the homography, and finally applying a Levenberg-Marquardt optimiser [103] to improve the homography.

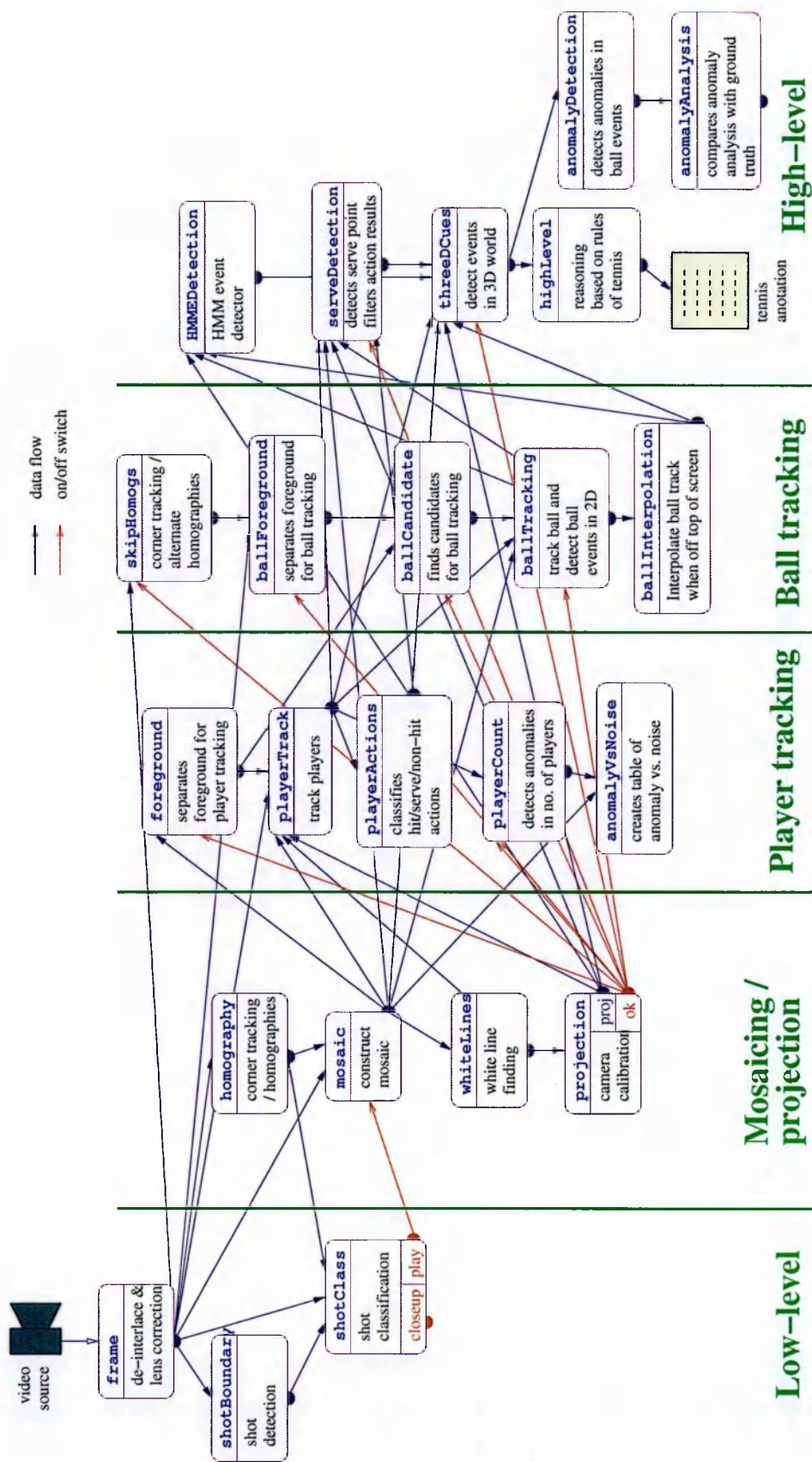


Figure 3.1: A detailed diagram of the tennis video analysis system

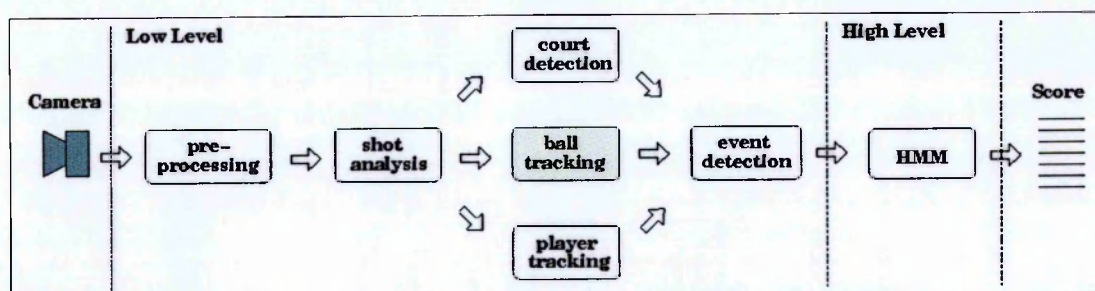


Figure 3.2: A simplified diagram of the tennis video analysis system.

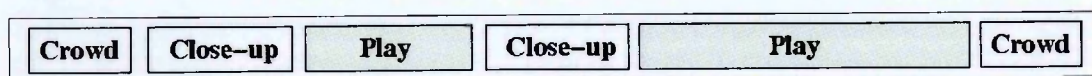
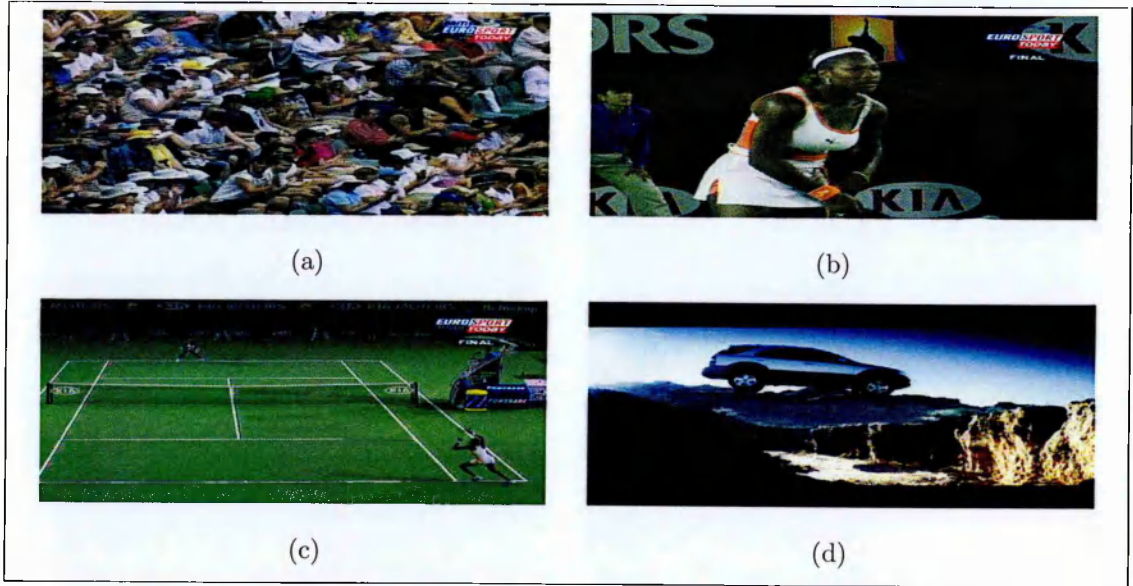


Figure 3.3: An illustrative example of the composition of a tennis video. The length of each shot is proportional to the width of the corresponding block in the figure.

**Shot Analysis** A broadcast tennis video is composed of shots, such as play, close-up, crowd, commercial. An illustrative example of the composition of a tennis video is shown in Figure 3.3. Example frames from different types of shots can be found in Figure 3.4. In the “shot analysis” block of Figure 3.2, shot boundaries are detected using colour histogram intersection between adjacent frames; shots are then classified into appropriate types using a combination of color histogram mode and corner point continuity. For our purposes, some shots are incorrectly classified as “play”. This situation arises when “replays” are encountered. However, these detected false positives are eventually eliminated later on by the “projection” module (see Figure 3.1), which rejects the shot if it is unable to find the correct tennis court.

**Court Detection, Ball Tracking, and Player Tracking** For a play shot, the tennis court is detected through a combination of an edge detector and Hough transform (as explained in Chapter 2). The players are tracked using a particle filter, and player actions are detected (see [122, 144] for details). A more complete description of ball tracking [192] and player tracking (including player action recognition [83]) techniques involved in this system is presented in Section 4.3.2 where the output of these modules are used for anomaly detection and rule adaptation.





**Figure 3.4:** Shot types. (a) crowd. (b) close-up. (c) play. (d) commercial.

**Event Detection** By examining the tennis ball trajectories, motion discontinuity points are detected. These points are combined with player positions, player actions and court lines in the “event detection” module, to generate key events such as hit, bounce and net.

**High-Level Reasoning with HMM** Finally, the generated key events are sent to a high level module, where the tennis rules are incorporated into a Hidden Markov Model (HMM). The HMM is used as a reasoning tool to generate the annotation, i.e. outcome of play, point awarded, etc. (see [89] for details). HMMs are also employed for providing contextual prior for event detection.

It is this module that we propose to replace with a generic model able to – ultimately – learn rules of any input game. Figure 3.6 shows the non-hierarchical tennis rule model used by Kolonias *et. al.* in [86, 88] to determine game scores. Our aim is to autonomously learn (instead of pre-defining) such rule models in a hierarchical fashion that is also applicable to other domains in addition to tennis.

As can be seen in the detailed system diagram of Figure 3.1, the system is composed of 23 modules. A memory architecture is implemented to enable the modules to communicate with each other [89]. This system can carry out contextual reasoning at various

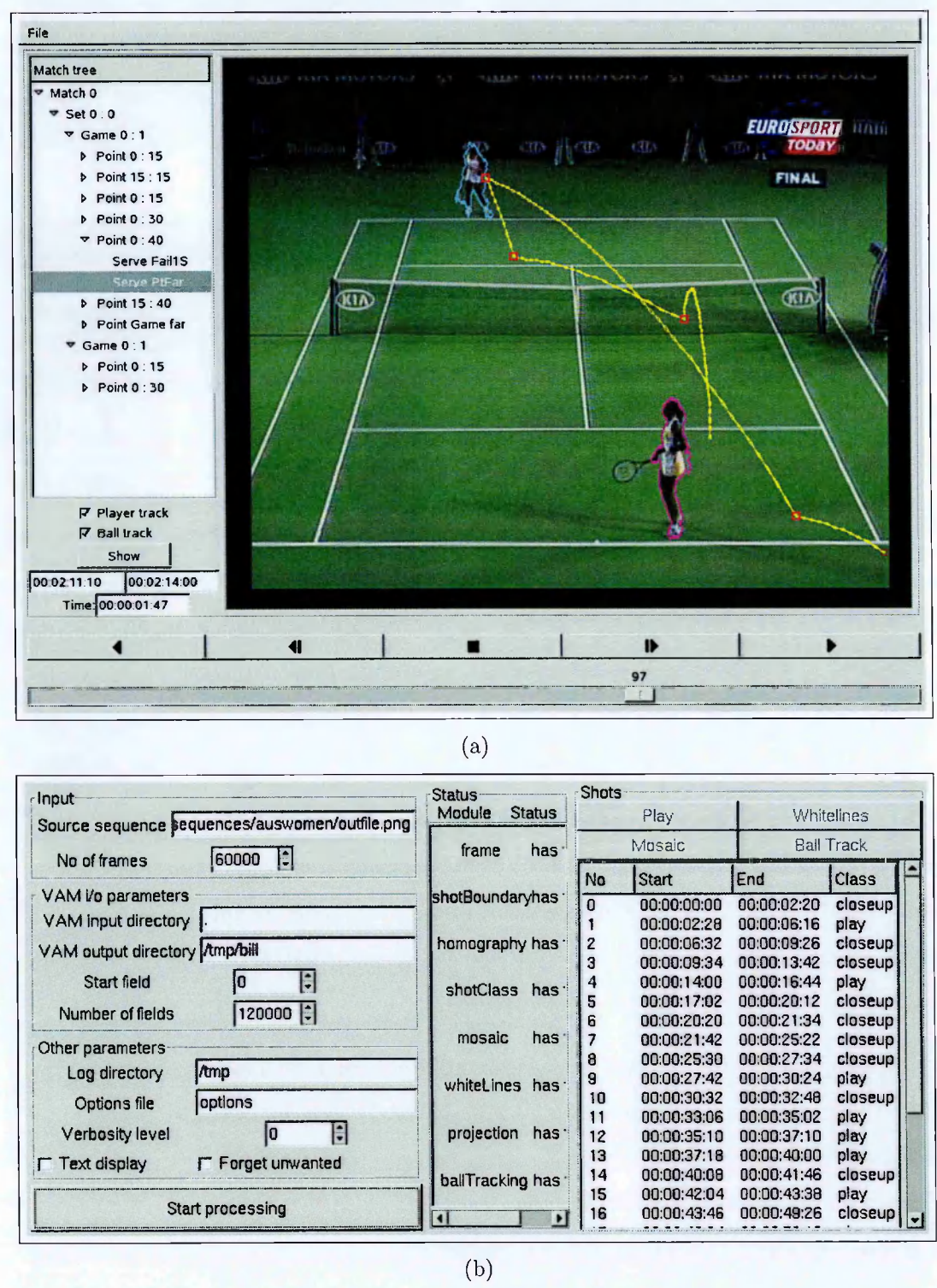
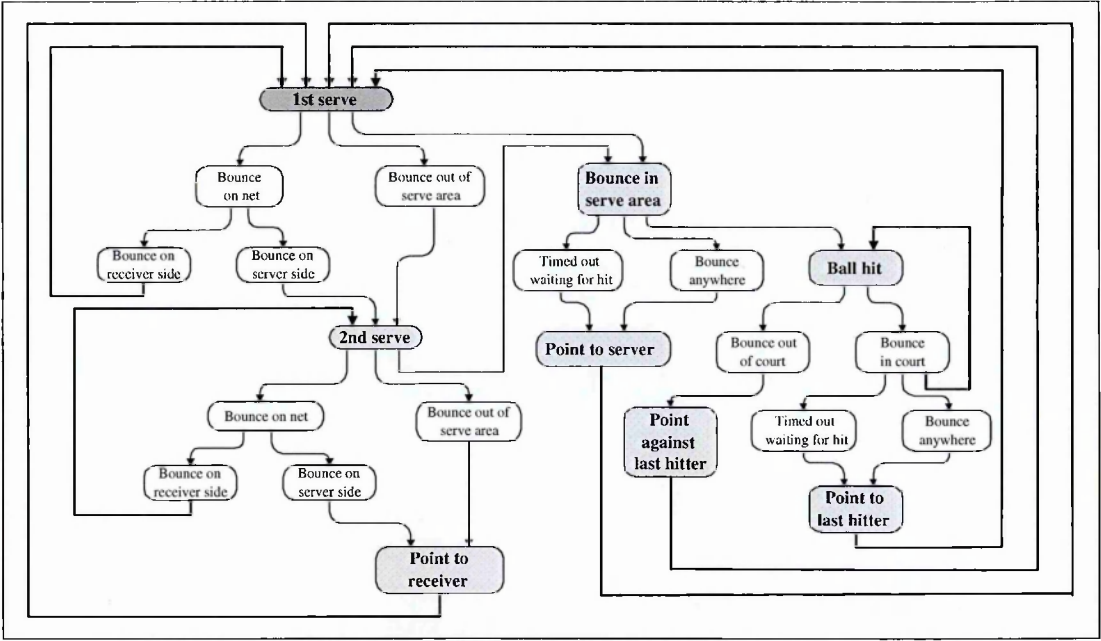


Figure 3.5: The GUI of the tennis video annotation system.





**Figure 3.6:** True tennis rule model as defined in [86, 88].

levels of interpretation for court games like Tennis with a unified apparatus with the raw video data at the lowest level and its semantic annotation of increasing abstraction at higher levels.

A graphical interface is also implemented for this system (see Figure 3.5) that shows the output of various modules in this system providing match related information such as high-level annotation e.g. game scores (displayed in the left panel), player positions, ball trajectories, and shot related information (such as showing whether the visible shot is a close-up, advert or play) etc. Also, temporally, as the game progresses, detected low-level features and the related key events are also displayed in the main window.

Table 3.1 shows all of the output labels from this system.

Moreover, an additional motivation of the work presented in this thesis is to provide a *generalized* high-level module for this system. The aim of this module is to enable the automated sports video annotation system to accommodate novel sports with rule adaptation and rule learning capabilities via anomaly detection.



**Table 3.1:** Summary of tennis events from [89]

| Event | Description   |
|-------|---|
| SFR   | Serve by Far player, Right side                     |
| SFL   | Serve by Far player, Left side                      |
| SNR   | Serve by Near player, Right side                    |
| SNL   | Serve by Near player, Left side                     |
| BIF   | Bounce Inside Far player’s half court               |
| BOF   | Bounce Outside Far player’s half court              |
| BIN   | Bounce Inside Near player’s half court              |
| BON   | Bounce Outside Near player’s half court             |
| HF    | Hit by Far player                                   |
| HN    | Hit by Near player                                  |
| BIFSR | Bounce Inside Far player’s Serve area on the Right  |
| BIFSL | Bounce Inside Far player’s Serve area on the Left   |
| BOFS  | Bounce Out of Far player’s Serve area               |
| BINSR | Bounce Inside Near player’s Serve area on the Right |
| BINSL | Bounce Inside Near player’s Serve area on the Left  |
| BONS  | Bounce Out of Near player’s Serve area              |
| NET   | Bounce on NET                                       |

### 3.3 Tennis and Badminton Ground Truth Annotation System

In order to set a baseline standard for the annotation system of Section 3.2 and to measure the accuracy levels of the novel methodologies presented in this thesis, it is important to have a tool that can be employed to generate error-free labellings (i.e. ground truth annotations). For this purpose, we build a system, capable of frame-wise and/or event-wise (i.e. every few frames when a “key event” takes place) annotating video frames with spatial information.

With this tool, a particular frame can be annotated with the following meta-data:

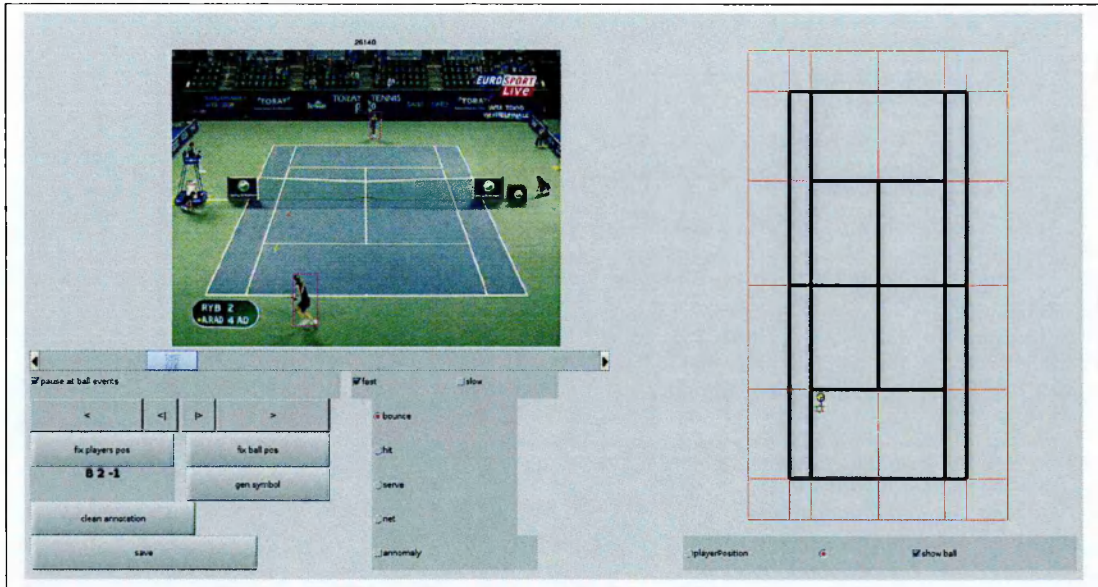


Figure 3.7: Tennis and Badminton Annotation Toolbox

- Ball and players locations in the image plane i.e. the input image co-ordinates.
- Ball and players locations in the court co-ordinates i.e. the top view of the tennis or badminton court.
- Event descriptors such as *Serve*, *Hit*, *Bounce* and *Net*.
- Using all of the above information, a *high-level* symbol is generated identifying the type and location of an event (see Tables 3.1 and 3.2)

In addition to the above features, this annotation tool is also capable of reading the output of the system introduced in Section 3.2. Thus, instead of newly annotating every single frame for player and ball locations, we can simply adjust any errors made by the system (or newly annotate only the missed frames), resulting in a speedy groundtruthing process.

We use this tool to annotate, not only tennis videos but also badminton videos where the play structure is similar. Note, these annotations are only event descriptors and do not contain information related to point allocations.

Table 3.2: Summary of Badminton events

| Event | Description                             |
|-------|---|
| SF    | Serve by Far player                     |
| SN    | Serve by Near player                    |
| BIF   | Bounce Inside Far player's half court   |
| BOF   | Bounce Outside Far player's half court  |
| BIN   | Bounce Inside Near player's half court  |
| BON   | Bounce Outside Near player's half court |
| HF    | Hit by Far player                       |
| HN    | Hit by Near player                      |

### 3.4 Driving Intention Manual Annotation System

In this section we present a similar system to the Tennis and Badminton Ground Truth Annotation System of Section 3.3 where human driving intentions are manually identified and are then labeled with spatial descriptors (i.e. annotations). We use the EU Project DIPLECS' dataset (also mentioned in [154]), where a camera-equipped car is driven across the city, to annotate driving events such as *Start*, *Turn*, *Signal* etc. We use this dataset as an additional domain in which we test the generality of the novel rule induction methodologies presented in Chapter 5.

A complete list of annotation labels (i.e. event descriptors) using this system is shown in Table 3.3.

### 3.5 Other Datasets

In addition to the data generated by the aforementioned annotation systems, we also employ two more datasets from the UCI repository, namely the website (MSNBC.com) dataset ([21]) and the human activity localization dataset ([72]). Both of these datasets are sequential with various labelings attached that describe page visits related to MSNBC.com such as *news*, *sports* and *weather* etc. for the website dataset and hu-

Table 3.3: Summary of Driving Events used in [154]

| Event | Description   |
|-------|---------------|
| LA    | LIGHTS Amber  |
| LG    | LIGHTS Green  |
| LR    | LIGHTS Red    |
| S     | Start         |
| SiLL  | Signal Left   |
| SiRR  | Signal Right  |
| Sp    | Stop          |
| TLe   | Turn Left     |
| TRi   | Turn Right    |
| TSt   | Turn Straight |

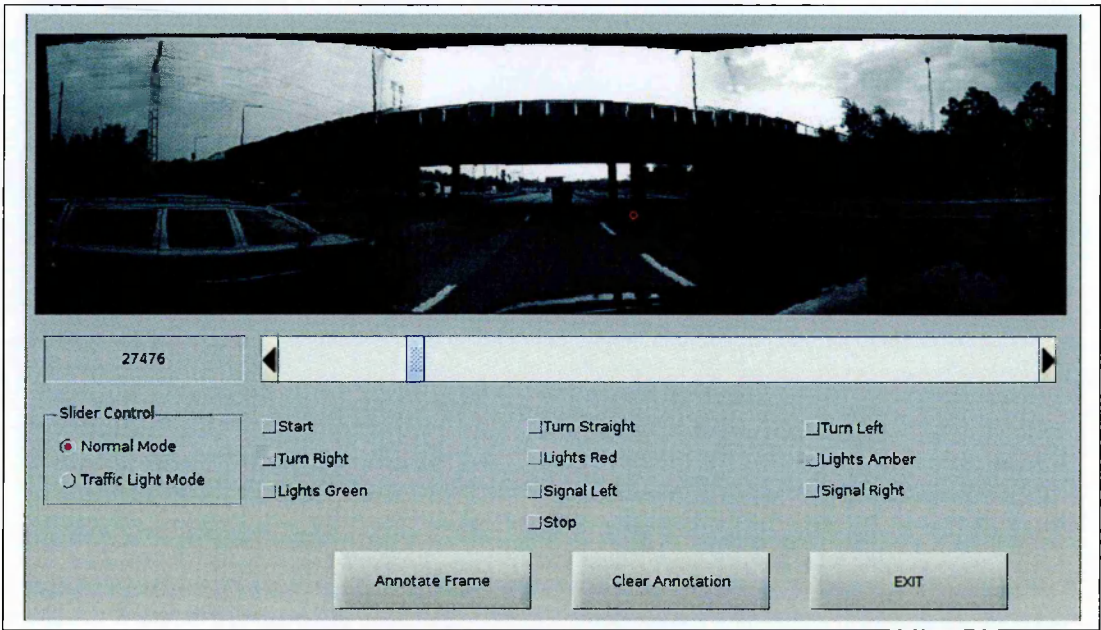


Figure 3.8: Driver’s Annotation System

man actions such as *walking*, *falling* and *sitting* etc. for the human activity localization dataset. These datasets are employed to test the generic nature of the novel rule induction methodologies presented in Chapter 5.

Lists of labels for both of these two datasets are shown in Tables 3.4 and 3.5.

**Table 3.4:** Summary of Website Events used in [21]

| Event | Description          |
|-------|----------------------|
| MH    | MSN-Home             |
| MNA   | MSN-News-ALL         |
| MNL   | MSN-News-Local       |
| MNW   | MSN-News-Weather     |
| MNB   | MSN-News-Business    |
| MNS   | MSN-News-Sports      |
| MIH   | MSN-Interests-Health |
| MIL   | MSN-Interests-Living |
| MIT   | MSN-Interests-Tech   |
| MITr  | MSN-Interests-Travel |
| MO    | MSN-Others           |

**Table 3.5:** Summary of Website Events used in [72]

| Event | Description                            |
|-------|--|
| F     | Falling                                |
| L     | Lying                                  |
| LD    | Lying Down                             |
| OF    | On all Fours                           |
| S     | Sitting                                |
| SD    | Sitting Down                           |
| SG    | Sitting on the Ground                  |
| StL   | Standing up from Lying                 |
| StS   | Standing up from Sitting               |
| StSG  | Standing up from Sitting on the Ground |
| W     | Walking                                |

---

## 3.6 Summary

In this chapter, we very briefly introduced the automated tennis video annotation system of [89] which outputs a set of tennis event labels. We also introduced a purpose-built tennis and badminton groundtruthing annotation system that exports similar set of event labels describing the sport game. In addition to this, we also introduce a manual human driving intention annotation system that outputs a set of driving events for a camera-equipped car driven across a city.

Furthermore, we finally introduced two other datasets from the UCI repository, that we employ for our experiments, describing human website browsing behavior and human action localization.

We employ these datasets for our experimental work to demonstrate the performance of the newly developed methodologies presented in Chapters 4 (Anomaly Detection and Rectification) and 5 (Rule Induction).

# Chapter 4

## Anomaly Detection and Rectification Methodologies for Knowledge Transfer in the Context of Automated Sports Video Annotation

### 4.1 Introduction

Adaptive and autonomous sports video annotation systems require the ability to switch between relevant knowledge bases depending on the input domain. When a new, but related, domain is introduced to the system, various changes are detected that are generally inexplicable in terms of the existing rule-base. We call these anomalies as defined in Chapter 2, where it is established that it is crucial to distinguish anomalies from mere errors in the input signal.

In this chapter, we present a series of methodologies to tackle the problem of anomaly detection in a sports environment i.e. detecting anomalous situations in the context of an adaptive system so as to be able to annotate different games. We specifically attempt to do this in a court-game environment (i.e. Tennis singles and doubles). In this category of sports, players follow certain rules of the game in a specified court

location providing not only feature level information but also contextual information. These aspects make a tennis game an important test bed for these methodologies.

Additionally, the associated problem of anomaly rectification is also addressed in this chapter. This involves re-structuring of the rule-base in such a way so as to accommodate the detected anomalies effectively, enabling recognition of the new domain (as discussed in Section 2.2.2 where we presented various approaches to dealing with meaningful novel events) and eventually autonomous annotation. Solutions to these problems are not only crucial in developing an autonomous video annotator but also for knowledge transfer i.e. the capability of knowing the amount of information to be shared between two different domains.

For this purpose, we first introduce a lattice-based method to address the problem of anomaly detection and rectification that exploits the implicit court structure of sports games like tennis. This is achieved via measuring court box activities for two different but related modalities (singles and doubles tennis).

The related problem of anomaly *rectification* is also tackled using this methodology (enabling adaptation of the existing learning mechanism to the change of domain). Thus, as a concrete instantiation of this notion, we investigate a novel court structure-based HMM (Hidden Markov Model) induction strategy for arbitrary court-game like environments. We show test results in real and simulated domains to demonstrate the ability of the method to identify a change in the rule base going from tennis singles to tennis doubles.

We also present another methodology for anomaly detection that is based on the disparity between the low-level vision based classifiers and the high-level contextual classifiers [184]. We then propose another approach to address the problem of anomaly rectification via the Convex hulling of localized anomalous states. We show experimental results using datasets extracted from the vision based annotation system and ground-truth tennis annotator introduced in Chapter 3.

In the next section, we thus introduce the methodological details of the lattice-based anomaly detection and rectification method and demonstrate relevant experimental results. We then outline the classifier disparity based anomaly detection method as



---

well as the convex hulling of detected anomalies approach for anomaly rectification in Section 4.3. A discussion and summary of this chapter is presented in Section 4.4.

## 4.2 Lattice Based Anomaly Detection and Rectification

Anomaly detection has received considerable interest in the literature (Chapter 2) as a means of determining whether a learning domain has changed in a fundamental way. Furthermore, this also may require continuous adaptive learning to be abandoned and a new learning process initiated in the new domain i.e. switching the knowledge base. Thus, in this section, we introduce a lattice-based method for anomaly detection specifically in the context of court game environments such as Tennis. We also address the related problem of anomaly *rectification* using this methodology; the adaptation of the existing learning mechanism to the change of domain. As a concrete instantiation of this notion, an HMM (Hidden Markov Model) based induction strategy is investigated in this context. We test (in real and simulated domains) the ability of the method to adapt to a change of rule structures going from tennis singles to tennis doubles. In the following section, we introduce the method following its methodological formulation and experimental results.

### 4.2.1 Introduction

There is a well-established requirement for detecting and treating anomalies in machine learning for creating an adaptive system. Artificial cognitive systems, in particular, should be able to autonomously extend capabilities to accommodate anomalous input as a matter of course (humans are known to be able to establish novel categories from single instances [173]). Typically, the anomaly detection problem is one of distinguishing novel (but meaningful) input from misclassification error within existing models i.e. by defining a new learning domain. By extension, the *treatment* of anomalies so determined typically involves the attribution of suitable class designators to the novel input, along with an appropriate method for extending (i.e. generalizing) this categorization. The composite system should thus be capable of inferring novel representations

— ‘bootstrapping’ — via the interaction between the bottom-up processes of anomaly detection and the top-down processes of novel object categorization. Such composite techniques have been applied, for example, to the problem of segmentation [95]. Often, bottom-up description will also explicitly consider *context*, rather than specific objects of classification interest as means of generating high-level domain description [98, 126]. For evaluation and the proof of concept, we consider anomaly rectification in the context of sporting events, focusing on Markovian modeling of anomalous high-level (i.e. abstract, rule-like) state transitions such that the inference system must detect *how* the rules of game-play should change. As a test-bed for this idea, we start with a system trained on ‘singles’ tennis matches, and then change the input material for doubles tennis matches. On the assumption that a suitable detection system has already flagged the game-play anomaly and collected suitable quantities of data in the newly defined domain, the problem then is to adapt the existing rule structure accordingly. We define our approach in terms of observed state transition probabilities defined in the two different rule domains, initially testing the method on simulated state transition data and later on testing on real data derived from an existing system that employs court line detection, homography, player/serve detection, and ball detection via tracklet propagation for singles tennis annotation [81, 192].

In the following section we present the problem formulation, with the methodology described in Section 4.2.3. The implementation protocol and experimental validation on real and simulated data are discussed in Section 4.2.4 with methodological conclusion offered in Section 4.2.5.

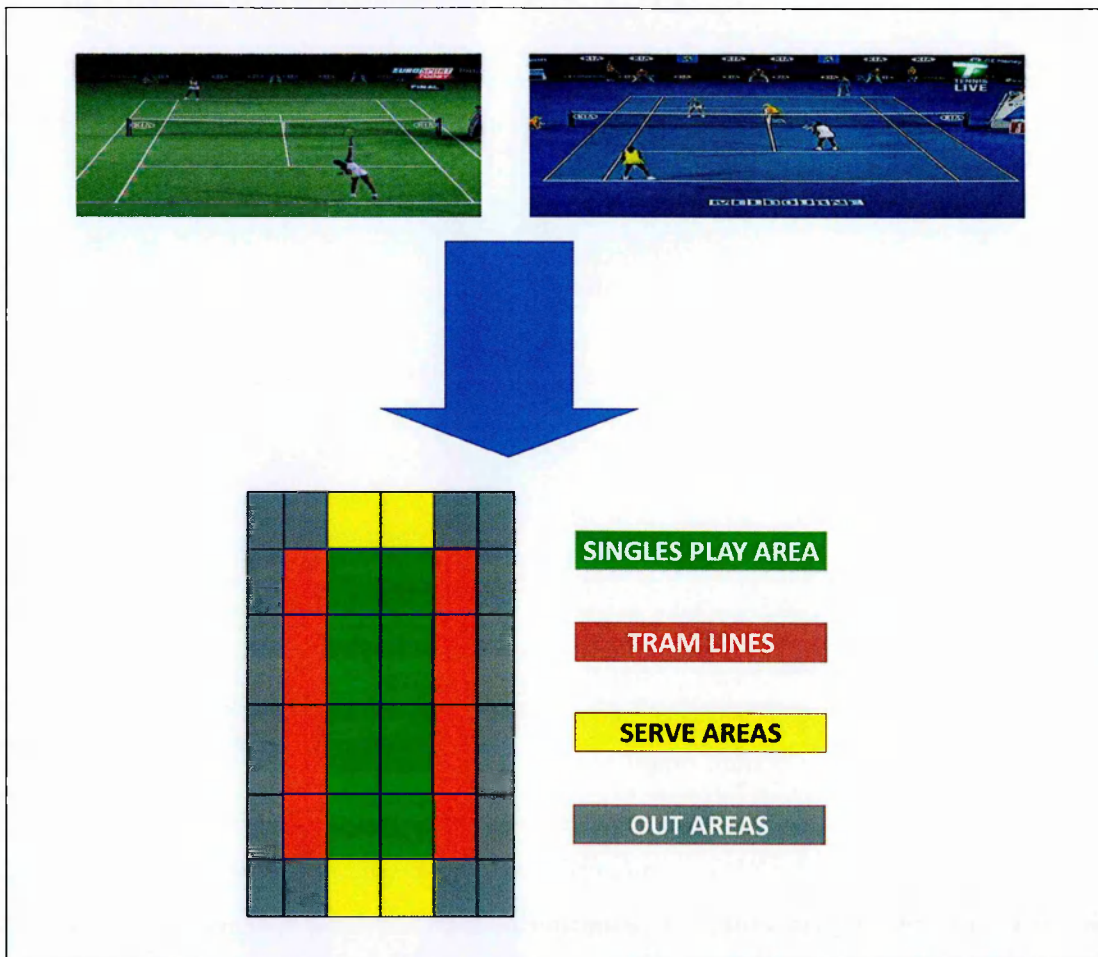
#### 4.2.2 Anomaly Detection and Rectification in Court Game Environments

The tennis annotation system described in [81, 192] and Section 3.2 does not identify individual players, so that scoring is primarily determined via ball movements with respect to designated play areas. We employ a simplified version of Table 3.1 introduced in Chapter 3 as follows:

- Play Area (PA),

- Near Play Area (NPA),
- Far Play Area (FPA),
- Near Serve Area (NSA),
- Far Serve area (FSA),
- Ball Out Area on Far side of the court (BOF), and
- Ball Out Area on Near side of the court (BON).

Figure 4.1 highlights the singles tennis' playing area and also the tram lines (included in the playing area for doubles tennis).



**Figure 4.1:** Highlighted playing areas for singles and doubles tennis including the out areas

Each of these areas is associated with a 4-tuple box designation,  $b$ , given in terms of the ordered set of horizontal,

$$H = \{(h_1, h_2, \dots, h_{n_h})\} \quad (4.1)$$

and vertical screen lines,

$$V = \{(v_1, v_2, \dots, v_{n_v})\} \quad (4.2)$$

Thus,

$$b \in \{(h_\alpha, v_\beta, h_\zeta, v_\omega)\} \quad (4.3)$$

with  $h_\alpha, h_\zeta \in H$  and  $v_\beta, v_\omega \in V$ .

Applying the constraints,

$$v_\beta < v_\omega \text{ and } h_\alpha < h_\zeta, \quad (4.4)$$

each box has a unique  $b$  designated in terms of its bottom left and top right corner coordinates; i.e.

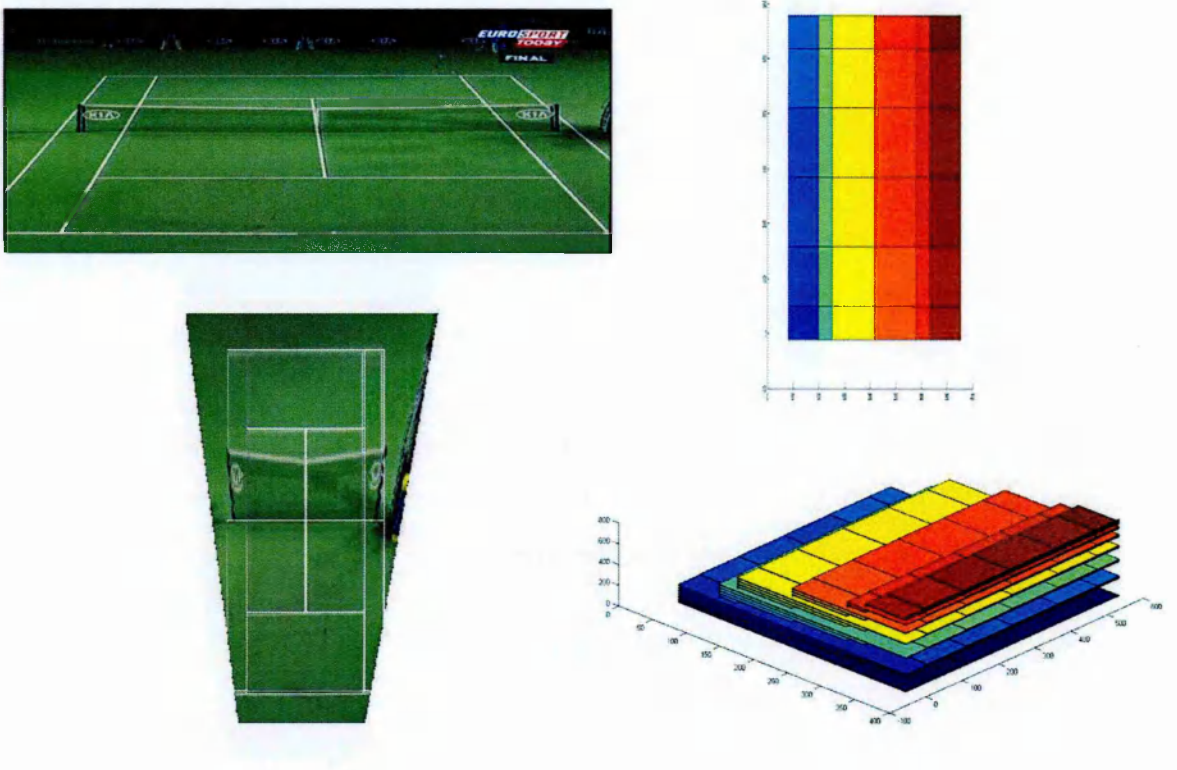
$$(h_\alpha, v_\beta) \text{ and } (h_\zeta, v_\omega) \quad (4.5)$$

The complete set of boxes  $\{b\}$  forms a *lattice* (see Figure 4.2), having *joins* (also known as least upper bounds or suprema) and *meets* (also known as greatest lower bounds or infima) analogous to intersection, union and complementation etc in set theory<sup>1</sup>. This allows for complex relationships between designated play areas, e.g. overlaps and subset relations (such as PA/FPA=NPA i.e. near play area is the difference between the total play area and the far play area). This notion of a lattice clearly generalizes to any

---

<sup>1</sup>A set equipped with a partial order relation for all elements is automatically a lattice if this relation is reflexive, antisymmetric and transitive.

other rectilinear court structures such as those of badminton (and indeed our method as a whole is intended to generalize to any such domain, so that learning-transfer is possible between superficially different game types).



**Figure 4.2:** Tennis court lattice with a Tennis court mosaic and projection images on the left and two views (2-D and 3-D) of the constructed court lattice showing various levels of court box sizes.

From this perspective, the distinction between singles and doubles tennis is characterized by a change in definition of the play area (PA);

$$(PA \rightarrow b_o) \rightarrow (PA \rightarrow b_n) \quad (4.6)$$

with  $b_o, b_n \in \{(h_\alpha, v_\beta, h_\zeta, v_\omega)\}$ .

As a step towards a fully general sport-rule annotation induction system capable of transferring learning from one domain to another, our aim is to detect this transition and thereby identify both the old and new play area definitions i.e.  $b_o$  and  $b_n$ .

This situation is made inherently complex by the fact that ball state transitions in terms of which the high-level game description is given e.g. for a typical serve ,

$$NSA \rightarrow PA \rightarrow BO \quad (4.7)$$

are not directly observed. Instead, we see only transitions in the occupancies of the various boxes within the lattice, which looks like,

$$b_1 \rightarrow b_2 \rightarrow b_3 \quad (4.8)$$

and to which these high-level (contextual) game descriptions correspond, such that the high-level state space can be regarded as the *hidden* states of a Hidden Markov Model (HMM). Moreover (making this analogy exact), we find that the transition structure is inherently ambiguous within the observable state space because of the possibilities of inclusion and intersection within the lattice. We will thus in general have a large set of box transitions within the lattice,

$$\{(b_1 \rightarrow b_2 \rightarrow b_3), (b_5 \rightarrow b_8 \rightarrow b_3), \dots\} \quad (4.9)$$

which are consistent with any given sequence of key play areas. The task of determining which high-level play area has undergone redefinition in the transition from singles to doubles game-play requires that we obtain a method for treating this ambiguity. We do this via a Minimum Description Length (MDL)-like [141] approach in which we favor the smallest parametric change, required to bring-about the appropriate high-level re-description of the game mechanics (key areas being the main rule-designated areas of play). A single *key area* transformation is represented by,

$$(key\_area \rightarrow b_A) \rightarrow (key\_area \rightarrow b_B) \quad (4.10)$$

Note that key area transitions in an arbitrary court game can be between any boxes of *any* size, for instance a transition that goes from the serve area (SA) to the far play area (FPA) is generally a transition from an area of 1 ‘court unit’ to an area of several

court units in size (if a court unit is the smallest delineatable region defined by the court lines).

Consistent with a fully-unsupervised approach, we will initially assume no prior knowledge of the injective mapping,

$$\mathcal{P} \rightarrow \{b\} \quad (4.11)$$

where  $\mathcal{P}$  is the set of play areas,

$$\mathcal{P} = \{PA, NPA, FPA, NSA, FSA, BOF, BON\} \quad (4.12)$$

and  $b \in \{(h_\alpha, v_\beta, h_\zeta, v_\omega)\}$

(i.e. we will not assume knowledge of even the initial single play areas). However, for the purposes of experimental application, we will later relax this assumption in order to recast the approach as one of learning transfer (which can be treated as a subset of the above problem).

### 4.2.3 Methodology

We assume that game play can be modeled via an HMM in which the hidden states are the rule-designated play areas  $\mathcal{P}$  and the emission states are the least elements of the lattice  $\mathcal{P}$  (i.e. the ‘smallest’ indivisible boxes of the court) such that,

$$b \in \{(h_\alpha, v_\beta, h_{\alpha+1}, v_{\beta+1})\} \quad (4.13)$$

The game play is thus described by key points of the ball’s trajectory (serves, hits and bounces) which are described by the system of Section 3.2 as having occurred at a particular time within one of these ‘small’ (i.e. indivisible) court units. An HMM-based game-play description of this kind is sufficient to enable the existing hardwired tennis annotation system of Section 3.2 to provide accurate score annotation of singles games.

Within such a Markovian framework we consequently assume that there exists a transition probability matrix  $M_P(\mathcal{P}_{in}, \mathcal{P}_{out})$  describing the probability of transition between key play areas. In particular, this matrix is sufficient to capture the notion that a certain fraction of the serves will be returned, with the remainder resulting in either a point award (i.e.  $FPA \rightarrow BO$ ) or an ‘out ball’ (i.e.  $SA \rightarrow BO$ ). The returned balls will either go out, be awarded a point or enter into a further rally recursion, with some particular probability captured by the matrix  $M_P$ .

In addition to this matrix, game-play characterization also requires the injective mapping,

$$f(\mathcal{P}) \rightarrow \{b\} \quad (4.14)$$

that gives the actual *definitions* of the play areas ( $f$  is thus the mapping between the key-area labels and the corresponding boxes within the lattice). Consequently, the transition from singles to doubles tennis game-play may be characterized by a transition from this mapping to some other specific mapping i.e.  $f \rightarrow f'$  (i.e we assume that the basic game-play structure remains the same in terms of the key-area transition probabilities, with only the mapping into the lattice undergoing change). In our later simulation of the single to doubles transition, only a single element of this mapping (relating to the play area) will undergo change: i.e.

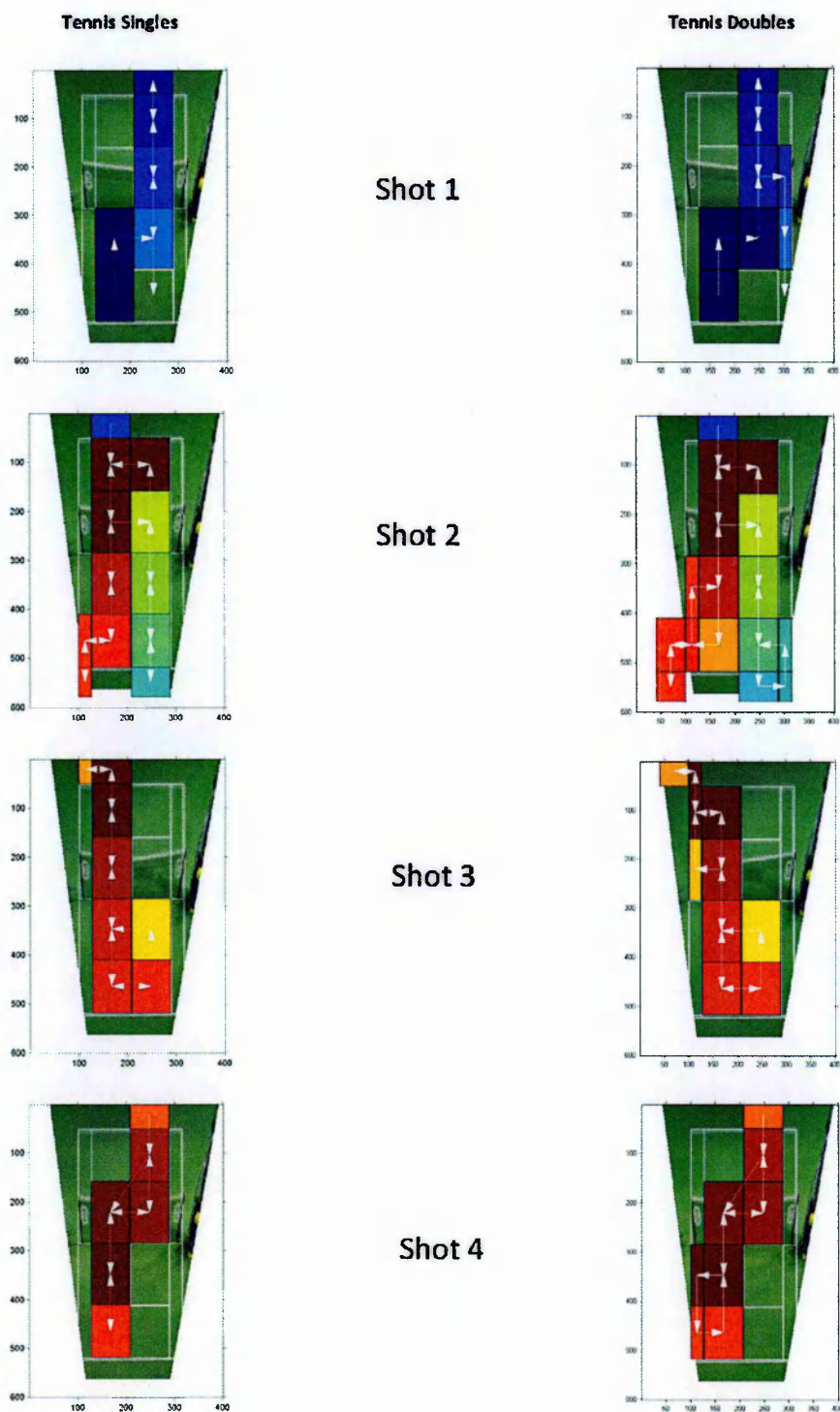
$$(PA \rightarrow b_o) \rightarrow (PA \rightarrow b_n) \quad (4.15)$$

where  $b_o$  is the *old play area* and  $b_n$  is the *new play area*, with

$$b_o, b_n \in \{(h_\alpha, v_\beta, h_\zeta, v_\omega)\} \quad (4.16)$$

However, the only evidence for this transition in the definition of play area (PA) that we are presented with is in terms of the *observed* matrix of box transitions defined over the entire lattice  $M(b_1 \in \{b\}, b_2 \in \{b\})$  (note that  $M$  is the histogram of lattice transitions for a given set of play sequences, rather than a true row-normalized transition matrix





**Figure 4.3:** Singles and Doubles Tennis court box transition at the smallest (i.e. indivisible) units

like  $M_P$ ). An example illustration of this is shown in Figure 4.3 where court box transitions are shown only at the lowest level of the court lattice.

A change to the transition matrix  $M_P$  can thus only be detected by compiling multiple observations in the differing domains (resulting in e.g. a singles transition matrix,  $M^s$ , and a doubles transition matrix,  $M^d$ ). However, even with sufficient sampling of  $M^s$ , we do not directly know which play-area box-mapping has undergone transition, since in general a large fraction of other boxes in the lattice will experience correlated activity as a result of the transition. In order to determine precisely which key area redefinition has taken place our first goal is thus to determine the matrix transform,  $T$ , parameterized by the key play-area transform,  $(b_o \rightarrow b_n)$ , that brings about  $M^d$  i.e. we require a  $T$  such that,

$$T(M^s, b_o, b_n) = M^d \quad (4.17)$$

Without further analysis, it is not clear *a priori* that this is a well-posed problem, in the sense that the transform may be non invertible if the resultant matrix,

$$M^d = T(M^s, b_o, b_n) \quad (4.18)$$

loses information about the individual lattice components  $b_o, b_n$ . In addition to this difficulty, we also have the potentially inadequate sampling of the probabilities in the underlying Markovian play-area transitions of  $M_P$  manifested in  $M^s$  and  $M^d$ . We therefore seek instead to minimize the *residual* of the parameterized transform  $T(M^s, b_A, b_B)$  with respect to  $M^d$ , rather than directly inverting it:

$$(b_o^{est}, b_n^{est}) = \underset{(b_A, b_B)}{\operatorname{argmin}} \left[ D(T(M^s, b_A, b_B), M^d) \right] \quad (4.19)$$

where  $D$  is an appropriate distance measure (see below); the superscript *est* denotes the estimated lattice value. The transform  $T$ , itself, is derived as follows:

The aggregate ‘ball-event activity’ associated with any given box  $b$  in the lattice can be separated into ‘into’,  $M(., b)$ , and ‘out-of’,  $M(b, .)$  transition components. We can also

define a coarse aggregate activity measure  $A(x)$  by summing over all of the observed transitions *into* and *out of* the box  $x$  for every single box within the lattice,

$$A(x) = \sum_{m=1}^{\mathcal{L}} (M(x, m) + M(m, x)) \quad (4.20)$$

where,

$$\mathcal{L} = \frac{h_{n_h}(h_{n_h} + 1)}{2} \cdot \frac{v_{n_v}(v_{n_v} + 1)}{2} \quad (4.21)$$

The rationale for doing so is we can thereby obtain an approximate means for estimating the effect of redefining a key area (e.g.  $(PA \rightarrow b_o) \rightarrow (PA \rightarrow b_n)$ ) by *translating* the activity associated with a box  $b_o$  to  $b_n$ : i.e. such that,

$$A^{new}(b_n) = A(b_o) \quad (4.22)$$

However, it is not simply the case that we can transfer activity in this way without also explicitly considering interactions *within* the lattice structure.

A measure of this lattice interaction can be defined in terms of the proportional overlap of one box with respect to another. The expectation of the coarse activity measure  $A$  in box  $b_1$  due to activity in box  $b_2$  for uniformly distributed ball events is thus:

$$E[A(b_1|b_2)] = \frac{|b_1 \cap b_2|}{|b_2|} \cdot A(b_2), \quad |b| = (h_\alpha - h_\zeta)(v_\beta - v_\omega) \quad (4.23)$$

This is also true for both the 'into' and 'out of' of activity components. Thus, for example, given an isolated 'into' component  $M(., b)$ , we expect a second, potentially overlapping, box  $b'$  to have an 'into' component

$$\frac{|b' \cap b|}{|b|} M(., b) \quad (4.24)$$

Consequently, to a first order of approximation, the play area redefinition  $(PA \rightarrow b_o) \rightarrow (PA \rightarrow b_n)$  has the effect on the matrix  $M$  of subtracting a '*lattice interaction*'

matrix,  $M_{sub}$ , that removes activity attributable to box  $b_o$ , while adding another lattice interaction matrix,  $M_{add}$ , that displaces this activity to box  $b_n$ . Hence:

$$M_{add}^{(o,n)}(x, y) = M(x, b_o)E[A(y|b_n)] + M(b_o, y)E[A(x|b_n)] \quad (4.25)$$

and

$$M_{sub}^{(o,n)}(x, y) = M(x, b_o)E[A(y|b_o)] + M(b_o, y)E[A(x|b_o)] \quad (4.26)$$

That is, we obtain  $M_{add}$  and  $M_{sub}$  by multiplying all ‘into’ and ‘out of’ transitions of the box in question by the expected overlap of activity. To the first order, the transform  $T$  can, thus, be approximated by:

$$T(M, b_o, b_n) = M(., .) + M_{add}^{(o,n)}(., .) - M_{sub}^{(o,n)}(., .) \quad (4.27)$$

However, this does not take into account the fact that activity in  $b_o$  and  $b_n$  have a certain likelihood of influencing each other at the outset; i.e. we cannot say that *all* of the activity in  $M$  attributable to  $b_o$  should be transferred to  $b_n$ . Moreover, we cannot say that all activity in  $b_o$  is attributable *specifically* to  $b_o$ ; it could equally apply to an intersecting box. We therefore introduce a free parameter representing the appropriate proportion of activity to transfer for inclusion within the optimization i.e. we specify:

$$T(M, b_A, b_B, \gamma) = M(., .) + \gamma(M_{add}^{(A,B)}(., .) - M_{sub}^{(A,B)}(., .)) \quad (4.28)$$

such that the optimization function becomes:

$$(b_o^{est}, b_n^{est}) = \underset{b_A, b_B}{\operatorname{argmin}} \left[ \underset{\gamma}{\operatorname{argmin}} D(T(M^s, b_A, b_B, \gamma), M^d) \right] \quad (4.29)$$

The ready optimisability of the above equation lies in the fact that the matrices  $M^s$  and  $M^d$  are essentially sparse when the effects of lattice interaction are removed from consideration, with occupancy dictated by the size of the game-play transition matrix,  $M_P(\mathcal{P}, \mathcal{P})$  (i.e.  $\mathcal{P} \times \mathcal{P}$ ), rather than the size of the lattice transition matrix,  $|b| \times$

$|b|$ . We can thus regard  $M$  as a convolution of the individual components  $(g_x^i, g_y^i)$  of  $M_P(f(\mathcal{P}), f(\mathcal{P}))$  with an activity ‘point-spread function’,  $E[A(x|g_x^i)] \cdot \delta(y - g_y^i) + E[A(y|g_y^i)] \cdot \delta(x - g_x^i)$ .

The full optimization function for the transform, using an activity-normalized RMS (root mean square) residual difference measure, is thus:

$$(b_o^{est}, b_n^{est}) = \underset{b_A, b_B}{\operatorname{argmin}} \left[ \underset{\gamma}{\operatorname{argmin}} \operatorname{RMS}((M^d - (M^s + \gamma(M_{add}^{(A,B)} - M_{sub}^{(A,B)}))) \circ M_{\text{norm}})) \right] \quad (4.30)$$

where the normalization matrix is defined by,

$$M_{\text{norm}}(a, b) = \left( \frac{|h_\alpha^a - h_\beta^a| \cdot |v_\alpha^a - v_\beta^a| \cdot |h_\alpha^b - h_\beta^b| \cdot |v_\alpha^b - v_\beta^b|}{(|h_1 - h_{n_h}| \cdot |v_1 - v_{v_h}|)^2} \right)^{-1} \quad (4.31)$$

( $\circ$  is the Hadamard product).

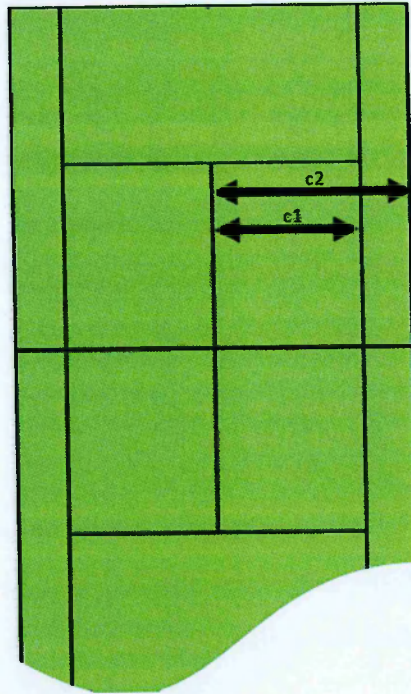
The above optimization is still based on finding a single, optimal substitution  $b_o \rightarrow b_n$ ; however, denoting this optimization  $O(p_o^{est}, p_n^{est})$ , it can be seen that a fully general optimization function for arbitrary matrix transforms,  $O_{gen}$ , can be obtained by concatenating sequences of individual box redefinitions:

$$O_{gen}(M_1, M_2) = \sum_{k=1}^m O(b_{2k-1}, b_{2k}) \quad (4.32)$$

However, in this case it is necessary to balance the allocation of parametric freedom (essentially governed by  $|m|$ ) with the cumulative RMS residuals. This requires an empirical cross-validation or *a priori* MDL-like criterion to accomplish. Such a generalization of the current approach could potentially transfer learning from tennis to badminton, since much of the serve/return game-play structure is consistent between the two, with only the court area definitions differing between them.

#### 4.2.4 Implementation Protocol and Experimental results

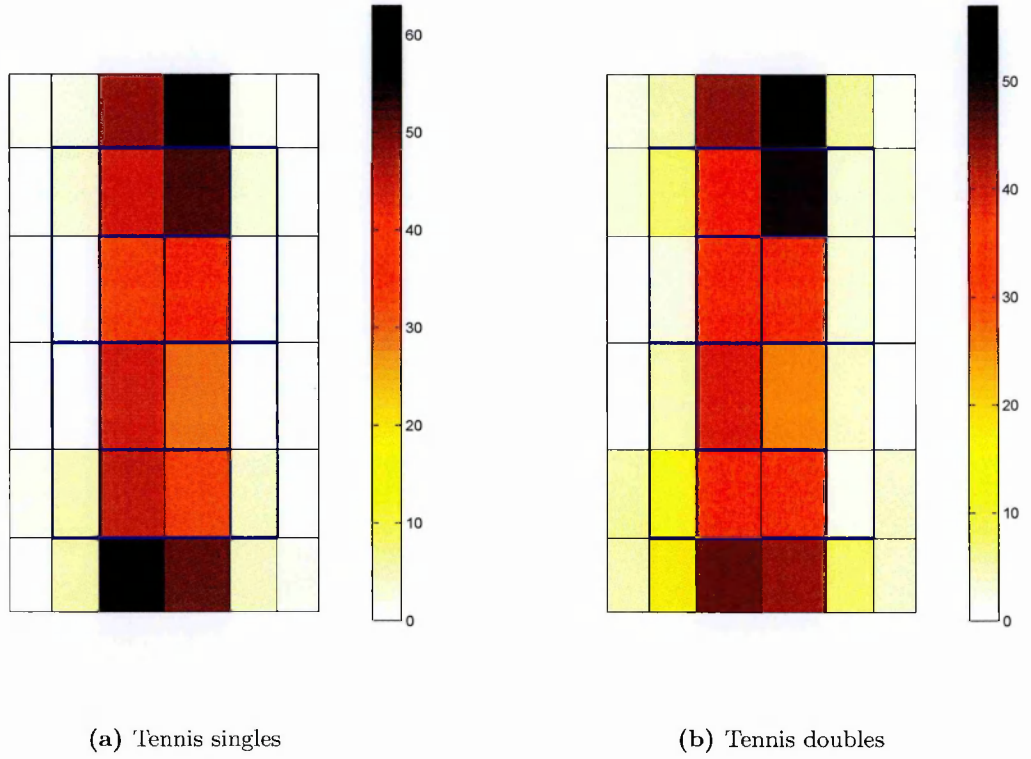
We employ simulated tennis singles and doubles games for evaluation using 100 playing shots starting with serve and ending with a BO. We also test on real data derived from the Toray Pan Pacific Open 2009 womens singles match between M. Rybarikova and A. Radwanska with a total of 58 play-shots with 58 Serves, giving 343 events in total (excluding hits) and 285 Bounces. The doubles game is simulated using the real data by expanding the playing area via multiplying all the points with the factor  $\frac{c_2}{c_1} = 1.33$  (see Figure 4.4).



**Figure 4.4:** Singles to Doubles Tennis extension factor

Comparative lattice box activities in the smallest (i.e. indivisible) units can be seen in Figure 4.5 for the original singles tennis and derived doubles tennis. Note, the box activities shown only represent the bounce type of events.

To start the experiments, we first simulate a simplified tennis game by choosing  $M_P(\mathcal{P}, \mathcal{P})$  with the following transition probabilities (we omit NPA and FPA transitions to give a single-box lattice-transformation problem):



**Figure 4.5:** Comparative lattice box activities for singles tennis (left) and derived doubles tennis (right) using the expansion factor shown in Figure 4.4

$$p(NSA \rightarrow PA) = 0.9, \quad (4.33)$$

$$p(NSA \rightarrow BOF) = 0.1, \quad (4.34)$$

$$p(PA \rightarrow PA) = 0.2 \quad (4.35)$$

$$p(PA \rightarrow BOF) = 0.7, \quad (4.36)$$

$$p(PA \rightarrow BON) = 0.1, \quad (4.37)$$



$$p(others) = 0 \quad (4.38)$$

i.e. we capture the possibility that a serve may or may not be returned; a low rally probability is also included. We also have the following ordered 4-tuple play area definitions for simulated ‘doubles’ play (omitting center lines for simplicity):

$$NSA \rightarrow (1, 1, 3, 2), \quad (4.39)$$

$$PA \rightarrow (2, 2, 5, 5) \quad (4.40)$$

$$BOF \rightarrow (1, 5, 6, 6) \quad (4.41)$$

$$BON \rightarrow (1, 1, 6, 2) \quad (4.42)$$

For 100 simulated serves this generates the lattice transition matrix depicted in Figure 4.6 .

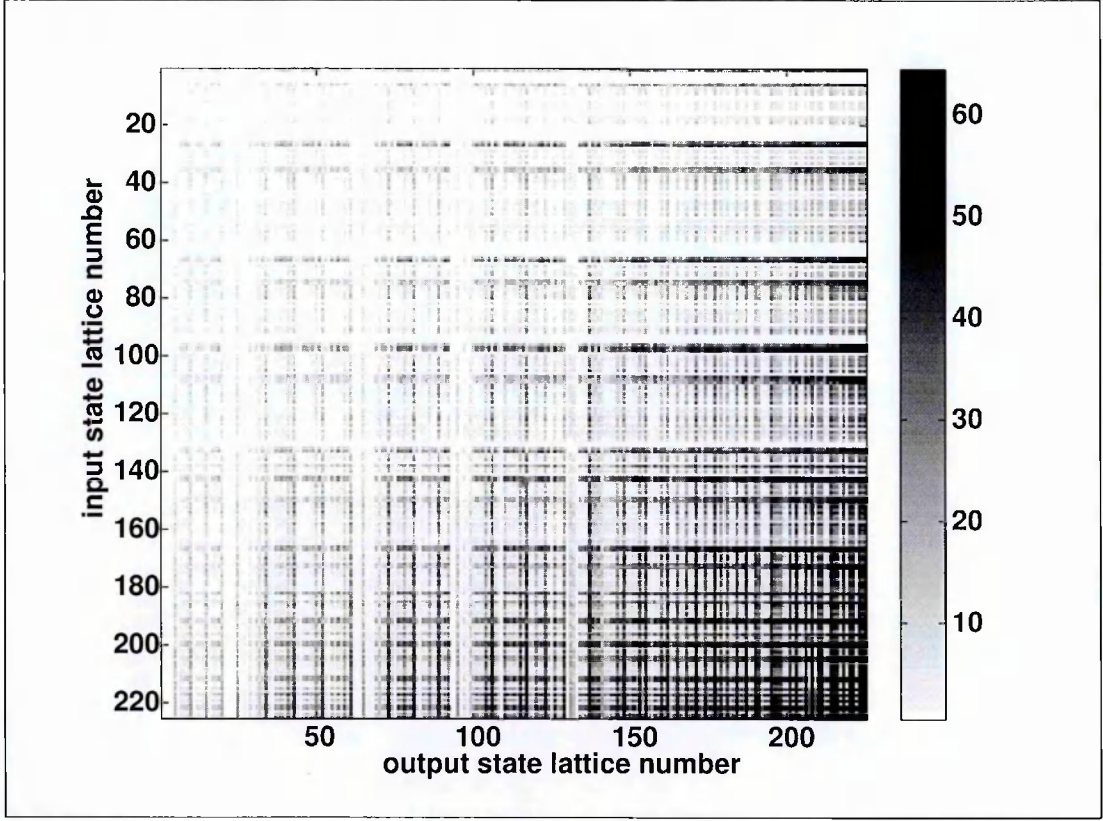
Singles play is simulated (in this simplified scenario) by changing the PA key area description to  $(PA = (3, 2, 4, 5))$  and keeping all the remaining values. This represents the fact that the ‘tram-lines’ are no longer part of the legitimate play area, so that any ball bouncing in this area is not automatically out. The resulting lattice transition matrix depicted for observations of 100 simulated serves is depicted in Figure 4.7.

Carrying out the optimization in Equation 4.30 by considering all possible transitions,

$$(PA \rightarrow b_o^{est}) \rightarrow (PA \rightarrow b_n^{est}) \quad (4.43)$$

and iterating over  $\gamma$ , we obtain an estimate of this game-play area redefinition. (Note that for the transfer learning problem, we need only consider the redefinitions of *known* play areas such that the search space is of size  $|\mathcal{P}|$  rather than  $|b|$ , i.e.  $b \in f(\mathcal{P})$  ).





**Figure 4.6:** Gray-scale histogram of Singles ( $M^s$ ) transition counts over the lattice (ordered by box size and count-number, respectively)

A general performance metric for proposed play-area redefinitions of this type can be obtained by taking the total ordinal difference between proposed and actual transitions. Thus, for a ‘ground-truth’ box redefinition:

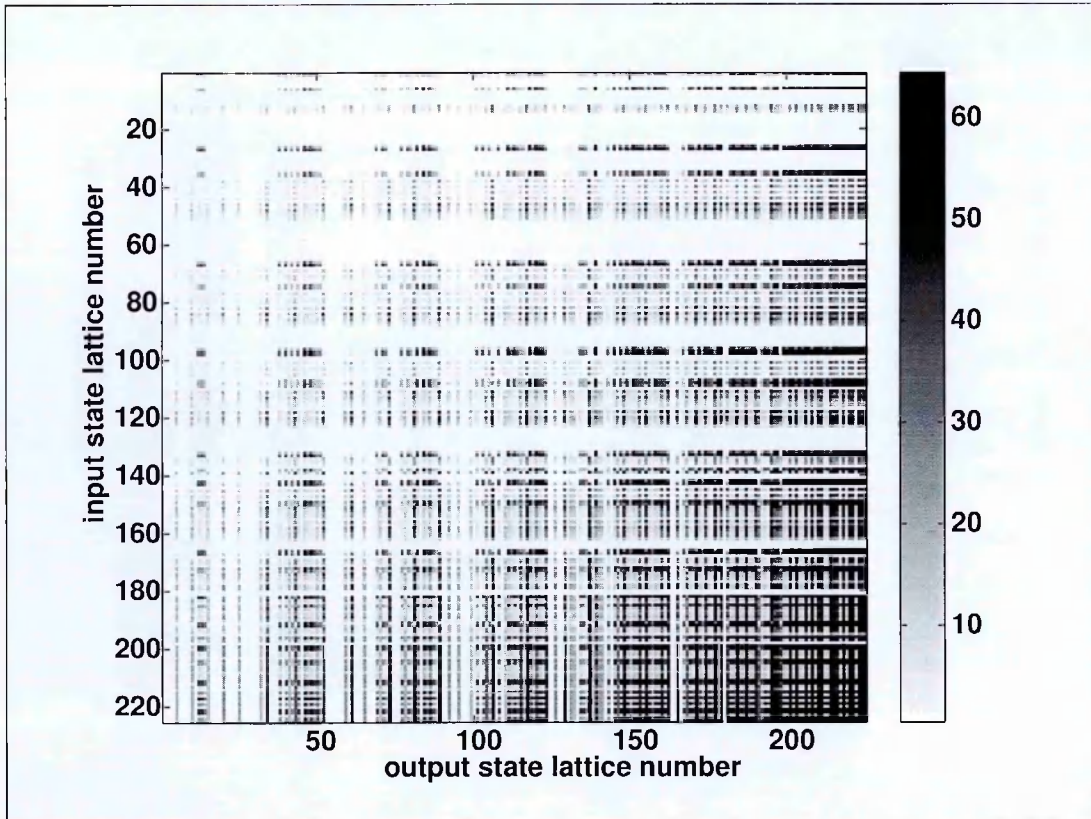
$$PA(h_1^a, v_1^a, h_2^a, v_2^a) \rightarrow PA(h_3^a, v_3^a, h_4^a, v_4^a) \quad (4.44)$$

and a proposed box redefinition supplied by the optimization method:

$$PA(h_1^p, v_1^p, h_2^p, v_2^p) \rightarrow PA(h_3^p, v_3^p, h_4^p, v_4^p) \quad (4.45)$$

We have:

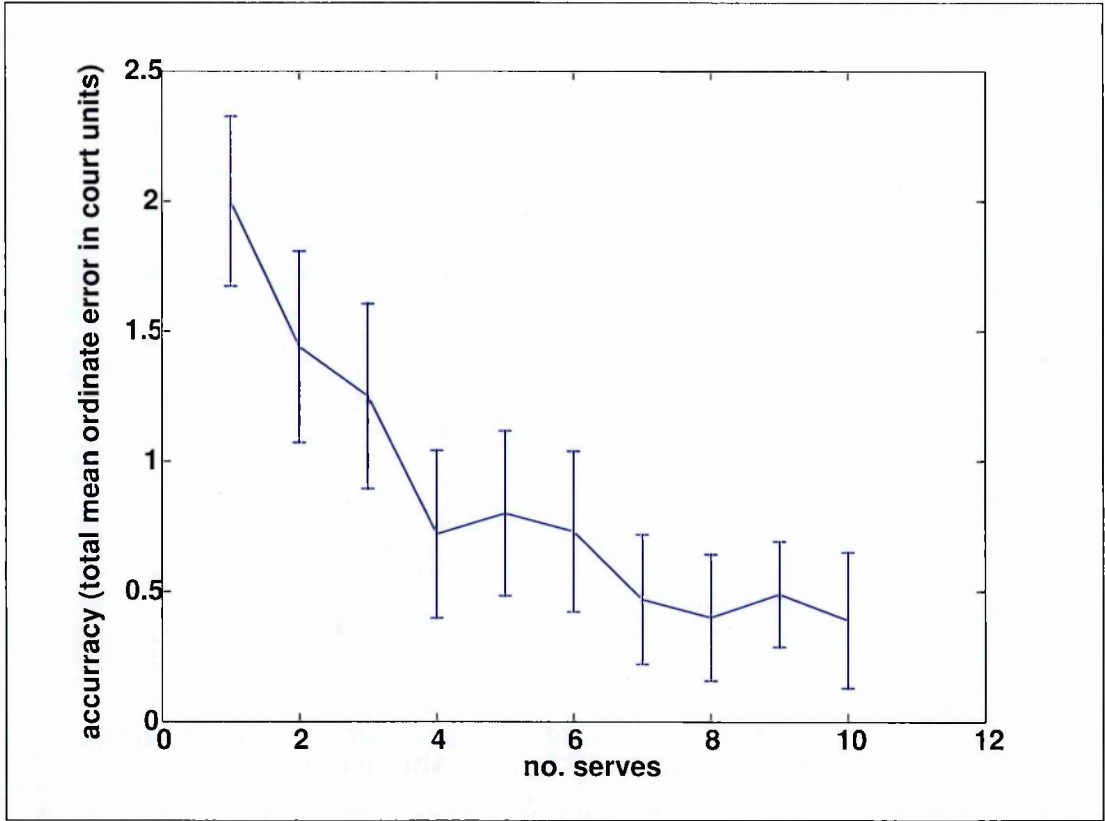
$$Error = \left( \frac{1}{\max(Error)} \right) \sum_{x=1}^4 [|h_x^a - h_x^p| + |v_x^a - v_x^p|] \quad (4.46)$$



**Figure 4.7:** Gray-scale histogram of Doubles ( $M^d$ ) transition counts over the lattice (ordered by box size and count-number, respectively)

Figure 4.8 thus gives the resulting average prediction error for a given number,  $x$ , of complete Markov chains obtained by Gibbs sampling [23] of the indicated singles and doubles play area transition matrices (with error bars given by the standard error of mean determined from 20 samples). It may be observed that  $x \approx 10$  complete game-play sequences is sufficient to identify the play area redefinition involved in transiting from singles to doubles for the specified game parameters.

For the real data, Doubles play is simulated using the real data by multiplying the baseline ( $x$ -axis) by 1.33 (centralized at the court center) so as to extend the legal play into the tram-lines. We fold the court along its symmetric  $x$  and  $y$  axes around the court center to provide better statistical sampling (generating a lattice of 36 elements) and also introduce a weighting proportional to the physical size of court box to ensure the validity of the ‘within box’ uniform distribution assumption as far as possible. For this, we use the following horizontal and vertical ordinate values:



**Figure 4.8:** Mean prediction error of simulated game area transformation for a given number of complete singles/doubles serve sequences (x-axis)

$$horizontalLineSet = [40, 100, 127, 208, 289, 316, 376]; \quad (4.47)$$

$$verticalLineSet = [-10, 50, 158, 284, 410, 518, 578]; \quad (4.48)$$

In the above experiment the method returns the estimated transform (in the folded coordinate system):

$$(PA \rightarrow (2, 3, 4, 4)) \rightarrow (PA \rightarrow (2, 2, 4, 4)) \quad (4.49)$$

That is, the system has correctly identified the original play area and made a correct identification of its redefinition (differing in no ordinate values); a residual graph is



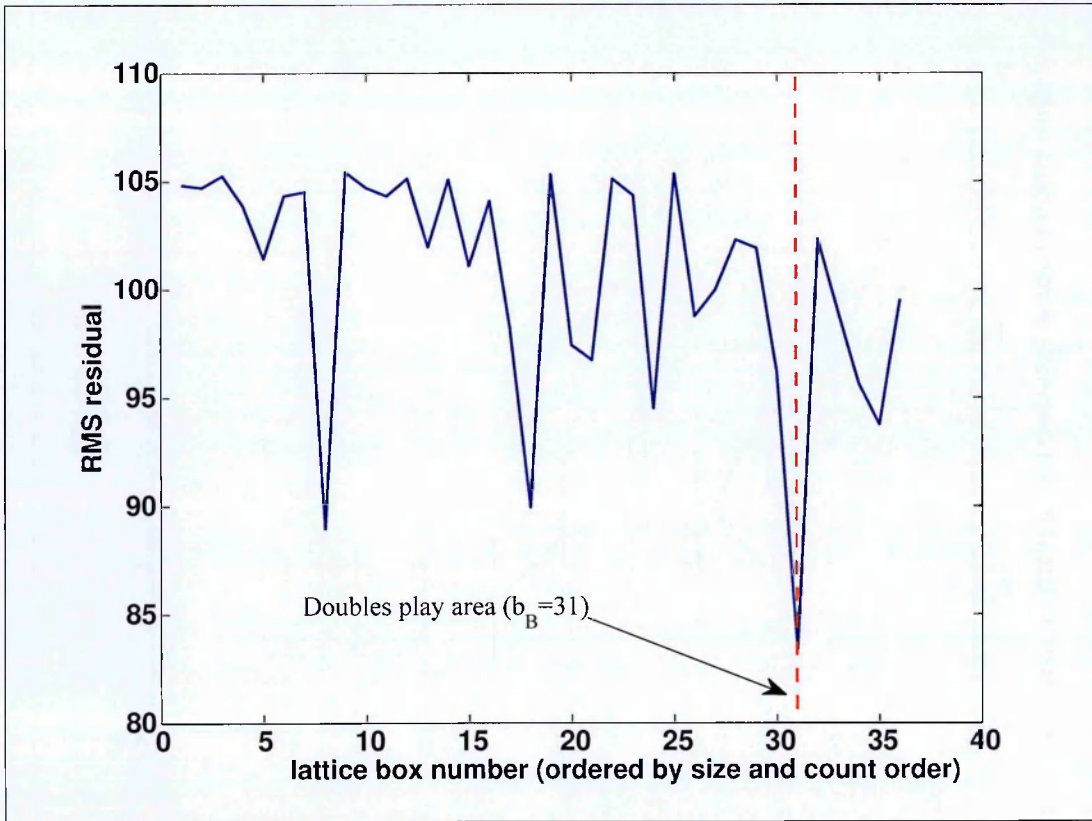


Figure 4.9: RMS residual over  $b_B$  for  $b_A = \text{play area}$ .

given in Figure 4.9. The method is hence sufficiently robust to accommodate any systematic deviations from uniformity in the play sequence distribution.

#### 4.2.5 Conclusion

In this section, we set out, within the context of sport video annotation, to address the problem of anomaly *rectification*; the adaptation of an existing learning mechanism to a change of domain. Consequently, we proposed a novel HMM induction strategy tuned for court-game environments that maps ‘hidden’ game-play states into a court lattice using a deconvolution-like strategy. The system was able to correctly determine transitions in the definition of a play area on both real and simulated data of tennis singles and doubles.

In the next section, we introduce another approach to address the problem of anomaly detection with a relevant rule adaptation mechanism.

### 4.3 Classifier Disparity based Anomaly Detection and Convex Hulling of Anomaly States for Anomaly Rectification

As discussed earlier in Chapter 2 and in Section 4.2, a key concept in machine perception is how to adaptively build upon existing capabilities so as to permit novel functionalities. Implicit in this are the notions of *anomaly detection* and *learning transfer*. A perceptual system must firstly determine at what point the existing learned model ceases to apply, and secondly, what aspects of the existing model can be brought to bear on the newly-defined learning domain. *Anomalies* must thus be distinguished from mere *outliers*, i.e. cases in which the learned model has failed to produce a clear response; it is also necessary to distinguish novel (but meaningful) input from misclassification error within the existing models. We thus apply a methodology of anomaly detection based on comparing the outputs of strong and weak classifiers [184, 202] to the problem of detecting the rule-incongruence involved in the transition from singles to doubles tennis videos. We then demonstrate how the detected anomalies can be used to transfer learning from one (initially known) rule-governed structure to another. We use a convex-hulling approach to address the notion of rule adaptation i.e. rule updating (an application of anomaly rectification). Framework for carrying out this method is introduced in the next section following its methodological details and experimental results.

#### 4.3.1 Introduction

As discussed in Section 4.2, autonomous systems should be able to accommodate novel inputs as a matter of course like humans (as in [173]). The anomaly detection problem, as defined in Chapter 2, is typically one of distinguishing novel (but meaningful) input from misclassification error within existing models. By extension, the *treatment* of anomalies so determined involves adapting the existing domain model to accommodate the anomalies in a robust manner, maximizing the transfer of learning from the original domain so as to avoid over-adaptation to outliers (as opposed to merely incongruent

events). That is, we seek to make conservative assumptions when adapting the system.

The composite system for detecting and treating anomaly should thus be capable of bootstrapping novel representations via the interaction between the two processes. In this section, we aim to demonstrate this principle with respect to the redefinition of key entities designated by the domain rules, such that the redefinition renders the existing rule base *non-anomalous*. We thus implicitly designate a new domain (or context) by the application of anomaly detection.

Our chosen framework for anomaly detection is that advocated in [184, 202] which distinguishes outliers from anomalies via the disparity between a generalized context classifier (when giving a low confidence output) and a combination of ‘specific-level’ classifiers (generating a high confidence output). The classifier disparity leading to the anomaly detection can equally be characterized as being between strongly constrained (contextual) and weakly constrained (non-contextual) classifiers [20]. A similar approach can be used for model updating and acquisition within the context of tracking [200] and for the simultaneous learning of motion and appearance [201]. Such tracking systems explicitly address the *loss-of-lock* problem that occurs without model updating.

In this section we consider anomaly detection in the context of sporting events. What we propose here is a system that will detect when the rules of tennis matches change. We start with a system trained to follow singles matches, and then change the input material to doubles matches. The system should then start to flag anomalies, in particular, events relating to the court area considered to be “in play”.

The system is based on an existing tennis annotation system [81, 192] (also discussed in Section 3.2), which is used to generate data for the anomaly detection. This system provides basic video analysis tools: de-interlacing, lens correction, and shot segmentation and classification. It computes a background mosaic, which it uses to locate foreground objects and hence track the players. By locating the court lines, it computes the projection between the camera and ground plane using a court model. It is also able to track the ball; this is described in more detail in the next section.

In the next section we describe the weak classifiers and their integration. In Section 4.3.3 we discuss the anomaly detection mechanism. We describe some experiments to validate

the ideas in Section 4.3.4, incorporating the results into the anomaly-adaptation/rule-update stage in the immediately following section. We conclude in Section 4.3.6.

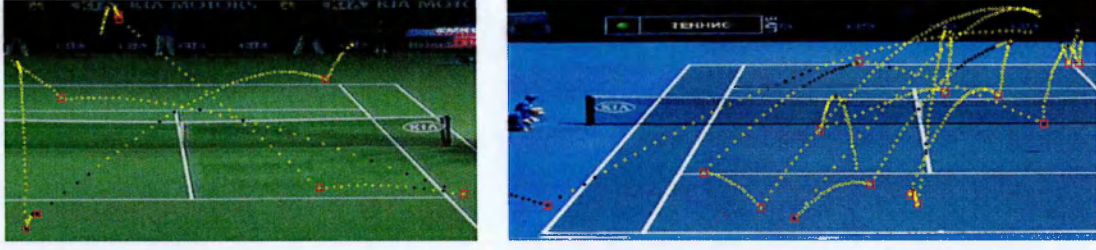
### 4.3.2 Weak Classifiers

#### Ball event recognition

Ball event recognition is one of the weak classifiers we employ. In the following, we first briefly describe the tennis ball tracker, then introduce an HMM-based ball event classifier.

**Tennis ball tracking:** To detect the key ball events that describe how the match progresses, e.g. the tennis ball being hit or bouncing on the ground, the tracking of the tennis ball in the play shots is required. This is a challenging task: small objects usually have fewer features to detect and are more vulnerable to distractions; the movement of the tennis ball is so fast that sometimes it is blurred into the background, and is also subject to temporary occlusion and sudden change of motion direction. Even worse, motion blur, occlusion, and abrupt motion change tend to happen together: when the ball is close to one of the players. To tackle these difficulties, we propose a ball tracker based on [192] with the following sequence of operations:

- (i) Candidate blobs are found by background subtraction.
- (ii) Blobs are then classified as ball / not ball using their size, shape and gradient direction at blob boundary.
- (iii) “Tracklets” are established in the form of 2nd-order (i.e. roughly parabolic) trajectories. These correspond to intervals when the ball is in free flight.
- (iv) A graph-theoretic data association technique is used to link tracklets into complete ball tracks. Where the ball disappears off the top of the frame and reappears, the tracks are linked.
- (v) By analyzing the ball tracks, sudden changes in velocity are detected as “ball events”. These events will be classified in an HMM-based classifier, to provide information of how the tennis game progresses.



**Figure 4.10:** Two examples of the final ball tracking results with ball event detection. Yellow dots: detected ball positions. Black dots: interpolated ball positions. Red squares: detected ball events. In the left example, there is one false positive and one false negative in ball event detection. In the right example, there are a few false negatives.

**HMM-based ball event recognition:** The key event candidates of the ball tracking module need to be classified into serve, bounce, hit, net, etc. The higher the accuracy of the event detection and classification stage the less likely it is that the high level interpretation module may misinterpret the event sequences. A set of continuous-density left-to-right first-order HMMs, i.e.;

$$\Lambda = \{\lambda_1, \dots, \lambda_k, \dots, \lambda_K\} \quad (4.50)$$

are used to analyze the ball trajectory dynamics and recognize events regionally within the tracked ball trajectory, based on [5], but using the detected ball motion changes to localize events.  $K$  is the number of event types in a tennis game, including a null event needed to identify false positives in event candidates. An observation,  $\mathbf{o}_t$ , at time  $t$ , is composed of the velocity and acceleration of the ball position in the mosaic domain:

$$\mathbf{o}_t = \{\dot{\mathbf{x}}_t, \ddot{\mathbf{x}}_t\} \quad (4.51)$$

To classify an event at a time  $t$ , a number of observations, i.e.,

$$\mathbf{O}_t = \mathbf{o}_{t-W}, \mathbf{o}_{t-W+1}, \dots, \mathbf{o}_{t+W} \quad (4.52)$$

are considered within a window of size  $2W + 1$ . Each HMM is characterized by three probability measures: the state transition probability distribution matrix  $\theta$ , the obser-



vation probability distribution  $\eta$  and the initial state distribution  $\pi$ , defined for a set of  $N$  states i.e.,

$$S = (s_1, s_2, \dots, s_N) \quad (4.53)$$

and ball information observation sequence  $\mathbf{O}_t$ . Each state  $s_j$  is represented by a number,  $G_j$ , of Gaussian mixture components. Given a set of training examples corresponding to a particular model, the model parameters are determined by the Baum-Welch algorithm [195]. Thus, provided that a sufficient number of representative examples of each event can be collected, an HMM can be constructed which implicitly models the sources of variability in the ball trajectory dynamics around events.

Once the HMMs are trained, the most likely state sequence for a new observation sequence is calculated for each model using the Viterbi algorithm [195]. The event is then classified by computing

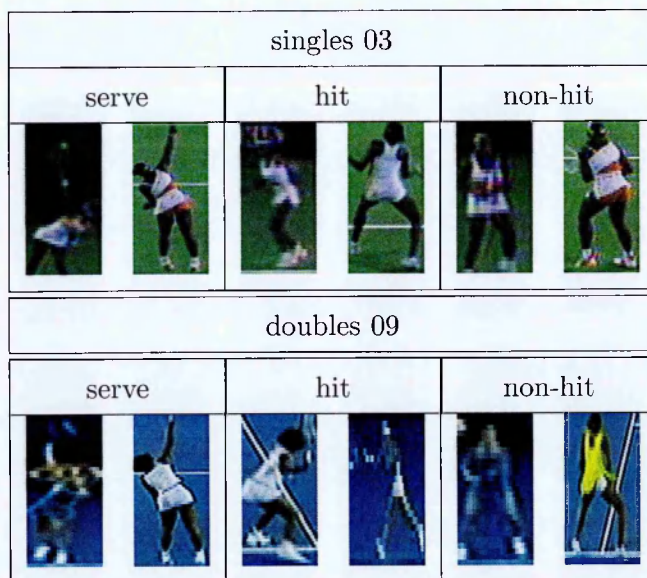
$$\hat{k} = \arg \max_k (P(\mathbf{O}_t | \lambda_k)) \quad (4.54)$$

For every ball trajectory, the first task is to identify when the serve takes place. The first few key event candidates are searched; once the serve position and time are determined, the subsequent key events candidates are classified into their most probable categories, and the null events, considered false positives, are ignored. For recognized bounce events, the position of the ball bounce on the court is determined in court coordinates.

This process can have some false negatives due to ball occlusion and smooth interpolation etc. [192]. This happens often when there is a long time gap between two recognized events. To recover from such suspected false negatives, an exhaustive search is used to find other likely events in such gaps.

### Action recognition

In tennis games the background can easily be tracked, which enables the use of heuristics to robustly segment player candidates, as explained in [81]. To reduce the number of



**Figure 4.11:** Sample images and detected players performing each primitive action of tennis.

false positives, we extract bounding boxes of the moving blobs and merge the ones that are close to each other. Next, geometric and motion constraints are applied to further remove false positives. A map of likely player positions is built by cumulating player bounding boxes from the training set. A low threshold on this map disregards bounding boxes from the umpires and ball boys/girls. In subsequent frames, the players are tracked with a particle filter. Figure 4.11 shows some resulting players detected in this manner, performing different actions.

Given the location of each player, we extract a single spatio-temporal descriptor at the center of the player's bounding box, with a spatial support equal to the maximum between the width and height of the box. The temporal support was set to 12 frames. This value was determined using the validation set of the KTH dataset. As a spatio-temporal descriptor, the 3DHOG (histogram of oriented gradients) method of Klaser et al. [83] is chosen. This method gave state-of-the-art results in recent benchmarks of Wang et al. [182] and has a number of advantages in terms of efficiency and stability over other methods. Previously, 3DHOG has only been evaluated in bag-of-visual-words (BoW) frameworks.

In preliminary experiments, we observed that if players are detected, a single 3DHOG feature extraction followed by classification with kernel LDA gives better performance

than an approach based on BoW with key-point detection.

Three actions are classified: *serve*, *hit* and *non-hit*. A *hit* is defined by the moment a player hits the ball, if this is not a *serve* action. *Non-hit* refers to any other action, e.g. if a player swings the racket without touching the ball. Separate classifiers are trained for near and far players. Their location w.r.t. the court lines is easily computed given the estimated projection matrix. For training, we used only samples extracted when a change of ball velocity is detected. For the test sequences, we output results for every frame. Classification was done with Kernel LDA using a one-against-rest set-up.

We determine the classification results using a majority voting scheme in a temporal window. We also post-process them by imposing these constraints that are appropriate for court games:

- (i) Players are only considered for action classification if they are close to the ball, otherwise the action is set to *non-hit*;
- (ii) We assume that the detected *hits* are actually *serves* at the beginning of a play shot;
- (iii) At the later moments, serves are no longer enabled, i.e., if a serve is detected later in a play shot, the action is classified as a *hit*.

This enables overhead-hits (which are visually the same as *serves*) to be classified as *hits*. In order to provide a confidence measure for the next steps of this work, we use the classification scores from KLDA (normalized distance to the decision boundary).

### Bounce position uncertainty

As the ball position measurements and camera calibration are subject to errors, the probability values near the boundaries of parts of the court will bleed into the neighboring regions. This can be modeled by a convolution between the probability function

$$\hat{P}(\text{bounce}_{in}|\mathbf{x}_t), \quad (4.55)$$

where  $\mathbf{x}_t$  is the ball position, and the measurement error function  $p(e)$  which is assumed to be Gaussian with zero mean and standard deviation  $\sigma_{ball}$ .

$$P(bounce_{in}|\mathbf{x}_t) = \int_{\psi} \hat{P}(bounce_{in}|\psi)p(\mathbf{x}_t - \psi)d\psi \quad (4.56)$$

Finally, the probability of  $bounce_{out}$  is given by:

$$P(bounce_{out}|\mathbf{x}_t) = 1 - P(bounce_{in}|\mathbf{x}_t) \quad (4.57)$$

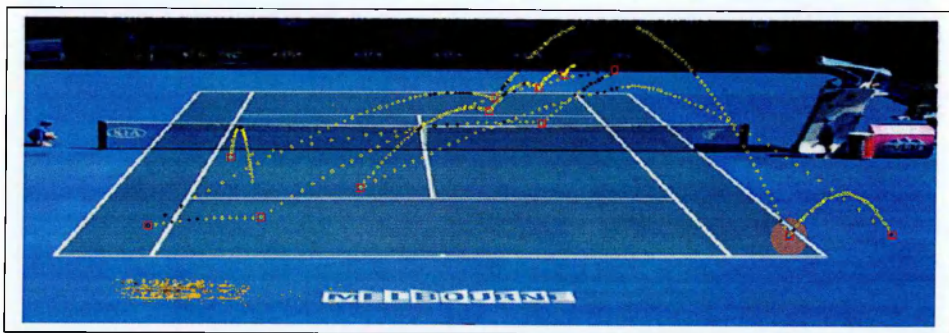
### Combining evidence

The sequence of events is determined by the output of the ball event recognition HMM. Firstly a serve is searched for. If “serve” is one of the 4 most probable HMM hypotheses of an event, that event is deemed to be the serve, and the search is terminated. The remaining events are then classified on the basis of the most probable HMM hypothesis. The entire ball trajectory is then searched for possible missed events: e.g. if consecutive events are bounces on opposite ends of the court, it is likely that a hit was missed.

Sequences of events that start with a serve are passed to the context classification stage. Event sequences are composed of some 17 event types (see [87]). Each event is assigned a confidence, based on the HMM posterior probabilities and, for hits, the action confidences. The combination rules are at present a set of Boolean heuristics, based on human experience. Bounce events are also assigned a separate confidence, based on the bounce position uncertainty.

#### 4.3.3 Context Classification

To detect incongruence, we devise an HMM stage similar to the high-level HMM used to follow the evolution of a tennis match as described in [87]. Here, each sequence of events starting with a serve is analyzed to see if it is a failed serve or a point given to one of the two opponents. The aim is to find sequences of events in which the temporal context classifier reaches a decision about awarding a point before the end of play. Thus, in an anomalous situation, a number of events will still be observed after the



**Figure 4.12:** Doubles tennis ball tranjectory in a complete play-shot with a highlighted anomaly.

decision has been taken by this awarding mechanism. However, the observed sequence of events will only be considered as anomalous when the confidence associated with its events by the weak classifier is high enough. If the reported events are correct, the only event that will create such anomaly is a bounce outside the play area. In the case of the ball is clearly out in singles tennis, the play will stop either immediately or after few ball events.

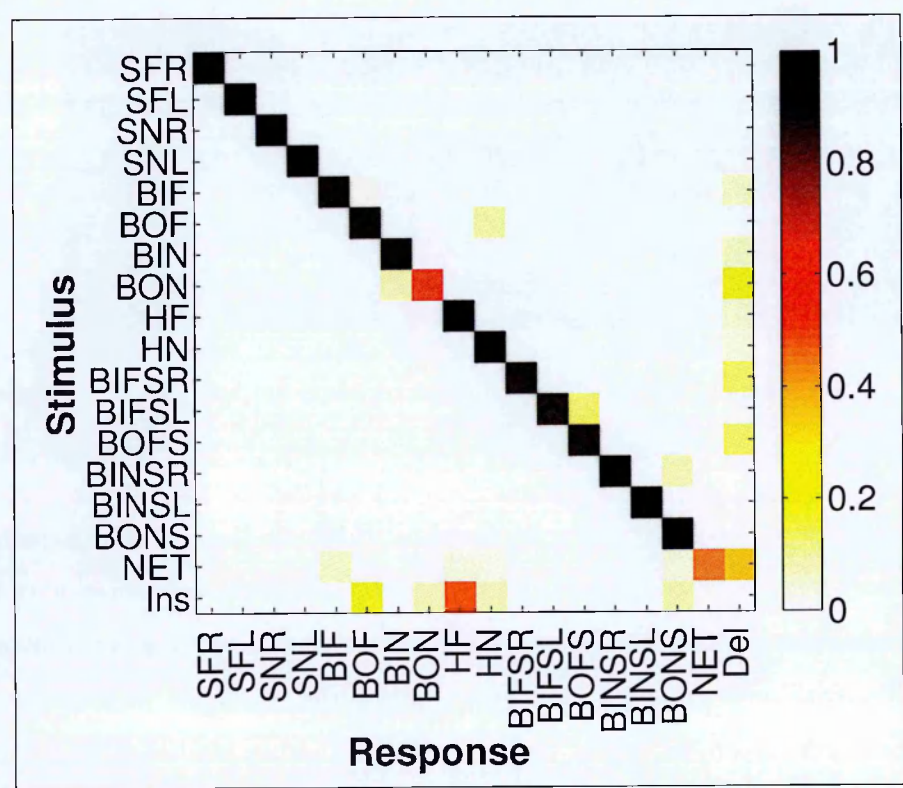
In the doubles tennis, however, the tram lines are part of the play area and bounces in the tram lines will be seen as anomalous for an automatic system that is trained on singles tennis. (Note that the sequence of events that goes into the context classifier does not have multiple hypotheses except when there is uncertainty about bounce in or out (Eq. 4.56)). Through direct observation of singles tennis matches, we have established that the number of events reported subsequent to a clear bounce occurring outside of the legitimate play area does not appear to exceed four events. This is consequently our basis for classification of context.

An example anomaly is highlighted in Figure 4.12 where a complete ball trajectory of a play-shot in doubles tennis with a tramline bounce is shown.

#### 4.3.4 Experiments

Experiments were carried out on data from two singles tennis matches and one doubles match. Training was done using 58 play shots of Women's final of the 2003 Australian Open tournament while 78 play shots of Men's final of the same tournament were used

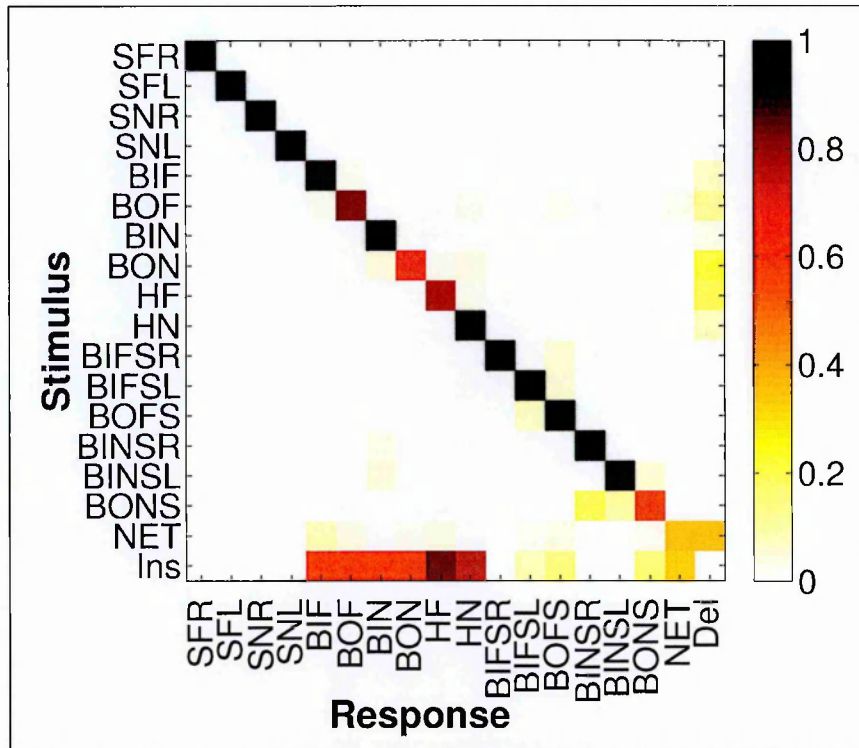




**Figure 4.13:** Singles (validation dataset) - Confusion matrix of recognized events, K. We use the same labels as [5].

for validation. The test data is composed of 163 play-shots of doubles Women’s match of the 2008 Australian Open tournament. The data is manually annotated and 9 HMMs with 3 emitting states and 256 Gaussian mixture components per state modeling ball events were trained using the training data. The performance and parameterization of these HMMs was optimized on the validation data. A window size of 7 observations is selected ( $W = 3$  in Sec. 4.3.2). An accuracy of 88.73% event recognition was reached on the validation data, see Figure 4.13. The last column of the matrix represents the number of deletions and the last row represents the number of insertions.

The confidence measures for the validation data were then used to find appropriate thresholds for rejecting sequences of events that are anomalous due to processing errors rather than genuine bounces out of the play area (Figure 4.15). The x-axis shows the minimum confidence reported on events up to the point of decision made about the event sequence while the y-axis shows the minimum confidence reported on bounces in



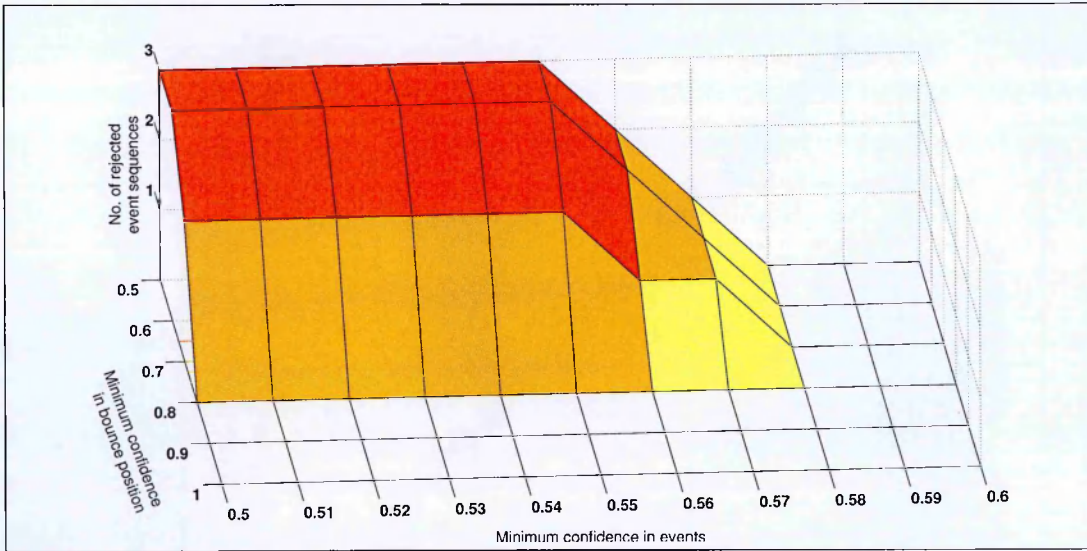
**Figure 4.14:** Doubles (test dataset) - Confusion matrix of recognized events, K

or out the play area up to that point. The number of event sequences where the score decision is taken before the play ends are shown on the z-axis.

It can be seen that a threshold of 0.8 in the bounce position confidence and 0.58 in the event recognition confidence lead to no false positives on sequences from singles. Applied on the doubles data, an accuracy of 83.18% event recognition was obtained using the parameters optimized on the singles data (Figure 4.14). The anomaly detector was able to detect 6 event sequences that contain anomaly, i.e. evident bounce in the tram lines followed by 5 or more events, out of a total of 21 anomaly sequences.

The number of detections are largely limited by the decision confidence filter set on tennis singles. Most of the anomalous bounces take place very close to the inner tramlines triggering a low confidence in the bounce position uncertainty or are followed by very few further exchanges resulting in a high confidence in the context classifier. However, the system is still able to detect a significant number of anomalies that are able detect a domain change.





**Figure 4.15:** Number of event sequences of the validation set that contain errors with varied confidence thresholds

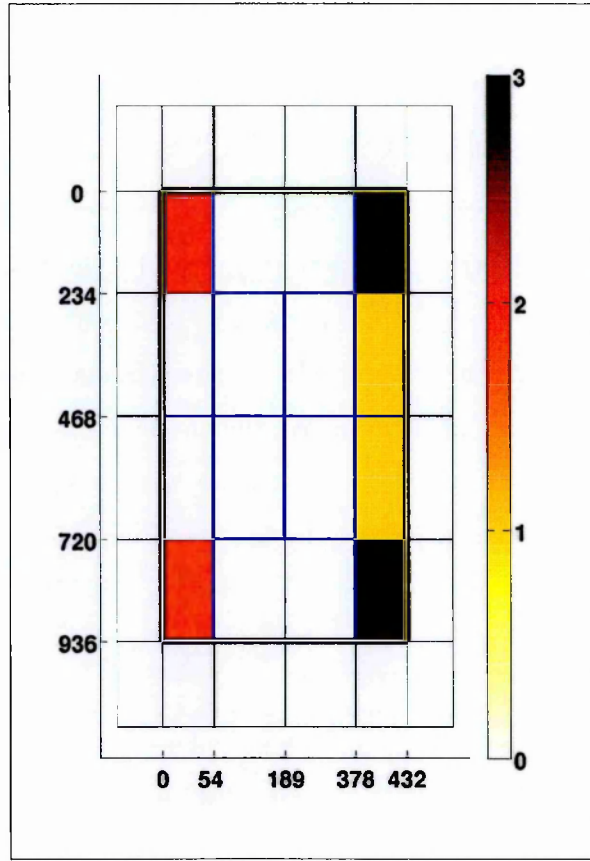
#### 4.3.5 Data association/rule updating

Having identified a discrete set of anomalous events in the manner indicated above, we proceed to an analysis of their import. A histogram of the detected anomalies superimposed on the court delineation obtained by extrapolation of detected horizontal and vertical court-line sets is given in Figure 4.16. We also assume two axes of symmetry around the horizontal and vertical mid lines.

The determination of changes to play area definitions via events histogram is generally complex, requiring stochastic evaluation across the full lattice of possibilities [76] as also discussed in Section 4.2. However, in the absence of false positives, and given the relatively complete sampling of the relevant court area it becomes possible to simplify the process. In particular, if we assume that only the main play area is susceptible to boundary redefinition, then the convex hull of the anomaly-triggering events is sufficient to uniquely quantify this redefinition.

In our case, this redefines the older singles play area coordinates  $\{0,936,54,378\}$  to a new play area with coordinates  $\{0,936,0,432\}$  shown in Figure 4.16, where rectangles are defined by  $\{\text{first row, last row, first column, last column}\}$ , in inches. This identification of new expanded play area means that rules associated with activity in those areas





**Figure 4.16:** Detected anomaly triggering events of the test set for Tennis Doubles

now relate to the new area. Since the old area is incorporated within the novel area according to lattice inclusion, all of the detected anomalies disappear with the play area redefinition. In fact, the identified area  $\{0,936,0,432\}$  corresponds exactly to the tennis doubles play area (i.e. the area incorporating the ‘tram-lines’).

#### 4.3.6 Conclusions

In this section, we set out to implement an anomaly detection method in the context of sport video annotation, and to build upon it using the notion of *learning transfer* in order to incorporate the detected anomalies within the existing domain model in a conservative fashion.

In our experiments, the domain model consists of a fixed game rule structure applied to detected low-level events occurring within delineated areas (which vary for different

game types). On the assumption that anomalies are due to the presence of a novel game domain, the problem is consequently one of determining the most appropriate redefinition of play areas required to eliminate the anomaly.

We thus applied an anomaly detection methodology to the problem of detecting the rule-incongruence involved in the transition from singles to doubles tennis videos, and proceeded to demonstrate how it may be extended so as to transfer learning from the one rule-governed structure to another via the redefinition of the main play area in terms of the convex hull of the detected anomalies. We thereby delineate two distinct rule domains or *contexts* within which the low-level action and event detectors and classifiers function. This was uniquely rendered possible by the absence of false positives in the anomaly detection phase; a more stochastic methods (such as that of Section 4.2) would be considered were this not the case.

## 4.4 Discussion and Summary

In this chapter, we presented methodologies for addressing the problems of anomaly detection and rectification for court based sports such as Tennis singles and doubles. We demonstrated the ability of these methodologies in the context of an adaptive and autonomous court based sports video annotation system.

For this purpose, we first introduced a court-model based method that uses the court structure in the form of a lattice for two related modalities of singles and doubles Tennis to tackle the problems of anomaly detection and rectification. This was achieved by using a novel HMM induction strategy, proposed in this context. Experimental results were shown using real and simulated tennis datasets to demonstrate adaptability of this method by identifying the change in playing area when going from singles to doubles Tennis.

We also introduced another anomaly detection methodology in this chapter, based on the disparity between the low-level vision based (weak) classifiers and the high-level contextual (strong) classifier. Another approach to address the problem of rule adaptation is also proposed that employs Convex hulling of the anomalous states. Experimen-

---

tal results using the datasets extracted from the vision based annotation system and the ground-truth Tennis annotator were also shown, where the new playing location is identified for doubles tennis.

Methodologies introduced in this chapter are important in the context of an automated and adaptive sports video annotation system. Anomaly detection and rectification features within the system can raise flags when the input domain is changed, and accordingly relevant changes can be made in the knowledge base thereafter to accommodate the new domain rules resulting in an adaptive system.

For a more general approach, a framework is required that is capable of learning the rules (i.e. rules of the game in sports) within the input domain by using its observations. This will enable generalization of the “high-level” contextual analysis module of Figure 3.1, which is originally built using hard-wired tennis game rules. Therefore, we introduce a generic rule induction framework in the next chapter capable of learning hierarchical rule structures of tennis and other domains.

In the context of anomaly detection, this means, anomaly flags could also be triggered in the rule structures and rectified via altering relevant sections of the learnt rule model to accommodate novel domains. Rule adaptation in such a scenario can also be achieved leveraging an already existing rule structure by allowing transfer learning. However, anomaly detection using rule induction is beyond the scope of this thesis.

## Rule Induction in the Context of Automated Sports Video Annotation

In this chapter, we propose four variants of a novel hierarchical HMM strategy for rule induction in the context of automated sports video annotation including a Multi-Level Chinese Takeaway Process (MLCTP) based on the Chinese Restaurant Process and a novel Cartesian Product Label-based Hierarchical Bottom-up Clustering (CLHBC) method that employs prior information contained within label structures. Our results show significant improvement by comparison against the flat Markov model: optimal performance is obtained using a hybrid method which combines the MLCTP generated hierarchical structures with CLHBC generated event labels. We also show that the methods proposed are generalizable to other rule-based environments including human driving behavior and human actions.

For an adaptive and automated annotation system, a generic rule induction framework can help in establishing a reliable anomaly detection system for knowledge transfer. Using the induced rule structures, measuring the knowledge shared between various input domains can be quantified in a robust manner reducing the need for re-training every time a new domain is introduced. If, at all, the new domain follows a completely disjoint rule model then the induction framework presented in this chapter, is able to infer related new rule structures.

## 5.1 Introduction

As introduced in Chapter 1, multimedia data production has grown exponentially in recent time. Sports videos have a high demand for automated annotation as there is considerable interest in browsing key events (such as goals in football). Complete annotation may also be used to extract match statistics and to construct performance analysis of teams. As established in Chapter 2, it would be useful to understand contents of the video automatically in sports [35, 130, 196]. For this purpose, a generic rule induction framework can provide a robust context analysis module.

Sports videos consist of rich multimedia content, as well as contextual details. Key temporal event information is critical in understanding sports videos. Sports games in general have a rule structure, built around low level visual events which are further interpreted as game events and similar high level contextual information. Events can thus be expressed in the form of a hierarchical structure. For example, in a game of tennis, initial low level visual events include tennis ball transitions within the court, and player movements enacting game play on the court surface. These transitions can be interpreted in a more contextualized form such as a “hit” taking place at a particular location on a court box. These high-level events can then be combined to describe the tennis game incorporating all the rule salient temporal details as annotations.

As introduced in Chapter 2, Hidden Markov Models (HMMs) [134] are often used to represent stochastic processes and can effectively model temporal sequences of data (such as stock market [61], audio/video signals [96], and patient’s Electrocardiography (ECG) [91] etc.). However, as indicated, some domains such as sports games in general are hierarchical in nature, with a clear delineation between low-level visual representations and progressively higher levels of contextual interpretations. If a game is to be modeled stochastically, with various levels of progressive abstractions, this implies the use of the *hierarchical* Hidden Markov Models (hHMMs)[49] to model game transitions at different levels of contextual interpretations.

However, a particular disadvantage of the classical HMM framework is that it generally requires the number of states to be fixed a priori, and in practical applications they are usually fixed heuristically. Teh et al. [168] have proposed a non-parametric Bayesian

implementation of the HMM in which the hierarchical Dirichlet process (HDP) provides a prior distribution over countably infinite state spaces resulting in a generalized version of HMM. Hierarchical Dirichlet process Hidden Markov Models (HDP-HMMs) have been effectively employed in tackling different problems such as visual scene recognition [82], and the modeling of genetic recombination [189] etc, as established in Chapter 2.

Our aim is to achieve automated stochastic rule induction for a rule-based sport game environment. We make use of the non-parametric Chinese Restaurant Process (CRP) [4] to produce hierarchical structures with states and a Stick-Breaking construction [153] to generate their probabilistic state transitions i.e., we systematically parametrize hHMMs to build a game rule model. As a variant on this approach, we also propose a novel label-based hierarchical method to build hHMMs and show the significance of having prior knowledge of a labeled system in the construction of the hierarchy.

We thus compare a number of derived hHMM models against the flat Markov Model which serves as the baseline for all our methodological variants;

Firstly, we propose a new label-based method in Section 5.2, that takes into account the actual label structure that defines a particular game play sequence in order to define an hHMM generation method that proceeds in a bottom-up, data driven fashion. We call this methodological variant, Cartesian Product Label-based Hierarchical Bottom-up Clustering (CLHBC).

A further variant is introduced via a novel implementation of the Chinese Restaurant Process called the Multi-Level Chinese Takeaway Process (MLCTP). This is a constrained version of the standard CRP that is more relevant to applications with a limited state space i.e., where the number of rule-defining events are known and a limited rule depth is present i.e., rule induction occurs under a certain unknown, but relatively limited number of levels.

MLCTP does not intrinsically exploit labeled states and we speculate that the highest likelihood inferred rule structure given a set of hyper-parameters representing the MLCTP model can be further improved via employing the label structures. Thus, we also propose two hybrid methods in Section 5.4, that combine the unlabeled MLCTP with the labeled structure from the Flat Markov model and CLHBC. The main idea is

---

thus to combine the hierarchical topological structures of MLCTP with the event label transition probabilities.

We show comparative results of all the proposed methods in Section 5.5 using the following various datasets from different domains:

- **Badminton Rules Domain:** This dataset contains ground truth annotated events in Badminton (Mens Singles, Czech Vs GB, Beijing Olympics, 2008).
- **Tennis Rules Domain:** This dataset contains ground truth annotated events extracted from a complete Tennis match (Serena Williams VS Venus Williams, Final, Womens singles, Australian Open 2003).
- **Tennis Rules Domain:** This dataset contains labeled sequential events in Tennis obtained via a computer vision based annotator ([89, 108]) (Serena Williams VS Venus Williams, Final, Womens Singles, Australian Open 2003 and Andre Agassi VS Rainer Schttler, Final, Australian Open 2003).
- **Highway Rules Domain:** This dataset contains ground truth annotated events obtained from a camera-equipped car driven across a city [154].
- **Website Domain:** This sequential dataset contains visit counts of different pages for MSNBC.com on September 28, 1999 by many different users (used in [21]).
- **Human Activity Dataset:** This sequential dataset contains recordings of five sensor-tagged people performing different actions [72] such as sitting, walking, falling and lying etc.

Summary of this chapter is presented in Section 5.6.

## 5.2 Cartesian Product Label-Based Hierarchical Bottom-up Clustering

### 5.2.1 Introduction

Sports games have a specific rule structure built around temporal events that are based on transitions between labeled states according to structured game rules e.g.:

#### Football

kick, pass left, pass right etc

#### Tennis/Badminton

Serve near left, hit far etc

#### Cricket

Square drive, straight drive, full length delivery etc.

Generally, labeled events contain not only temporal information but also spatial details, for instance in tennis, a *Serve* followed by a *Bounce* taking place at the *Out* and *Far* side of the court can be represented by a concatenated descriptor “*BOFS*” [86], as introduced in Chapter 3 (see Table 3.1). Thus, each event label is constructed by incorporating relevant sub-labels providing detailed spatio-temporal information related to game-play which are crucial for inferring rule structures.

By taking the whole sequence of event labels into account, we can thus represent rule-related information by using the Cartesian combinations of these sub-labels where they collectively constitute a *lattice* in which coarse-grained event labels are clustered bottom-up to form a hierarchical topology that can potentially represent abstract rule structures.

Thus, various hierarchical label clusters obtained using Cartesian products of sub-labels produce different, but meaningful topological structures that are potentially capable of modeling the underlying abstract rule structure of the game. This is autonomously achieved by taking all possible permutations of the label order that constitutes these



hierarchical structures and, with a predefined selection criterion, a rule-like topological structure is chosen. Methodological details of this method are formulated in the next section and a comparative analysis against other methodologies is conducted in Section 5.5.

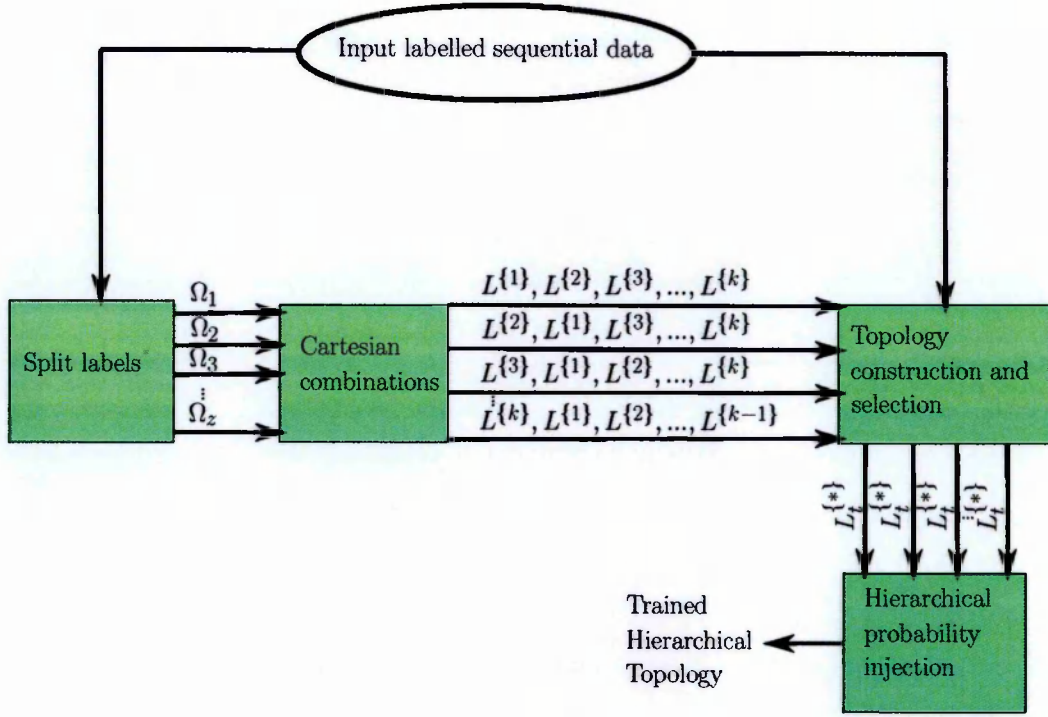
### 5.2.2 Methodology

Sports games have a repetitive rule structure such that a particular sequence of events often repeats during the course of the game-play. For example, a *Serve* followed by a *Bounce* repeats very frequently at the start of every *play-shot* in tennis where a *play-shot* is defined as a sequence of events that starts with a *Serve* and ends with the *point allocation* to one of the players in the case of court games [86]. Additionally, game exchanges can also be interpreted via contextual notations such as game transitions between two players. In the middle of a *play-shot*, they can be represented as *rally*, or a *bounce out* of the court area can be represented as a *point allocation* to either of the players.

Such behavior, with various levels of abstractions, can be modeled using a hierarchical state structure i.e. hierarchical Hidden Markov Models. We propose the *Cartesian Product Label-Based Hierarchical Bottom-up Clustering* method to generate different hierarchical HMMs capable of producing rule structures representing sports games.

Our input to the system is a set of event labels shown in Table 5.1 for badminton and tennis, extracted from [86] and translated into a Cartesian Product notation. We argue, more generally, that in most situations complex labeling scenarios can be treated in this fashion (usually with the proviso that we can introduce a null value,  $\phi$ , where there exists incomplete factorizability of the labels intrinsically - see below).

Labels are thus constituted of various sub-labels which can represent *event* types,  $\Omega_E$ , *distance* from the camera,  $\Omega_D$ , *sides* of the court area  $\Omega_S$ , and *position* with respect to the court lines  $\Omega_P$  etc.



**Figure 5.1:** Cartesian Product Label-Based Hierarchical Bottom-up Clustering

$$\Omega_E = \begin{Bmatrix} S \\ H \\ B \end{Bmatrix}, \Omega_D = \begin{Bmatrix} N \\ F \\ \phi \end{Bmatrix}, \Omega_S = \begin{Bmatrix} L \\ R \\ \phi \end{Bmatrix}, \Omega_P = \begin{Bmatrix} I \\ O \\ \phi \end{Bmatrix} \quad (5.1)$$

where  $\phi$  is the null value in the argument description (henceforth we shall omit this).

To train the model, we divide the stream of input event labels into groups of play-shots (that starts with a serve and ends with a point allocated to either player), for example:

$$SN\phi\phi \rightarrow HF\phi\phi \rightarrow HN\phi\phi \rightarrow HF\phi\phi \rightarrow HN\phi\phi \rightarrow BIF\phi$$

There are repeated sequences within almost every play-shot ( $HF\phi\phi \rightarrow HN\phi\phi$  is repeated twice in the example above): these can potentially form hidden states representing common *meta-labels*, on the next hierarchical level. The method achieves this by combining labels in a manner similar to an explicitly hierarchical Lempel-Ziv-Welch (LZW) encoding [185] i.e. with common labels combined together sequentially to form a parent node e.g. different types of *serves* ( $SN$  and  $SF$ ) can be combined to represent a node labeled as  $S$  representing the *Serve* meta-label (see Figure 5.2). Similarly, an-

**Table 5.1:** Summary of Badminton and Tennis events extracted from [86]

| Event              | Description                             |
|--------------------|---|
| SF $\phi\phi$      | Serve by Far player                     |
| SN $\phi\phi$      | Serve by Near player                    |
| HF $\phi\phi$      | Hit by Far player                       |
| HN $\phi\phi$      | Hit by Near player                      |
| BIF $\phi$         | Bounce Inside Far player's half court   |
| BOF $\phi$         | Bounce Outside Far player's half court  |
| BIN $\phi$         | Bounce Inside Near player's half court  |
| BON $\phi$         | Bounce Outside Near player's half court |
| BOFS (Tennis only) | Bounce Out of Far player's Serve area   |
| BONS (Tennis only) | Bounce Out of Near player's Serve area  |

other combination (by changing label order) can also be formed combining the  $N$  and  $F$  meta-labels to form two separate nodes at the parent level which consequently shall represent game transitions between the *Near* and *Far* side of the court.

These Cartesian meta-labels form the parent level nodes, clustering sets of un-omitted labels beneath it. For example, the string above in terms of *event type* labels,  $\Omega_E$  (achieved via the omission of  $\Omega_D$  labels) looks like:

$$S \rightarrow H \rightarrow H \rightarrow H \rightarrow H \rightarrow BI$$

Play-shots can be represented in the form of other Cartesian label type subsets by changing the label order. Figure 5.1 shows a block diagram representing the Cartesian Product Label-Based Hierarchical Bottom-up Clustering method in context.

Input events sequence contains individual labels,  $L_t$ , describing an event at time  $t$ , and constituted of  $z$  label components drawn from  $\Omega_i$ , such that,

$$\Omega_1 = \begin{Bmatrix} \omega_1^1 \\ \omega_1^2 \\ \omega_1^3 \\ \vdots \\ \phi \end{Bmatrix}, \Omega_2 = \begin{Bmatrix} \omega_2^1 \\ \omega_2^2 \\ \omega_2^3 \\ \vdots \\ \phi \end{Bmatrix}, \Omega_3 = \begin{Bmatrix} \omega_3^1 \\ \omega_3^2 \\ \omega_3^3 \\ \vdots \\ \phi \end{Bmatrix}, \dots, \Omega_z = \begin{Bmatrix} \omega_z^1 \\ \omega_z^2 \\ \omega_z^3 \\ \vdots \\ \phi \end{Bmatrix} \quad (5.2)$$

So, at any given time  $t$ ,  $L_t$  represents the Cartesian products of all the  $\Omega_z$  labels at each instant in the sequence, defining the base of a  $z$ -dimensional *lattice* (i.e. the lattice formed from differing subsets of  $\Omega$  labels).

$$L_t \in \{\Omega_1 \times \Omega_2 \times \Omega_3 \times \dots \times \Omega_z\}^t \quad (5.3)$$

Thus, various Cartesian combinations can be formed within the lattice by progressively omitting  $\Omega_z$  labels, such that, e.g.

$$L_t^{\{k\}} \in \{\Omega_1 \times \dots \times \Omega_{k-1} \times \Omega_{k+1} \times \dots \times \Omega_z\}^t, \quad (5.4)$$

where the omitted label set  $k \subseteq \{1, 2, 3, \dots, z\}$ .

Hence,  $L_t^{\{2\}}$ , with  $z = 3$ , represents Cartesian combination of all of the three labels with the exception of  $\Omega_2$  i.e.,

$$L_t^{\{2\}} \in \{\Omega_1 \times \Omega_3\}^t \quad (5.5)$$

$L_t^{\{k\}}$  is thus composed of a sequence of ordered pairs,  $l_i : i = 1, \dots, t$ , derived from the remaining  $\omega$  labels, such that, in this form, a particular event might look like:

$$l_i^{\{2\}} = (\omega_1^2, \omega_3^1) \quad (5.6)$$

However, note that because label omission is carried out sequentially, not all of the hierarchies within the lattice space are sampled; in fact only a unique hierarchical subset is selected for a particular input label ordering.

### Topology selection criterion

Before sampling the resultant hierarchical structure, we repeat the hierarchy generation process above under different orderings i.e.  $L_t$  is represented via other permutations of  $\Omega_i$ . In the case of the example sequence above,  $SN$  can be represented as  $NS$  and so on (omitting  $\phi$  for simplicity). This results in various other hierarchies which may or may not approximate the domain rules. For this purpose, a selection criterion is introduced via counting the number of *nodes with non-mono child nodes* (excluding the leaf nodes). Resultant hierarchies are ranked according to this criterion e.g. Figure 5.2 has a rank of 4 and a differently ordered *near-far* model has a rank of 3. The hierarchical topology with the highest rank is selected for training by sampling the space of transition probabilities in the hierarchy i.e., by explicitly modeling hierarchical transitions (explained in the next section).

Note that, usually a human annotator implicitly follows a certain label order (typically general-to-specific) that results in a particular form of rule structure. In case of the tennis/badminton games, the label order followed (see Table 5.1) contains an implicit rule structure that results in the topology shown in Figure 5.2. In order to generalize the method's capability and assuming no prior knowledge about label order, a selection criterion that explores all label permutations can autonomously choose a richer rule structure.

### Modeling hierarchical transitions

In the following analysis, we will model transitions within the lattice hierarchy (chosen with the criterion above) on a Markovian basis. However, this means that the model as a whole is not consistent with the Markov property (the higher level 'hidden' hierarchical state transitions effectively constitute a memory). It is, though, still possible to represent the entire hierarchy as an *Implicitly Markovian Model*. This differs from the standard 'flat' Markovian in which  $P_f$  represents a transition likelihood between states  $Q_{n-1}$  and  $Q_n$ , derived by histogramming over components of an observed sequence (or set of sequences),  $S(j), j = 1, \dots, T$ , i.e.:

$$P_f(Q_n|Q_{n-1}) = \frac{1}{\mathbb{F}} \sum_{j=1}^{T-1} f(S(j-1), S(j)) \quad (5.7)$$

where  $f = \begin{cases} 1 & S(j-1) = Q_{n-1}, S(j) = Q_n \\ 0 & \text{otherwise} \end{cases}$  and  $\mathbb{F}$  represents the normalization factor. In the following analysis this flat model will serve as our baseline.

We define this *Implicitly Markovian Model* as follows. In a  $z$ -dimensional CLHBC-generated lattice space,  $q$  levels can be formed (depending on the Cartesian combinations) where  $q \leq z$ , such that a resultant *augmented likelihood*  $\bigwedge^C$ , of event transitions can be computed by considering transitions at all the levels of the constructed hierarchy. The concept of augmented likelihood centres on the modification of observed event likelihoods in order to explicitly favour hierarchicality (i.e. by sampling events at all the levels of the hierarchy).

We introduce a bijective mapping of the constructed hierarchy's leaf states to observations which we use to compute transition likelihood between observations  $\mathcal{E}_{X-1}$  to  $\mathcal{E}_X$ ;  $X = 1, 2, 3, \dots, G$  where  $G$  is the total number of leaf nodes (Figure 5.2 has  $G = 8$ ). This is achieved using the normalized products of all the super-lying parent state transitions via connected nodes resulting in augmented likelihood of state transitions i.e.,

$$\bigwedge^C(\mathcal{E}_X|\mathcal{E}_{X-1}) = \frac{1}{\mathbb{C}} \prod_{h=1}^q \left\{ \frac{1}{\mathbb{N}} \sum_{i=1}^T g(S_h(i-1), S_h(i)) \right\} \quad (5.8)$$

$$\text{where } g = \begin{cases} 1 & S_h(i-1) = Q_{n-1}^h, S_h(i) = Q_n^h \\ 0 & \text{otherwise} \end{cases}$$

$Q_n^h$  is the observed state at level  $h$  of the hierarchy (i.e. under progressive label omission);  $\mathbb{C}$  is a Cartesian normalization factor and  $\mathbb{N}$  is the level-based normalization factor. The *hierarchical probability injection* step computes the augmented likelihood (see Figure 5.1). Probabilities in the hierarchy are computed top-down and injected per level based on Equation 5.8 resulting in a single matrix representation of observation state transitions that are bijectively mapped onto the bottom level leaf nodes.

Note, the label space at the bottom level of the hHMM needs to be fully sampled by the data i.e. such that at least a single instance of each label has been observed

(however, there are no such restriction higher up in the hierarchy). We have not directly distinguished label uncertainty from state uncertainty, since the latter is fully capable of modelling the former.

The Markovian model thus defined differs from the flat model in that transition likelihoods for observed states are biased by progressively higher-level hidden state transitions, for which there exist better sample-statistics (due to coarser-grained transition likelihoods). We thus influence low-level, rapidly-changing, potentially more noise-influenced transitions by higher-level, more slowly-transitioning states. Consequently, we retain all of the advantages associated with the Markov assumption (in particular, the ability to rapidly model sequence likelihoods via transition matrices), while leveraging the descriptive potential of hierarchical modeling.

### Worked example

Consider an example sequence of events, with  $z = 3$  types of labels,

$$\mathcal{E} = l_1 \rightarrow l_2 \rightarrow l_3 \rightarrow l_4 \rightarrow l_5 \rightarrow l_6$$

A three-dimensional lattice of labels is formed. Each event label  $l_t$  may look like  $(\omega_1^1, \omega_2^1, \omega_3^1)$ . After analysing the whole sequence above  $l_1$  to  $l_6$ , different common combinations are extracted at the sub-label level. For example, in tennis or badminton, a sequence of “Hits” can be combined to produce a hidden state semantically equivalent to a “Rally” (these sub-labels are identified and decomposed into a series of  $\Omega_i$  represented in Equations 5.2 and 5.3):

$$(\omega_1^1, \omega_2^1, \phi) \rightarrow (\omega_1^2, \omega_2^2, \phi) \rightarrow (\omega_1^2, \omega_2^1, \phi) \rightarrow (\omega_1^2, \omega_2^2, \phi) \rightarrow (\omega_1^2, \omega_2^1, \phi) \rightarrow (\omega_1^3, \omega_2^1, \omega_3^1)$$

The above sequence can also be represented in its  $\{k\} = \{2, 3\}$  sub-label form (see Equation 5.4) as,

$$(\omega_1^1) \rightarrow (\omega_1^2) \rightarrow (\omega_1^2) \rightarrow (\omega_1^2) \rightarrow (\omega_1^2) \rightarrow (\omega_1^3)$$

Common sequential sub-labels are thus extracted as a meta-label that constitutes a node in the next highest level. In this example, 3 nodes are formed for the sequence such that the augmented likelihood  $\bigwedge^C$  of event transitions (see Equation 5.8) can be

---

computed with  $q = 3$  representing the number of labels and resultant levels,  $G = 4$ , and  $T = 6$ .

In applying the CLHBC model to a Badminton game we find that Cartesian labeling can split the labeled sequential data into various categories of *play shot sequences* demonstrating the applicability of the method with regards to the label structure of events e.g. we autonomously combine labels according to event types (*Serves, Hits etc*). In Figure 5.2 an example of the bottom-up labeling with colors indicating hereditary of states is shown. Events are delineated in accordance with the *play structure* by combining starts, rallies and ends together, in turn constituted by serve, hit and bounce meta-states, respectively. The two transition matrices represent non-zero transition probabilities at each level of the hierarchy (using badminton as an example).

## 5.3 Multi Level Chinese Takeaway Process

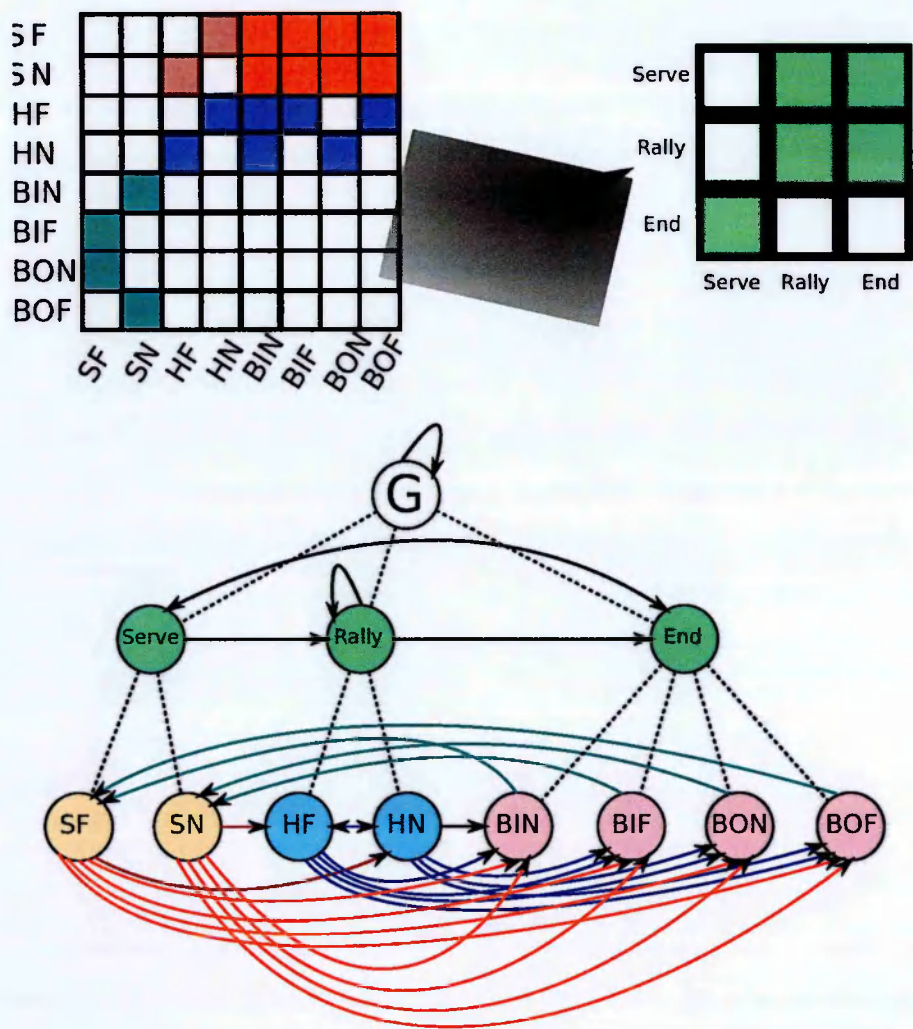
### 5.3.1 Introduction and Motivation

As discussed in Section 5.2.1, court-games are inherently hierarchical in nature and we attempt to create stochastic approximations of the game rules using hierarchical Hidden Markov Models (hHMM) for contextual game description covering various levels of abstractions, ultimately, giving rise to meaningful annotations. As explained in Section 5.2.2, our input observations are a set of events that occur over a temporal sequence marked when a particular event starts happening (Table 5.1).

In a rule based environment, these events contribute meaningful attributes on a contextual level; thus events like *Serves* and *Hits* relate to *Player's actions* while *Near* and *Far* correspond to *Court Locations*. Events can be either described in terms of *Player's actions* or *Rule-Defined* combinations such as *Rallies* and *Game points*.

To build a generic hierarchical HMM framework suitable for characterizing these environments we propose a constrained variant of the widely used Chinese Restaurant Process (CRP) first introduced in [4] that allows us to establish rule structures that are capable of describing the sports game in a compact and efficient fashion. The proposed





**Figure 5.2:** Three Level Cartesian Product Label-Based Hierarchical Bottom-up Clustering with Transition Matrices generated at each level (colored so as to indicate heredity)

method does not intrinsically exploit labeled information (unlike CLHBC of Section 5.2) making it more suitable for applications with limited meta-data.

We refer to it as the Multi-Level Chinese Takeaway Process (MLCTP) and in the next section we explore the methodological details of this particular variant of the classical CRP and its application to rule-based environments primarily sport games (i.e. Tennis and Badminton).

### 5.3.2 Methodology

The Chinese Restaurant Process is a non-parametric stochastic process that is naturally capable of representing grouped sequential data. In a rule-based environment, data can be grouped together in a hierarchy and thus we require a hierarchical CRP for stochastic approximation of rules induced via input observations. CRP's hierarchical version is referred to as the Chinese Restaurant Franchise (CRF) first coined by Teh *et. al.* in [168]. Due to the limited state-space hierarchy of sport rule structures evident from the types and number of events, it is desirable to implement a hierarchical, but also constrained, version of the classical Chinese Restaurant Process which we call the Multi-Level Chinese Takeaway Process. To understand this particular variant of the CRP, we step-wise explore the methodological details of MLCTP. To intuitively understand the process, we make use of an analogy similar to the Chinese Restaurant Process (CRP) [4].

There are three main methodological steps in generating MLCTP-based hierarchical topologies (i.e. hierarchical rule structures); the state generation phase, transition probabilities generation phase, and the hierarchical state transition matrices injection phase;

#### State Generation Phase

This phase is similar to CRP where the number of states is defined by the process via the number of tables. The notion of tables in MLCTP is replaced with takeaways to leverage re-visits and further recommendations to other takeaways (further explained in Section 5.3.2). For the sake of consistency, we replace the notion of tables with takeaways in the first phase.

To start the process, people (tokens) enter a city with infinite number of takeaways and choose a particular takeaway to visit. First person visits the first takeaway in the city with the initial probability equal to 1. The takeaway visit probability,  $v_i$  for the  $i$ th person is thus defined as;

$$P(v_i = c | v_{1:i-1}) = \begin{cases} \frac{o_c}{i-1+\alpha} & \text{if } c \leq \mathcal{C} \\ \frac{\alpha}{i-1+\alpha} & \text{otherwise } c \text{ is the new takeaway} \end{cases} \quad (5.9)$$

Where  $o_c$  is the number of people who have visited the takeaway  $c$ .  $\mathcal{C}$  is the number of takeaways for which  $o_c > 0$  i.e. visited.  $\alpha$  is the concentration parameter. Intuitively, high  $\alpha$  implies more visited takeaways with fewer customers.

We initialize the process of state generation assuming one top level state. We henceforth call the top level the first level. For the second level, we follow the takeaway visit process expressed in Equation 5.9, and generate this level with  $\mathcal{C}^2$  states defined by  $\alpha$ . For each state at this level, Equation 5.9 is followed recursively to generate the third level, where a total number  $\mathcal{C}^3$  states are created and so on. The process continues until the maximum truncation point is reached which is defined by the number of event types in the training dataset. Note that,  $\mathcal{C}^H > \mathcal{C}^{H-1} > \dots \mathcal{C}^2 > \mathcal{C}^1$ , where  $H$  represents the total number of levels i.e. a hierarchy is formed.

At the end of this phase, we thus establish a hierarchical topology with states generated top-down with vertical edges (i.e., representing connections not transitions). Note, this phase is precisely controlled based on the number of events. As such, as soon as the number of states generated by CRP in the next level to be generated exceeds the termination criterion, the process halts and a new topology is generated. Otherwise, the process continues and, if matched, the process proceeds to phase 2, the transition probability generation phase. An example topology is shown in Figure 5.3.

### Topological State Transition Matrix Generation Phase

This second step for generating the state transition matrix involves two major sub-steps; firstly we extract state transition probabilities, defined by Equation 5.9 for all the levels. We define each takeaway visit self transition probability as  ${}_{h'}\delta_{i_h}^h$  for state number  $i_h$  at  $h$ th level with  $h'$  its mother state;

$${}_{h'}\delta_{i_h}^h = \frac{\text{Total number of visits to takeaway } i_h}{\text{Total number of visits via } h'} \quad (5.10)$$

The remaining probability of transition,  ${}_h\psi_{i_h}^h = (1 - {}_h\delta_{i_h}^l)$  from the  $i_h$ th state to all the other states at level  $h$  is further re-distributed by executing a stick-breaking construction as follows. We use hyper-parameter  $\gamma$  for all the states controlling the redistribution of the state transitions. This can be intuitively represented by replacing tables in CRP with takeaways where people are recommended  $\mathcal{C}^h - 1$  other takeaways to try additionally in city  $h$ .

Note, in theory, this phase can also be represented using another implementation of CRP. However, the purpose of introducing SB-construction to represent MLCTP's state transition matrix generation is to differentiate between the two phases so as to make the analogy exact (i.e., an individual's visit and post-visit recommendations).

We start with the stick of length 1. The stick is broken  $(\mathcal{C}^h - 2)$  times to create  $(\mathcal{C}^h - 1)$  partitions representing all other transitions where  $\mathcal{C}^h$  represents the total number of takeaways/states at level  $h$ . Equation 5.12 represents the stick-breaking construction weights:

$${}_h\pi_k^h = {}_h\beta_k^h \prod_{j=1}^{\mathcal{C}^h-2} (1 - {}_h\beta_j^h) \quad (5.11)$$

and, the final weight is defined (due to the finite number of states) as:

$${}_h\pi_{\mathcal{C}^h-1}^h = \sum_{c=1}^{\infty} {}_h\pi_c^h - \left( \sum_1^{\mathcal{C}^h-2} {}_h\pi_k^h \right) \text{ where } \sum_{c=1}^{\infty} {}_h\pi_c^h = 1 \quad (5.12)$$

${}_h\pi_k^h$  represents  $k$ th weight at level  $h$  for state  $i_h$ .

$${}_h\beta_k^h \sim \text{Beta}(1, \gamma) \quad (5.13)$$

Within the levels so generated,  ${}_h\psi_{i_h}^h$  is partitioned in the manner indicated and transitions to all the other states (left to right indexed) are represented with weights:

$${}^*\pi_1^*\psi_*, {}^*\pi_2^*\psi_*, {}^*\pi_3^*\psi_*, \dots \text{ etc.}$$

For each level,  $h$ , state transition matrix is built using self-transitions (i.e. the takeaway visit probability of Equation 5.9) and the remaining probability is distributed across all the other states at level  $h$ , using the stick-breaking construction of Equation 5.12.

**Hyper-parameters:** MLCTP is governed by two hyper-parameters  $\alpha$  and  $\gamma$ .  $\alpha$  controls the number of new states at each level as employed in the classical CRP model. Additionally,  $\alpha$  also defines the self-transition probability for each state (i.e the take-away visit probability). The second hyperparameter  $\gamma$ , is employed in the Beta distribution of Equation 5.13. This controls the size of the stick-break defined in Equation 5.12 which furthermore controls the contribution of the remaining probability.

Various combinations of  $\alpha$  and  $\gamma$  hyper-parameters are used to generate different types of topologies with varying transition probabilities. These parameters are selected based on the number of events and the number of correct-width topological structures. The main focus of this paper has been to infer a rule structure given a sequence of observations and as such, a large range of  $(\alpha, \gamma)$  hyper-parametric pairs are sampled in order to generate potential rule structures.

### Hierarchical State Transition Matrix Injection Phase

The next step is to form a state transition matrix for the whole topological structure. We do that by first forming state transition matrices for each level using all the state transitions extracted in Section 5.3.2 and then use the notion of *probability injection* introduced in Section 5.2.2. Equation 5.8 is employed again to represent the augmented likelihood of transitions between all the leaf states.

Note, the bottom-level states are associated with the input number of labels i.e. we introduce bijective mapping (similar to Section 5.2.2 for CLHBC model) of leaf states to observations which we use to compute the transition likelihood between observations  $\mathcal{E}_{\mathcal{O}-1}$  to  $\mathcal{E}_{\mathcal{O}}$ ;  $\mathcal{O} = 1, 2, 3, \dots, \mathcal{G}$  (see Figure 5.3 where  $\mathcal{G} = 7$ ). This is achieved using the normalized products of all the super-lying parent state transitions via connected nodes such that,

$$\bigwedge^U(\mathcal{E}_{\mathcal{O}}|\mathcal{E}_{\mathcal{O}-1}) = \frac{1}{\mathcal{D}} \prod_{V=H}^1 P(x\zeta_y^V(\mathcal{O})|_{x'}\zeta_{y'}^V(\mathcal{O}-1)) \quad (5.14)$$

where  $\bigwedge^U(\mathcal{E}_{\mathcal{O}}|\mathcal{E}_{\mathcal{O}-1})$  is the augmented likelihood for MLCTP generated state transition between events,  $\mathcal{E}_{\mathcal{O}-1}$  and  $\mathcal{E}_{\mathcal{O}}$ .  $\mathcal{D}$  is the normalization constant and  $H$  is the total

number of levels.  ${}_x\zeta_y^V$  represents state  $\zeta$ , indexed by  $y$ , with its parent state  $x$  and is at level  $V$ , where  $V = H, H - 1, \dots, 1$ , for an input observation index  $\mathcal{O}$ .

All these generated topologies have a few generic properties:

- Each child state has a unique parent but each parent can have one or more than one child state. A state, represented by  ${}_x\zeta_y^h$  has only one  $x$  for each  $y$  at level  $h$  i.e.  $x$  is unique for all  $y$  at  $h$ .
- Transition between two child states at level  $h$  of a single parent state represents self-transition at level  $h + 1$  of the corresponding parent state i.e.:

$$P({}_x\zeta_{y+1}^h | {}_x\zeta_y^h) \implies P({}_{x'}\zeta_x^{h+1} | {}_{x'}\zeta_x^{h+1}) \quad (5.15)$$

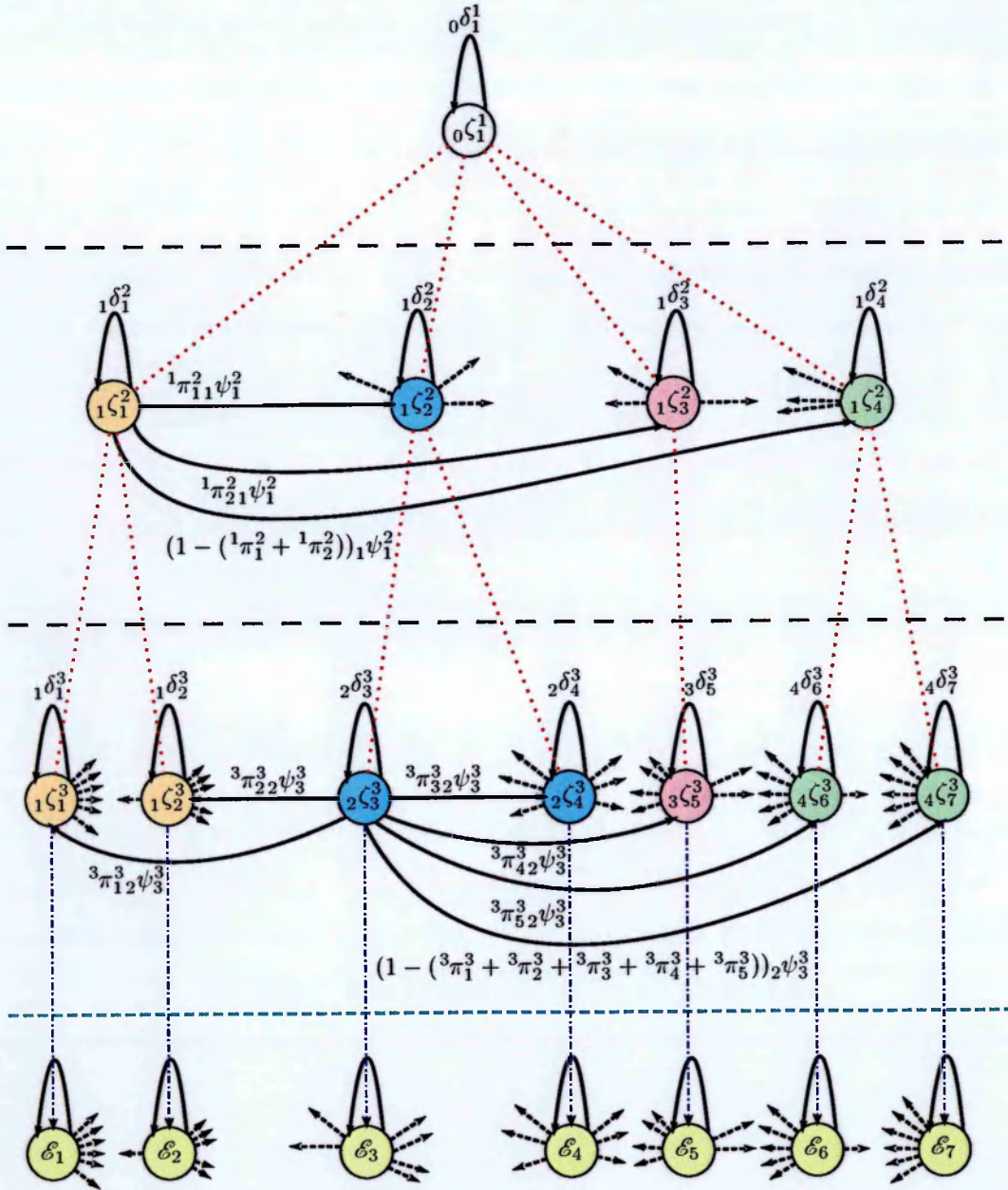
Note, the main difference between CLHBC and MLCTP lies in the construction of the rule structure. As such CLHBC is label based with probabilities computed using observations directly whereas in MLCTP this is achieved using recursive CRPs per state per level until truncation is reached, and SB-construction is then used for calculating topological transition probabilities. The hierarchical probability injection step for calculating augmented likelihoods of state transitions for both of these methods is similar.

### Worked Example

In this section, we initially present the topology construction process for a three-level ( $H = 3$ ), topological structure i.e. where people visit takeaways in three cities. We also present the topological states' transition matrices generation phase where people are recommended to visit takeaways in the same city including re-visiting the same takeaway (representing transition to other states and a self-transition, respectively).

Following is the step-wise instantiation of the process:














**Figure 5.3:** Multi-Level Chinese Takeaway Process; Example topology with  $H = 3$  and  $\mathcal{G} = 7$  (i.e.  $\mathcal{O} = 1, 2, 3, \dots, 7$ )

### Step 1

We begin the process by assuming that the top most level ( $h = 1$ ) has a single state and that the self-transition probability for this state is unity:

$$P(0\zeta_1^1 | 0\zeta_1^1) = 0\delta_1^1 = 1 \quad (5.16)$$

| Legend for Figure 5.3   |                                      |
|---|--------------------------------------|
|  | Initial state                        |
|  | Level delineations                   |
|  | State types                          |
|  | Parent-child state associations      |
|  | Illustrated transition probability   |
|  | Unillustrated transition probability |
|  | Bijective delineation                |
|  | Bijective mapping                    |
|  | Bijectively mapped observation       |

Where  ${}_x\zeta_y^h$  represents state  $\zeta$  number  $y$ , under the mother state number  $x$ . For the top level, represented by Equation 5.16,  $x = 0$  representing no mother node,  $y = 1$  the one and only state number and  $h = 1$  (the top level index). Intuitively, this level represents city number,  $h = 1$ , with one take away,  $y = 1$  and has no prior recommendations,  $x = 0$ .

${}_0\delta_1^1$  represents the self-transition probability for the top most state shown in Figure 5.3. The state transition matrix for this level is a single number representing the self-transition.

### Step 2

In this step, the first instantiation of the MLCTP takes place as formulated in Equation 5.9 and 5.12. The number of resultant generations (takeaways) represents the number of states under the mother node from Step 1. Analogically, people who have visited the takeaway  ${}_x\zeta_y^h$  in city,  $h = 1$ , are recommended takeaways in city,  $h = 2$  (Note, as we shall see, it is not always the case that  $h = x + 1$ ).

$\mathcal{C}^2$ , representing the number of takeaways in the second city ( $h = 2$ ) is in our example 4 (i.e.  $y = [1, 2, 3, 4]$ ) and their self-transition probability  ${}_x\delta_y^2$  i.e. the probability of



visiting the same takeaway, the next day, is extracted from Equation 5.9.

The remaining probability is broken  $\mathcal{C}^2 - 2$  times (i.e. 2 times in this example), so as to generate transitions to all the other states i.e. the probability of visiting another takeaway, the next day, having visited the current takeaway. This is achieved via the stick-breaking construction of Equation 5.12 and 5.13 and is repeated for all the 4 takeaways.

At this level, ( $h = 2$ ), we have that  $x = 1$ , representing the same mother node for all the states and  $y = [1, 2, 3, \dots, \mathcal{C}^2]$ .

Figure 5.3 shows all the possible transitions for the first state at the second level, i.e.  ${}_1\zeta_1^2$  :

$$P({}_1\zeta_1^2 | {}_1\zeta_1^2) = {}_1\delta_1^2 \quad (5.17)$$

$$P({}_1\zeta_2^2 | {}_1\zeta_1^2) = {}^1\pi_1^2 \cdot (1 - {}_1\delta_1^2) = {}^1\pi_1^2 \cdot {}_1\psi_1^2 \quad (5.18)$$

$$P({}_1\zeta_3^2 | {}_1\zeta_1^2) = {}^1\pi_2^2 \cdot {}_1\psi_1^2 \quad (5.19)$$

$$P({}_1\zeta_4^2 | {}_1\zeta_1^2) = (1 - ({}^1\pi_1^2 + {}^1\pi_2^2)) \cdot {}_1\psi_1^2 \quad (5.20)$$

Similarly, these transition probabilities are calculated for  ${}_1\zeta_2^2$ ,  ${}_1\zeta_3^2$  and  ${}_1\zeta_4^2$ . The resultant state transition matrix for this example is a  $4 \times 4$  matrix with 16 possible transitions.

### Step 3

Similar to Step 2, we generate new states via another instantiation of the MLCTP under each state at level 2. We do this for all  $\mathcal{C}^2$ -states generated in Step 2. The number of generations represents the number of states under each mother node. Analogically, people who have visited takeaways in city 2, are recommended to visit similar type of takeaways in city 3.

$x$  at this level is the total number of states indexed by  $y$  in the previous level in Step 2 representing the now-mother states while the length of  $y$  is determined by each instantiation of MLCTP for every  $x$ . The number of MLCTP instantiations is equal to the length of  $x$ .

Thus, at this level,  $h = 3$ ,  $x = [1, 2, 3, \dots, \mathcal{C}^2]$ , and  $y = [1, 2, 3, \dots, \mathcal{C}^3]$ , which is constituted via the vector of sets  $y'$ ,

$$y' = \begin{bmatrix} \{1, 2, 3, \dots, \mathcal{C}'_1\} \\ \{1, 2, 3, \dots, \mathcal{C}'_2\} \\ \{1, 2, 3, \dots, \mathcal{C}'_3\} \\ \vdots \\ \{1, 2, 3, \dots, \mathcal{C}'_{\mathcal{C}^2}\} \end{bmatrix} \quad (5.21)$$

such that,  $\sum_{r=1}^{\mathcal{C}^2} \mathcal{C}'_r = \mathcal{C}^3$ , representing the total number of states at this level.

In our example, at the lowest level,  $x = [1, 2, 3, 4]$ ,  $y = [1, 2, 3, 4, 5, 6, 7]$ , constituted via the vector of sets  $y'$ ,

$$y' = \begin{bmatrix} \{1, 2\} \\ \{1, 2\} \\ \{1\} \\ \{1, 2\} \end{bmatrix} \quad (5.22)$$

Similar to Step 2, the remaining probability is broken  $\mathcal{C}^3 - 2$  times (i.e. 5 times in this example), to generate transitions to all the other states i.e. the probability of visiting another takeaway, the next day, having visited the current takeaway. This is achieved via stick-breaking construction of Equation 5.12 and is repeated for all the 7 takeaways.

Figure 5.3 shows all the possible transitions for the third state at the third level under the second mother state of level 2, i.e.  ${}_2\zeta_3^3$ :

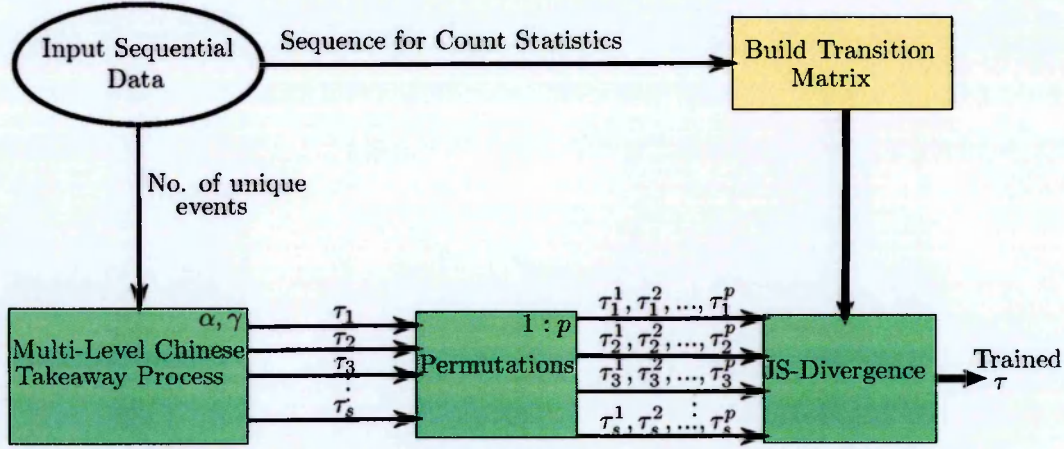


Figure 5.4: Block Diagram for Multi-Level Chinese Takeaway Process

$$P(2\zeta_3^3 | 2\zeta_3^3) = 2\delta_3^3 \quad (5.23)$$

$$P(1\zeta_1^3 | 2\zeta_3^3) = {}^3\pi_1^3 \cdot (1 - 2\delta_3^3) = {}^3\pi_1^3 \cdot 2\psi_3^3 \quad (5.24)$$

$$P(1\zeta_2^3 | 2\zeta_3^3) = {}^3\pi_2^3 \cdot 2\psi_3^3 \quad (5.25)$$

$$P(2\zeta_4^3 | 2\zeta_3^3) = {}^3\pi_3^3 \cdot 2\psi_3^3 \quad (5.26)$$

$$P(3\zeta_5^3 | 2\zeta_3^3) = {}^3\pi_4^3 \cdot 2\psi_3^3 \quad (5.27)$$

$$P(4\zeta_6^3 | 2\zeta_3^3) = {}^3\pi_5^3 \cdot 2\psi_3^3 \quad (5.28)$$

$$P(4\zeta_7^3 | 2\zeta_3^3) = (1 - ({}^3\pi_1^3 + {}^3\pi_2^3 + {}^3\pi_3^3 + {}^3\pi_4^3 + {}^3\pi_5^3)) \cdot 2\psi_3^3 \quad (5.29)$$

### 5.3.3 Induction Protocol

Figure 5.4 shows the block diagram of the experimental protocol for the Multi-Level Chinese Takeaway Process showing the training process. MLCTP is a stochastic process and to counter the issue of stochastic variations we first generate  $R$  topologies i.e. we execute the process  $R$  times given the hyper-parameters  $\alpha$  and  $\gamma$ . The total number of selected topologies according to the truncation parameter  $\mathcal{S}$  (applied such that when the exact number of leaf states is achieved the process stops and emits a topological structure), is  $s$ , where  $s \leq R$ . These  $s$  topologies are represented as transition matrices computed via Equation 5.14 and each transition matrix goes through a selection process for the best fit as the rule defining topology. This is achieved via measuring the distance

between the training matrix (using the count statistics of the training data) and the MLCTP-generated topological transition matrix.

MLCTP is a stochastic and unlabeled process, and thus a topology generated given a set of hyper-parameters does not necessarily correlate with the original training transition matrix. To better sample the topological state space, we generate many random permutations at the bijective mapping level i.e. observations indicated in Figure 5.4 (we employ *random* permutations to reduce the computation time). If  $b$  is the state-space defined by the number of leaf states and  $p$  is the number of random permutations then  $p \leq b!$ . Each topology,  $\tau_I$  ( $I = 1, 2, 3, \dots, s$ ), is expanded to  $p$  random permutations within the state-space, where  $\tau_1$  represents the first selected topology matrix and  $\tau_s$  represents the last selected topology matrix. We thus have an array:

$$\left\{ \begin{array}{c} \tau_1^1, \tau_1^2, \tau_1^3, \dots, \tau_1^p \\ \tau_2^1, \tau_2^2, \tau_2^3, \dots, \tau_2^p \\ \tau_3^1, \tau_3^2, \tau_3^3, \dots, \tau_3^p \\ \vdots \\ \tau_s^1, \tau_s^2, \tau_s^3, \dots, \tau_s^p \end{array} \right\} \quad (5.30)$$

Each of these topological transition matrices (total  $s \times p$ ) are then compared against the flat transition matrix built from the training sequence of events. This comparison is performed via the Jensen-Shannon Divergence.

### Jensen-Shannon Divergence

We use Jensen-Shannon Divergence [121] to measure the divergence between two probability distributions i.e. the output of permutations block and the flat Markov Model block in Figure 5.4. JS-Divergence is based on the Kullback-Leibler divergence and is defined as the average relative entropy of the source distributions to the entropy of the average distribution. Equation 5.31 represents the metric  $Y$  employed in Figure 5.4 for MLCTP's topological transition matrices  $J_Z$  (where  $Z = 1, 2, 3, \dots, s \times p$ ) and the training transition matrix  $J_{tr}$ .

$$Y(J_{tr}, J_Z) = \frac{1}{2}(KL(J_{tr} \parallel K) + KL(J_Z \parallel K)) \quad (5.31)$$

where  $K$  is the average distribution of the two sources i.e.,

$$K = \frac{1}{2}(J_{tr} + J_Z) \quad (5.32)$$

and the KL-divergence can be defined between two vectors,  $M_1$  and  $M_2$  as,

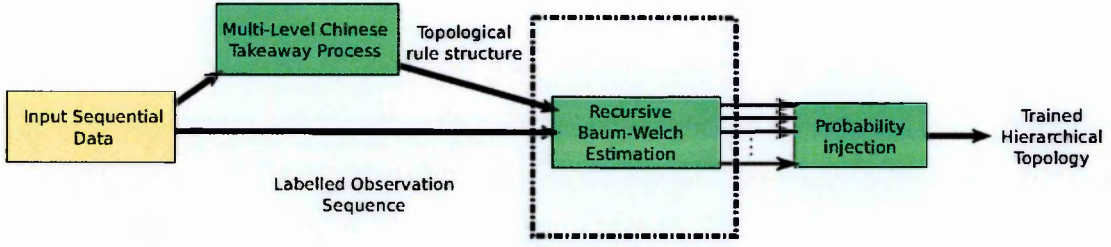
$$KL(M_1 \parallel M_2) = \sum_i M_1(i) \ln \frac{M_1(i)}{M_2(i)} \quad (5.33)$$

Each topological transition matrix ( $J_Z$ ) from the permutations block is compared against the training transition matrix  $J_{tr}$  and the closest topological structure (i.e. one with the smallest  $Y$  metric against the training matrix) is taken as the learned hierarchical topology with respect to the input training sequence of events. This trained hierarchical topology is used in the following experimental investigation (of Section 5.5) for predicting future events based on the input sequence.

## 5.4 Hybrid Models

### 5.4.1 Multi-Level Chinese Takeaway Process with Recursive Baum-Welch Estimated Hidden State Transitions (MLCTP-BW)

In addition to the above label-based and generative methods, we also propose a pair of hybridized methods suitable for stochastic inference of rule structures in sport videos. The first hybrid model extracts hierarchical structures using MLCTP's topological state generation process shown in Section 5.3.2. The topological state transition matrix generation phase is ignored in this model, so the hierarchical structure output from MLCTP is effectively just the arrangement of nodes, connected in a hierarchy. These topologies are built top-down and we similarly select topologies based on MLCTP's truncation parameter  $\mathcal{G}$  (defined as the number of differentiated states in the input sequential data).



**Figure 5.5:** Multi-Level Chinese Takeaway Process with Baum-Welch Hidden State Transition Estimation

In order to re-calculate transition probabilities on the topological edges for the hybrid model, we first compute the count statistics of the sequence of event transitions i.e. the flat Markov Model of Equation 5.7 that calculates the observed transition probabilities. Leaf states are thus mapped bijectively to observations, effectively leaving the number of states in the higher (i.e. non-observation or hidden) levels to be defined by the MLCTP-generated topological rule structures. This structure is then used to estimate a set of state transition probabilities at each level via *recursive Baum-Welch estimation* (Figure 5.5 has the block diagram representing the training process using this method).

We can characterize the model via the following notation:

The MLCTP emitted topological structure has  $h = 1, \dots, H$  levels and for each pair of levels, we can specify a set of HMM parameters;

$$\lambda_h = \{a_{ij}^h, e_i^h(\cdot), \eta^h(i)\} \quad (5.34)$$

where  $\eta^h(i)$  is the initial distribution of states for each level defined by MLCTP,  $e_i^h(\cdot)$  is the emission probability for each level i.e. the probability of state  $i$  at level  $h$  emitting a symbol at level  $h + 1$ , while the transition probability of a state transiting from  $i$  to  $j$  for level  $h$  is  $a_{ij}^h$ :

$$a_{ij}^h = P(Q_t^h = j | Q_{t-1}^h = i) \text{ and } \sum_{i=1}^{\mathcal{G}^h} a_{ij} = 1 \quad \forall j \quad (5.35)$$

(where  $Q_t^h$  is the current [hidden] state of a temporal sequence as represented at the hierarchical level  $h$ ).

The input sequence of labeled data is thus the observed sequence from which we obtain the parameters of the model via Maximum Likelihood Estimation. Utilizing the MLCTP-generated hierarchical topology, Baum-Welch algorithm is hence employed recursively to obtain the model parameters when the state path per level is unknown. Thus given a level-based sequence of observations  $\{Q_t^{h+1}\}$  for a given number of hidden states defined by MLCTP,  $\mathcal{E}^h$ , we compute  $\lambda_h = \{a_{ij}^h, e_i^h(\cdot), \eta^h(i)\}$ . The parameters that maximize the likelihood of the input data are thus chosen at each individual level.

The recursive Baum-Welch state transition estimation is thus defined:

$$A_{ij}^h = \text{BaumWelch}(\mathcal{E}^h, \{Q_t^{h+1}\}, \lambda_h) \quad (5.36)$$

or more specifically,

$$A_{ij}^h = \frac{1}{P(Q_t^{h+1}|\lambda_h)} \sum_t F(t, i) a_{ij}^h e_j^h(Q_{t+1}^{h+1}) B(t+1, j) \quad (5.37)$$

$$(i, j = 1, 2, \dots, \mathcal{E}^h),$$

Here,  $A_{ij}^h$  is the estimated state-transition probability of state  $i$  to  $j$  at level  $h$ ;  $F(t, i)$  represents the probability of the model emitting symbols,  $Q_1^{h+1} \dots Q_T^{h+1}$ , when in state  $i$  at time  $t$ , obtained using the Forward algorithm.  $a_{ij}^h$  and  $e_j^h(Q_{t+1}^{h+1})$  respectively represent the probability of transition from state  $i$  to  $j$  and emitting the  $t+1$ st emission symbol at level  $h+1$  (both arbitrarily instantiated and recursively updated). The Backward algorithm computes  $B(t+1, j)$  which is the probability of the model emitting the remaining sequence if the model is in state  $j$  at time  $t+1$ .

Thus, in this model, estimated hidden state transitions act as the observation level for the estimation of the next highest level hidden state transitions in the hierarchy and so on. After estimating state transition probabilities, a state sequence is generated i.e.  $\{Q_t^h\}$  which is used as input *observations* for the next level of MLCTP's hierarchy.

Finally, after computing state transition probabilities for each level, we perform the top-down hierarchical probability injection step (Equation 5.14) to obtain the learned



augmented likelihood of events for MLCTP-generated topological structure with recursive Baum-Welch estimated state transition probabilities (MLCTP-BW).

The MLCTP-BW hybrid could be considered as the methodology that is conceptually closest to the standard hHMM of [49] where the hierarchical topology is fixed based on the hierarchy established using MLCTP.

#### 5.4.2 Multi-Level Chinese Takeaway Process with Cartesian Product Label-Based Hierarchical Bottom-up Clustering Computed State Transitions (MLCTP-CLHBC)

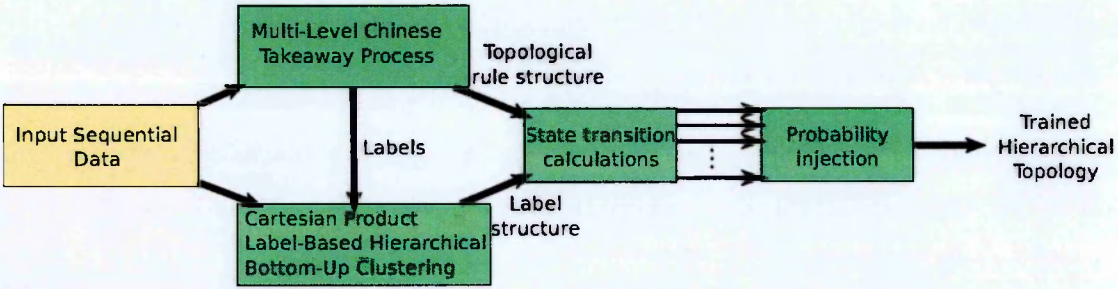
The second hybrid model variant similarly extracts the hierarchical topologies from MLCTP's topological state generation process. However, transition probabilities at the edges are then computed using the Cartesian Product Label-Based Hierarchical Bottom-Up Clustering (CLHBC) method.

In this model, MLCTP determines the number of levels (which in CLHBC is determined by the number of sub-labels defined in Section 5.2). Each labeled event is replaced by an arbitrary label, comprised of  $z$  types of labels,  $\Omega_z$ . In this hybrid,  $z$  for CLHBC is determined by MLCTP. Intuitively, bigger  $z$  implies a deeper hierarchy.

Thus, the input training sequential data provides the event labels to be bijectively mapped into the observation level of MLCTP generated hierarchy (e.g.  $Q_t^1 \rightarrow SF$  i.e. observation  $Q_t^1$  is associated with event label *Serve Far* as shown in Figure 5.3, where the leaf nodes are mapped into events  $\mathcal{E}_O$ ).

Given the number of levels in the MLCTP-generated topological structure ( $z = 3$  in Figure 5.3), event labels are then replaced with the Cartesian products of  $z$  arbitrary labels i.e.  $\Omega_z$  where  $z = 1, 2, 3$  such that the sub-label factors give rise to a hierarchy equivalent to that defined by MLCTP-generated topological structure. We thus reverse engineer Figure 5.3, where each event is re-represented with three (as  $z = 3$ ) labels e.g.  $\omega_1^1, \omega_3^2, \omega_5^3$ , with labeling associated with the observed states such that the common sequential factors result in the hierarchy generated using MLCTP. Following this re-association phase, the training event sequences are re-generated using the new label-





**Figure 5.6:** Multi-Level Chinese Takeaway Process - Cartesian Product Label-Based Hierarchical Bottom-up Clustering

structure and the CLHBC process executed resulting in state-transition probabilities at each level of the hierarchy generated by MLCTP.

Transition between states is thus governed by the input data at every level; however inter level associations are determined by MLCTP. The method thus populates the transition likelihoods ‘bottom up’ according to the MLCTP template. Figure 5.6 shows the block diagram for this hybrid model in which the input sequential data’s original labels are replaced with MLCTP-defined arbitrary labels. The hierarchical structure output from MLCTP is combined with this new label hierarchy and used to compute state-transition matrices for each level.

The trained augmented likelihood of events for the MLCTP-generated topological hierarchy and CLHBC formulated label structure is computed in a similar fashion to previous methods via *probability injection* block of Figure 5.6. Experimental results for MLCTP-CLHBC are shown in Section 5.5.

## 5.5 Experimental Results and Discussions

In this section, we evaluate the performance of all the four proposed variants of the novel hierarchical HMM strategy using six different datasets shown in Tables 5.2 and 5.4. In case of the Badminton dataset, we train the models using 77 play-shots (i.e. collections of sequences starting with the event *serve* and ending with a point-awarding event) and test using the remaining 20 play-shots. The number of unique events for Badminton is 8 (see Table 5.1). Similarly, datasets from other domains with varying

Table 5.2: Sports datasets with source information

| Source Information |            |        |           |                  |      |
|--------------------|------------|--------|-----------|------------------|------|
| Source label       | Sport Type | Gender | Game type | Competition      | Year |
| BMSB08             | Badminton  | Mens   | Singles   | Beijing Olympics | 2008 |
| TWSA03             | Tennis     | Womens | Singles   | Australian Open  | 2003 |
| TWSA03 [89]        | Tennis     | Womens | Singles   | Australian Open  | 2003 |
| TMSA03             | Tennis     | Mens   | Singles   | Australian Open  | 2003 |

Table 5.3: Sports datasets with the number of samples per event label

| Event Statistics |       |    |    |     |     |     |     |     |     |      |      |  |  |  |
|------------------|-------|----|----|-----|-----|-----|-----|-----|-----|------|------|--|--|--|
| Source           | Total | SF | SN | HF  | HN  | BIN | BIF | BON | BOF | BOFS | BONS |  |  |  |
| BMSB08           | 644   | 59 | 38 | 222 | 228 | 41  | 17  | 22  | 17  | -    | -    |  |  |  |
| TWSA03           | 532   | 32 | 42 | 98  | 92  | 104 | 120 | 11  | 11  | 13   | 9    |  |  |  |
| TWSA03 [89]      | 293   | 30 | 38 | 35  | 42  | 56  | 50  | 9   | 12  | 12   | 9    |  |  |  |
| TMSA03           | 1122  | 69 | 53 | 221 | 224 | 234 | 265 | 12  | 22  | 12   | 10   |  |  |  |

Table 5.4: Datasets Description

|                                 | Train Source        | Test Source         | Area          | Train Size      | Test Size           | No. of Unique Events |
|---------------------------------|---------------------|---------------------|---------------|-----------------|---------------------|----------------------|
| Badminton Dataset               | BMSO08              | BMSO08              | Sports        | 77 Play-shots   | 20 Play-shots       | 8                    |
| Tennis Dataset                  | TWSA03              | TMSA03              | Sports        | 74 Play-shots   | 122 Play-shots      | 10                   |
| Tennis (Annotation System [89]) | TWSA03              | TMSA03              | Sports        | 68 Play-shots   | 122 Play-shots      | 10                   |
| Human Activity Dataset          | UCI Repository [72] | UCI Repository [72] | Life          | 500 Events      | 275 Events          | 11                   |
| Human Driving Dataset           | EU DIPLECS [154]    | EU DIPLECS [154]    | Highway Rules | 135 Events      | 80 Events           | 10                   |
| MSNBC.com (Website Dataset)     | UCI Repository [21] | UCI Repository [21] | Web           | 1 → 8000 clicks | 8001 → 10000 Clicks | 11                   |

Table 5.5: Mean Accuracy

|                                 | Flat          | CLHBC         | MLCTP         | MLCTP-BW      | MLCTP-CLHBC   |
|---------------------------------|---------------|---------------|---------------|---------------|---------------|
| Badminton Dataset               | 89.04% ± 1.01 | 91.90% ± 0.64 | 92.94% ± 0.27 | 92.56% ± 0.49 | 93.23% ± 0.08 |
| Tennis Dataset                  | 63.80% ± 1.31 | 69.27% ± 1.16 | 73.28% ± 1.16 | 70.07% ± 0.62 | 73.43% ± 0.97 |
| Tennis (Annotation System [89]) | 52.73% ± 2.18 | 62.20% ± 2.24 | 73.44% ± 0.54 | 66.33% ± 0.21 | 73.80% ± 1.09 |
| Human Activity Dataset          | 81.34% ± 1.67 | 82.34% ± 1.57 | 84.27% ± 1.14 | 83.26% ± 0.77 | 84.56% ± 0.78 |
| Human Driving Dataset           | 55.77% ± 5.25 | 58.30% ± 3.34 | 66.75% ± 1.23 | 65.71% ± 0.93 | 67.21% ± 3.07 |
| MSNBC.com (Website Dataset)     | 44.21% ± 1.55 | 53.53% ± 1.48 | 60.00% ± 0.06 | 57.56% ± 1.61 | 64.17% ± 0.64 |
| Average Performance Gain        | -             | 5.11%         | 10.63%        | 8.10%         | 11.59%        |

test and training protocols are also employed, details of which are shown in Table 5.4. For experimental evaluation of the method we measure the prediction accuracy for the next event given all of the previous events. It is shown in Figure 5.8 for the Tennis dataset (extracted using automated sports video annotation system [89] introduced in Chapter 3) e.g. the model CLHBC correctly predicts the next event 74.52% of the time, if the current event is *HF* (see Table 5.1) and so on.

Figure 5.7 shows the comparative mean accuracies for the Badminton dataset using all of the methods employed namely, the Flat Markov Model (Flat MM), the Cartesian Product Label-Based Hierarchical Bottom-Up Clustering (CLHBC), the Multi-Level Chinese Takeaway Process (MLCTP), Hybrid I (MLCTP-BW) and Hybrid II (MLCTP-CLHBC). Mean prediction accuracies for all the methods applied to all of the datasets with individual standard deviations are shown in Table 5.5. Associated mean performance gains with respect to the baseline approach are also shown. It can be seen that all of the proposed hierarchical HMM generating methodologies demonstrate improvement relative to the Flat Markov Model.

Additionally, confusion matrices for the predicted events are also presented for all the methods applied to the Badminton dataset. Thus we see, for example, in Figure 5.9c, that *BON* (Bounce Out Near, see Table 5.1) is 50% of the time correctly predicted, while 50% of the time incorrectly predicted as *HN* (Hit Near).

As may be seen in Figure 5.7 and Table 5.5 optimal performance for all of the datasets is achieved using the hybrid model MLCTP-CLHBC (of Section 5.4.2). MLCTP-CLHBC hybrid leverages MLCTP's topological rule structure and consequently the label-based CLHBC to construct a rule model that is relatively more accurate. The average performance gain achieved using MLCTP-CLHBC compared with the flat Markov Model is 11.59% resulting in a significant improvement.

Relative performance gains in the context of a particular dataset vary depending upon the complexity of the data. It can be observed that in the case of more complex datasets – such as the MSNbc.com and the tennis [88] datasets – the average performance gains achieved by MLCTP-CLHBC are around 20%.

In a different setting (introduced in Chapter 3 and explained in [86, 88]) high-level



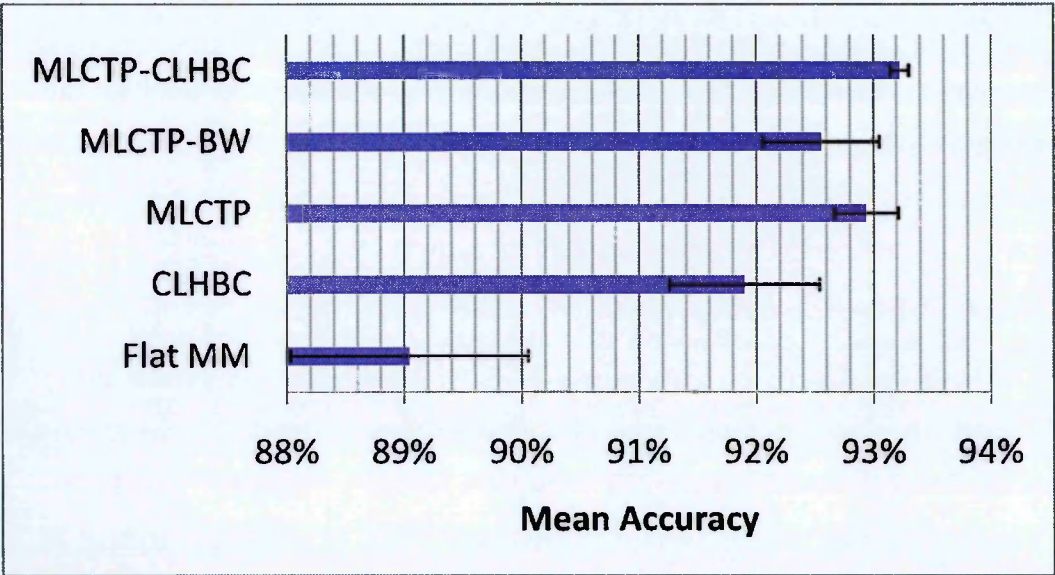


Figure 5.7: Badminton Dataset - Comparative mean prediction with standard deviation

reasoning, in terms of correctly awarded points, is performed using the hard-wired HMM based on Figure 3.6. Correct point recognition rates reported using TWSA03 and TMSA03 (see Table 5.2) were 87.5% and 73.75% respectively.

It is crucial to highlight again that HMM used in the work of [86, 88] requires the number of states to be fixed heuristically based on the exact rule of the game. A different domain cannot be directly introduced without manually altering the HMM topology as there is no capability in this framework to learn the rule model. Our generalized rule induction mechanism on the other hand is intrinsically adaptive, as evidenced by our demonstration of the approach in domains other than tennis.

5.6 Summary

In this chapter, we proposed four variants of the novel hierarchical Hidden Markov Model strategy for rule induction and applied them to the problem of automated sports video annotation. We firstly introduced a Cartesian Product Label-Based Hierarchical Bottom-Up Clustering method that employs the latent structure in the labels used to annotate videos. Labels are thus employed to build hierarchical structures based on various Cartesian Product based combinations of sub-labels such that a hierarchical

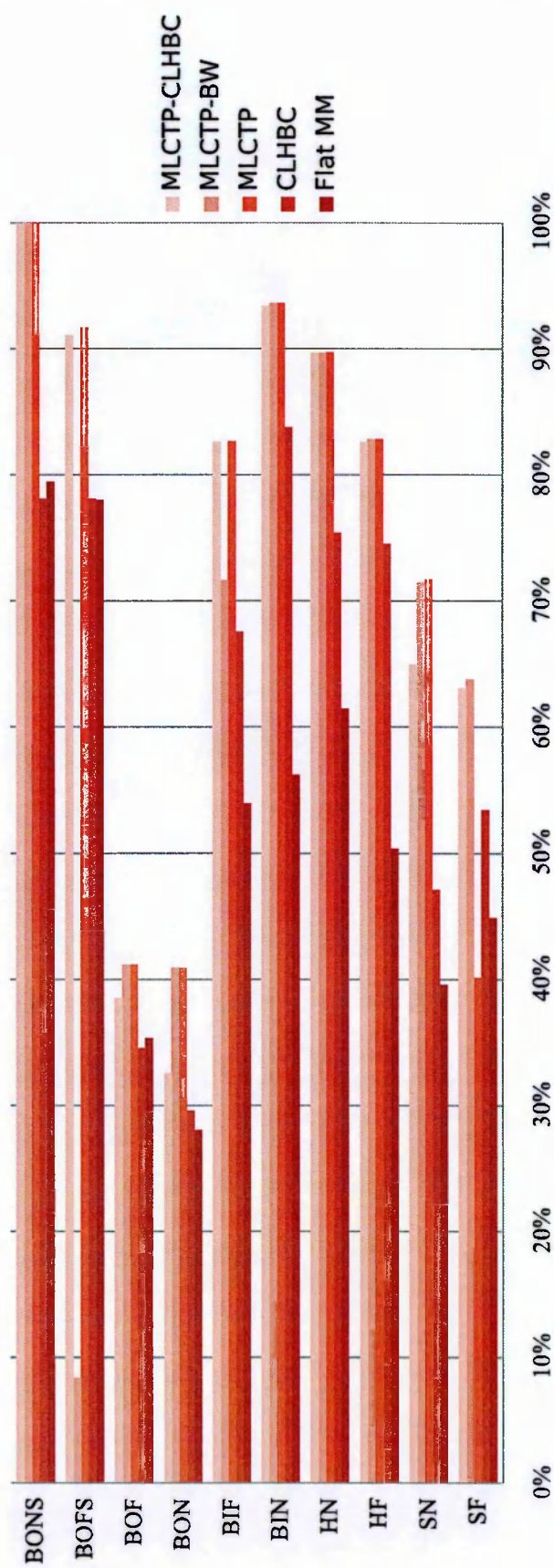
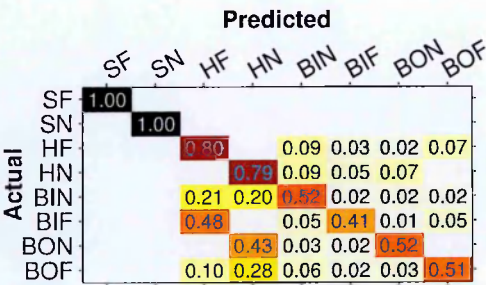
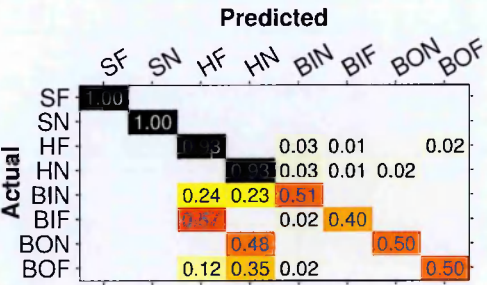


Figure 5.8: Tennis Dataset (Video Annotation System) - Comparative individual event prediction accuracies for all of the five methods.

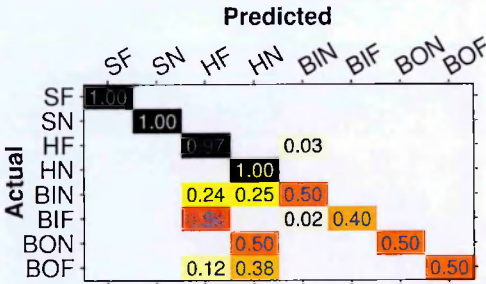




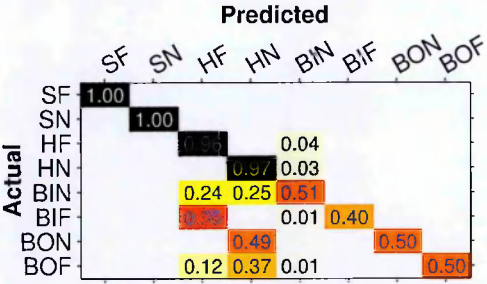
(a) Flat Markov Model



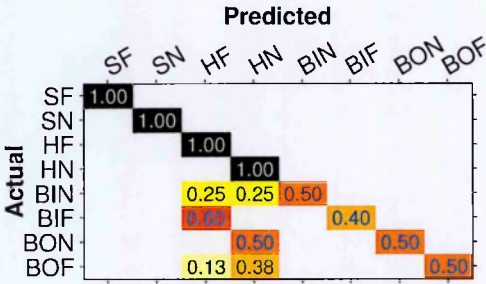
(b) Cartesian Product Label based Hierarchical Bottom-up Clustering



(c) Multi-Level Chinese Takeaway Process



(d) Hybrid - MLDP-BW



(e) Hybrid - MLDP-CLHBC

Figure 5.9: Badminton Dataset - Confusion matrices for all of the five methods

---

HMM of common repeated event structures is established (and which is used to evaluate the predictive capability of the method). The second proposed variant, the Multi-Level Chinese Takeaway Process, is based on the Chinese Restaurant Process with tables replaced by takeaways which may be re-visited within different cities representing levels in the hierarchy. This is a stochastic process, with many hierarchies generated for a given set of hyper-parameters, such that a distance measure (JS-Divergence) is employed to infer the highest likelihood stochastic rule structure.

We also introduced two hybrid variants namely MLCTP-BW and MLCTP-CLHBC, that leverage the stochasticity of MLCTP (whereby various latent hierarchical structures are produced), in conjunction with the label sequence to give a composite ‘top-down’ MLCTP-driven (topologically) and ‘bottom-up’ data-driven approach to hierarchical HMM inference. All of these methods finally generate finite intermediate-depth hierarchical HMMs that are well-suited to calculating the likelihood of event transitions taking place within sport video sequences typically governed by analogous hierarchical rule structures involving e.g. *matches*, *sets*, *points*, etc.

We conclude that leveraging the label information contained within sequential data (especially sports video sequences) in conjunction with our novel Multi-Level Chinese Takeaway Process provides a previously unexploited opportunity for rule-induction. Comparative prediction results for all of the proposed methods are shown relative to the flat Markov Model, with all of the hierarchical methods shown to perform better (with the most optimal method being the MLCTP-CLHBC hybrid).

In the context of an automated video annotation system, the rule induction framework thus provides a robust context analysis module in which rules are inferred from observations and predictions are made that can serve as logical priors on detections. Such a framework can be employed for tackling various problems beside prediction generation. In particular, it can address the issue of *anomaly detection*; when a new domain is introduced to the system, anomalous events (as opposed to outliers and errors) can be detected using the rule hierarchy triggering the domain change. In the context of an automated video annotation system, this may require switching the knowledge base by abandoning continuous adaptive learning and replacing it with a new learning process.



# Chapter 6

## Summary and Future Work

### 6.1 Thesis Summary

This thesis investigated various methodologies in the context of automated sports video annotation specifically aimed towards the goal of developing a generalized contextual analysis system. We experimentally demonstrated all of the proposed novel methodologies for problems like Anomaly Detection and Rectification which are important for a generalized adaptive system with changing input domains. Additionally, we also proposed new methods for rule-induction i.e. a generic framework that is able to contextually represent the input domain in terms of a hierarchical rule structure.

We commenced our discussion in Chapter 1 with the introduction to this thesis, comprising the motivation of this research work (potential research applications), problem statement, related task complexities and the contributions. The main objective of this thesis was to present state-of-the-art methodologies for problems like anomaly detection, anomaly rectification and stochastic rule-induction in the context of automated sports video annotation. This chapter concluded with a short summary comprising a systematic breakdown of the principal methodological contributions.

In Chapter 2, we presented a detailed literature review of key concepts related to automated sports video annotation, as well as, techniques deemed relevant to our research problems of anomaly detection, rule adaptation and rule induction. This provided

us with baseline knowledge and understanding to carry out a focused analysis of our methodologies for developing a generalized contextual analysis system and motivating our subsequent research problem.

We also very briefly, introduced the automated tennis video annotation system of [89] in Chapter 3, which constitutes a vehicle and experimental test bed for our work to design a generalized high-level contextual analysis system. In this chapter, we also presented other computer vision systems for deriving experimental datasets which we employ for our novel methodologies. This includes two ground-truth annotation systems for annotating tennis, badminton and human driving behavior. We also, very briefly, introduced two other datasets from the UCI repository i.e. website data and human activity localization data, which we employ for rule-induction methodologies. In Chapter 4 and Chapter 5, we presented our principal methodological contributions comprising the proposed solutions to our research problems in the context of automated sports video annotation.

In Chapter 4, we first introduced the problem of anomaly detection for changing domains in the context of automated sports video annotation and for knowledge transfer. We presented a novel lattice-based HMM induction strategy for detecting anomalies tuned for court-game environments that maps game-play states into a court lattice specifically when a sport domain is switched from singles tennis to doubles tennis. Furthermore, we also present the anomaly rectification strategy that changes the definition of events in the lattice-space for doubles tennis and hence rectifying all of the detected anomalies, i.e. successfully adapting to a change of rule structures going from singles to doubles tennis. We demonstrated the performance and ability of the methods in real and simulated tennis singles and doubles games.

In addition to the lattice-based anomaly detection and rectification method, we also presented another approach to address the problem of anomaly detection that is based on the disparity between the low-level vision based classifiers and the high-level contextual classifiers. A convex-hulling approach is then presented to rectify anomalies in this chapter. We demonstrated how the concept of anomaly detection may be extended so as to transfer learning from an initially known rule-governed domain to another via the

---

redefinition of the main play area in terms of the convex hull of the detected anomalies. We demonstrate experimental results using real tennis singles and doubles games.

In Chapter 5, we presented four novel methodologies that are capable of generating hierarchical HMMs for rule-induction primarily in the context of automated sports video annotation. These novel methods include the Multi-Level Chinese Takeaway Process (MLCTP) based on the classical Chinese Restaurant Process and the Cartesian Product Label-based Hierarchical Bottom-up Clustering (CLHBC) method that exploits information within the label structures. Furthermore, we also propose two hybrid methodologies, i.e. MLCTP-BW (MLCTP defined hierarchical topology with a recursive Baum-Welch estimated hierarchical state transition probabilities) and MLCTP-CLHBC (MLCTP defined hierarchical structures with CLHBC computed label associations). Our results showed significant improvement by comparison against the flat Markov model while the optimal performance is obtained using MLCTP-CLHBC hybrid. We also showed that the methods proposed are generalizable to other rule-based environments such as badminton, human driving, website clicking behavior and human activity localization.

## 6.2 Future Work

We have proposed and experimentally demonstrated methodologies for solving various problems required for a generalized adaptive system that is able to autonomously annotate videos. In order to achieve the goal of understanding videos, the current sport related experimental datasets are limited by data-quality i.e. most of the videos are either too old and/or improperly recorded. To overcome this difficulty, novel methods needs to be employed to contextually cater for missing (or immeasurable) video frames.

Our rule-induction methodologies, in this thesis, have been extensively evaluated in terms of event predictions. They can also be used to predictively and retrospectively identify unobserved video frames. Following this direction of research will introduce various implementation and methodological issues such as the need to replace the current Markovian structure. One method of dealing with this problem could be via analyzing associated rule-grammars.

In order to expand these models for practical implementation in future generalized video annotation systems, the proof-of-concept evaluations will need to be expanded to other domains comprising other sports such as cricket, football, table tennis and non-sporting domains such as characterizing surveillance footage, recorded meetings and lectures etc.

Rule induction framework can also be employed to address the problem of transferring knowledge from one domain to another; this can be achieved via analyzing various levels of the established rule hierarchies representing different levels of abstractions such that in a new (and related) domain, contextual inferences are transferred i.e. minimizing the need for re-training.

Additionally, these methodologies could also be effective in an online annotation system environment with faster processing abilities. Implementing the proposed rule-induction methods with speedy processing priorities will help the current sport broadcast in terms of commentating with methodologies learning player behaviors e.g. the types of shots a player is expected to repeat throughout game-play. With combined annotation history of players, the system will be able to provide other useful information e.g. player behavior not only in the normal games but in the finals etc.

Our rule-induction methods were able to produce rule structures for sport games representing various types of observations in a hierarchy of a limited depth. In other games, this limitation must be relaxed to model a complete rule model which might be present in a deeper and more extensive hierarchical topology.

Furthermore, in the context of broad research, rule-induction methods could also be influenced by human psychological learning behavior with human subjects exposed for the first time to a certain rule-domain. This could be achieved via presenting learning-based questionnaires among several exposures to the test videos and based on that, rule-induction strategies should be tuned for realistic modeling.

Moreover, beyond annotation systems, rule-induction methods can also be employed in other non-video based applications. There is a massive scope for learning gene behaviors for disease investigations, e.g. precisely modeling the likelihood of a human subject carrying a particular disease, given the subject carries a certain gene which can

---

help in avoiding major diseases.

Further modeling human actions (and expanding on what we show with the human localization data), we can explore behaviors resulting in muscle damage. Knowledge of a certain movement causing muscle damage may help in cases of arthritic patients with detectors warning the patients from carrying out risk afflicting actions.

Finally, but by no means least, rule-induction methodologies can also be employed across other scientific applications, such as, prosthetics, industrial and planetary exploration robots, driving safety systems, surgical tools, airline safety systems and hazardous environment analyzing robots. In all of these applications, rule-induction can potentially be used as a warning system if a certain rule-grammar is not followed correctly.

# Bibliography

- [1] Ball tracking and virtual replays for innovative tennis broadcasts. In *ICPR '00: Proceedings of the International Conference on Pattern Recognition*, page 4152, Washington, DC, USA, 2000. IEEE Computer Society.
- [2] Ryan Prescott Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems 23*, pages 19–27, Vancouver, British Columbia, 12/2010 2010.
- [3] M Agyemang, K Barker, and R Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10:6:521–538, 2006.
- [4] David Aldous, Illdar Ibragimov, Jean Jacod, and David Aldous. Exchangeability and related topics. In *cole d't de Probabilits de Saint-Flour XIII 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin / Heidelberg, 1985. 10.1007/BFb0099421.
- [5] I. Almajai, J. Kittler, T. de Campos, W. Christmas, F. Yan, D. Windridge, and A. Khan. Ball event recognition using HMM for automatic tennis annotation. 2010. In press.
- [6] S Ando. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *Proceedings of 7th International Conference on Data Mining*, pages 13–22, 2007.
- [7] J. Anemüller, J.-H. Bach, B. Caputo, M. Havlena, L. Jie, H. Kayser, B. Leibe,

- 
- P. Motlicek, T. Pajdla, M. Pavel, A. Torii, L. V. Gool, A. Zweig, and H. Hermansky. The DIRAC AWEAR audio-visual platform for detection of unexpected and incongruent events. pages pp. 289–293. Proc. International Conference on Multimodal Interaction (ICMI), 2008.
- [8] Jurgen Assfalg, Marco Bertini, Carlo Colombo, and Alberto Del Bimbo. Semantic Annotation of Sports Videos. *IEEE Multimedia*, 9(2):52–60, 2002.
- [9] Jurgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, and Walter Nunziati. Semantic Annotation of Sports Videos: Automatic Highlights Identification. *ELSEVIER Computer Vision and Image Understanding*, 92(2-3):285–305, November-December 2003.
- [10] Y. Bar-Shalom. *Tracking and data association*. Academic Press Professional, Inc., San Diego, CA, USA, 1987.
- [11] V Barnett and T Lewis. *Outliers in statistical data*. John Wiley and sons, 1994.
- [12] S Basu and M Meckesheimer. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11:2:137–154, 2007.
- [13] S. Beucher and F. Meyer. *The Morphological Approach to Segmentation : The Watershed Transformation*, chapter 12, pages 433–481.
- [14] Maria Bielikova, Pavol Navrat, and P. N Avrat. A multilevel knowledge representation of strategies for combining modules, 1997.
- [15] M J Black and A D Jepson. Eigentracking : Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.
- [16] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press, 2004.
- [17] M. Brand. Coupled Hidden Markov Models for modeling interacting processes. Technical Report TR #405, MIT Media Lab Vision and Modeling, June 1996.

- 
- [18] A. D. Brown and G. E. Hinton. Products of Hidden Markov Models. In *Proceedings of Artificial Intelligence and Statistics*, pages 3–11, 2001.
- [19] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky. Combination of strongly and weakly constrained recognizers for reliable detection of oovs. In *Proceedings 33rd International Conference on Acoustics, Speech, and Signal Processing ICASSP*, page 40814084. Available: <http://www.fit.vutbr.cz/research/viewpub.php?id=8494>, 2008.
- [20] Luká Burget, Petr Schwarz, Pavel Matejka, Mirko Hannemann, Ariya Rastrow, Christopher White, Sanjeev Khudanpur, Hynek Hermansky, and Jan Cernock. Combination of strongly and weakly constrained recognizers for reliable detection of oovs. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, page 4, 2008.
- [21] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Visualization of navigation patterns on a web site using model-based clustering. In *Proceedings of the sixth ACM SIGKDD*, pages 280–284, New York, NY, USA, 2000.
- [22] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, Nov. 1986.
- [23] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):pp. 167–174, 1992.
- [24] Graeme S. Chambers, Svetha Venkatesh, Geoff West, and Hung Bui. Segmentation of Intentional Human Gestures for Sports Video Annotation. In *Proceedings of the 2004 IEEE International Multimedia Modelling Conference (MMM)*, volume 1, pages 124–129, Jan 2004.
- [25] V Chandola and A Banerjee and V Kumar. Anomaly detection : A survey. *ACM Computing Surveys*, 41:15:1–15:58, 2009.



- 
- [26] V Chatzigiannakis, S Papavassiliou, M Grammatikou, and B Maglaris. Hierarchical anomaly detection in distributed large-scale sensor networks. In *ISCC'06: Proceedings of the 11th IEEE Symposium on Computers and Communications*, pages 761–767. IEEE Computer Society, Washington, DC, USA, 2006.
- [27] Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. In Thomas B. Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal, editors, *Visual Analysis of Humans*, pages 539–562. Springer London, 2011.
- [28] P A Crook and G Hayes. A robot implementation of a biologically inspired method for novelty detection. In *Proceedings of Towards Intelligent Mobile Robots Conference*, 2001.
- [29] P A Crook, S Marsland, G Hayes, and U Nehmzow. A tale of two filters: Online novelty detection. In *Proceedings of International Conference on Robotics and Automation*, pages 3894–3899, 2002.
- [30] R. Dahyot, A. Kokaram, N. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *ICASSP*, pages 561–564, 2003.
- [31] Rozenn Dahyot, Anil Kokaram, Niall Rea, and Hugh Denman. Joint audio visual retrieval for tennis broadcasts. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 3, pages 561–564, April 2003.
- [32] K Das and J Schneider. Detecting anomalous records in categorical datasets. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 2007.
- [33] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *NIPS: Twenty-Fifth Annual Conference on Neural Information Processing Systems*, Granada, Spain, 2011. NIPS Foundation.
- [34] H. Denman, N. Rea, and A. Kokaram. Content-based Analysis for Video from

- 
- Snooker Broadcasts. *ELSEVIER Computer Vision and Image Understanding*, 92(2-3):176–195, November-December 2003.
- [35] H. Denman, N. Rea, and Anil C. Kokaram. Content based analysis for video from snooker broadcasts. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 198–205, London, UK, 2002. Springer-Verlag.
- [36] C P Diehl and J B Hampshire. Real-time object classification and novelty detection for collaborative video surveillance. In *Proceedings of IEEE International Joint Conference on Neural Networks*, pages 2620–2625, 2002.
- [37] Ling-Yu Duan, Min Xu, Tat-Seng Chua, Qi Tian, and Chang-Sheng Xu. A mid-level representation framework for semantic sports video analysis. pages 33–44, 2003.
- [38] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, 1972.
- [39] F Y Edgeworth. On discordant observations. *Philosophical Magazine*, 23:5:364–375, 1887.
- [40] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic Soccer Video Analysis and Summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, July 2003.
- [41] E Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 255–262. Morgan Kaufmann Publishers Inc., 2000.
- [42] F Esponda, S Forrest, and P Helman. A formal framework for positive and negative detection schemes. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 34:1:357–373, 2004.
- [43] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. <http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html>.

- 
- [44] W Fan, M Miller, S J Stolfo, W Lee, and P K Chan. Using artificial anomalies to detect unknown and known network intrusions. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 123–130. IEEE Computer Society, 2001.
- [45] Usama Fayyad, Padhraic Smyth, N Weir, and S Djorgovski. Automated analysis and exploration of image databases: results, progress, and challenges. *Journal of Intelligent Information Systems*, (4), 1995.
- [46] James Ferryman and James L. Crowley, editors. *Proceedings of the Thirteenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Boston, USA, 29 August 2010. IEEE Computer Society (PAMI TC) and IEEE Signal Processing Society (IVMSP TC). In conjunction with the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance.
- [47] Shai Fine. The hierarchical hidden markov model: Analysis and applications. In *Machine Learning*, volume 32, pages 41–62, 1998.
- [48] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32(1):41–62, 1998.
- [49] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32(1):41–62, 1998.
- [50] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. Of the ACM*, 24:381–395, June 1981.
- [51] David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O’Brien, and Deva Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3):77–254, 2006.
- [52] D Gatica-Perez, Ming-Ting Sun Ming-Ting Sun, and Chuang Gu Chuang Gu.

- 
- Multiview extensive partition operators for semantic video object extraction, 2001.
- [53] Zoubin Ghahramani and Michael I. Jordan. Factorial Hidden Markov Models. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 472–478. MIT Press, 1995.
- [54] A Giordana and F Neri. Automated learning for industrial diagnosis. In *In P. Langley & Y. Kodratoff (Eds.), Fielded applications of machine learning*. Morgan Kaufmann.
- [55] Yihong Gong, Lim Teck Sin, Chua Hock Chuan, Hongjiang Zhang, and Masao Sakauchi. Automatic Parsing of TV Soccer Programs. In *Proceedings of the 1995 International Conference on Multimedia Computing and Systems (ICMCS'95)*, volume 2, pages 167–174, May 1995.
- [56] C Guilfoyle. Ten minutes to lay the foundations. expert systems user, 1986.
- [57] J.-Y. Guillemaut and A. Hilton. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *International Journal of Computer Vision*, 93(1):73–100, 2011.
- [58] K. Hariharakrishnan and D. Schonfeld. Fast object tracking using adaptive block matching. *Multimedia, IEEE Transactions on*, 7(5):853–859, Oct. 2005.
- [59] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.
- [60] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
- [61] Md. Rafiul Hassan and Baikunth Nath. Stockmarket forecasting using hidden markov model: A new approach. In *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications, ISDA '05*, pages 192–196, Washington, DC, USA, 2005. IEEE Computer Society.

- 
- [62] Z He, X Xu, and S Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24:9-10:1641–1650, 2003.
- [63] Z He, X Xu, J Z Huang, and S Deng. Mining class outliers: Concepts, algorithms and application in crm. *Expert Systems and Applications*, 27:4:681–697, 2004.
- [64] P Helman and J Bhangoo. A statistically based system for prioritizing information exploration under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics*, 27:449–466, 1997.
- [65] V Hodge and J Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:2:85–126, 2004.
- [66] P Huber. *Robust Statistics*. Wiley, New York, 1974.
- [67] N. Ikisler and D. Forsyth. Searching video for complex activities with finite state models. June 2007.
- [68] Michael Isard and Andrew Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [69] Yuri A. Ivanov and Aaron F. Bobick. Recognition of Visual Activities and Interactions by Stochastic Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, August 2000.
- [70] N Japkowicz, C Myers, and M A Gluck. A novelty detection approach to classification. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 518–523, 1995.
- [71] M V Joshi, R C Agarwal, and V Kumar. Predicting rare classes: can boosting make any weak learner strong? In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 297–306. ACM, New York, NY, USA, 2002.
- [72] Bostjan Kaluza, Violeta Mirchevska, Erik Dovgan, Mitja Lustrek, and Matjaz Gams. An agent-based approach to care in independent living. In *Ambient Intelligence*, volume 6439 of *LNCS*, pages 177–186. Springer Berlin / Heidelberg, 2010.

- 
- [73] N Karba. Expert system for the cold rolling mill of the steel works jesenice. In *Proceedings of the Thirteenth Symposium on Information Technologies*, 1989.
- [74] E Keogh, J Lin, and A Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 226–233. IEEE Computer Society, Washington, DC, USA, 2005.
- [75] E Keogh, J Lin, S-H Lee, and H V Herle. Finding the most unusual time series subsequence: algorithms and applications. *Knowledge and Information Systems*, 11:1:1–27, 2006.
- [76] A. Khan, D. Windridge, T. de Campos, J. Kittler, and W. Christmas. Lattice-based anomaly rectification for sport video annotation. In *Proc. ICPR*, 2010.
- [77] A. Khan, D. Windridge, and J. Kittler. Multi-level Chinese takeaway process and label-based processes for rule induction in the context of automated sports video annotation. *IEEE Transactions on Cybernetics*, 2013. Under review.
- [78] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Audiovisual integration for tennis broadcast structuring. In *International Workshop on CBMI*, pages 289–312, 2003.
- [79] Ewa Kijak, Lionel Oisel, and Patrick Gros. Hierarchical Structure Analysis of Sport Videos Using HMMs. In *Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP 2003)*, volume 2, pages 1025–1028, September 2003.
- [80] J. Kittler, W. Christmas, T. de Campos, D. Windridge, and F. Yan. Domain anomaly detection in machine perception: A framework and taxonomy. 2013. Under review.
- [81] J. Kittler, W J Christmas, F Yan, I Kolonias, and D Windridge. A memory architecture and contextual reasoning for cognitive vision. In *Proc. SCIA*, pages 343–358, 2005.
- [82] Jyri J. Kivinen, Erik B. Sudderth, and Michael I. Jordan. Learning multiscale

- 
- representations of natural scenes using dirichlet processes. In *ICCV*, pages 1–8, 2007.
- [83] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *19th British Machine Vision Conference*, pages 995–1004, 2008.
- [84] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman. Human focused action localization in video. In *International Workshop on Sign, Gesture, Activity*, 2010. (best paper award winner) in conjunction with ECCV.
- [85] E M Knorr, R T Ng, V, and Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8:3-4:237–253, 2000.
- [86] I. Kolonias. *Cognitive Vision Systems for Video Understanding and Retrieval*. PhD thesis, University of Surrey, 2007.
- [87] I. Kolonias. *Cognitive Vision Systems for Video Understanding and Retrieval*. PhD thesis, University of Surrey, 2007.
- [88] I. Kolonias, W. Christmas, and J. Kittler. A layered active memory architecture for cognitive vision systems. In *Proceedings of the 5th International Conference on Computer Vision Systems*, March 2007.
- [89] I. Kolonias, T. de Campos, F. Yan, W. Christmas, J. Kittler, A. Kostin, and D. Windridge. A bayesian reasoning system for sports video annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012. Under review.
- [90] Ilias Kolonias. *Cognitive Vision Systems for Video Understanding and Retrieval*. PhD thesis.
- [91] Antti Koski. Modelling ecg signals with hidden markov models. *Artif. Intell. Med.*, 8(5):453–471, October 1996.
- [92] J D Lafferty, A McCallum, and F C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

- 
- [93] Pat Langley and Herbert A. Simon. Applications of machine learning and rule induction. *Communications of the ACM*, 38:55–64, 1995.
- [94] W.J. Leech. A rule based process control method with feedback. In *Proceedings of the ISA/86*, pages 169–175, Research Triangle Park, NC 27709, 1986. The Instrumentation Society of America.
- [95] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1):259–289, May 2008.
- [96] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An Introduction to the Application of the theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell Sys. Tech. Journal*, 62(4), 1983.
- [97] Xiaohui Liu and Chin-Seng Chua. Multi-agent Activity Recognition Using Observation Decomposed Hidden Markov Model. In *Proceedings of the Third International Conference on Computer Vision Systems (ICVS 2003)*, pages 247–256, April 2003.
- [98] Y. Liu and K.N. Ngan. Embedded wavelet packet object-based image coding based on context classification and quadtree ordering. 21(2):143–155, February 2006.
- [99] Ying Luo, Tzong-Der Wu, and Jenq-Neng Hwang. Object-based Analysis and Interpretation of Human Motion in Sports Video Sequences by Dynamic Bayesian Networks. *Computer Vision and Image Understanding*, 92(2-3):196–216, November-December 2003.
- [100] D. Magee, C. J. Needham, P. Santos, A. G. Cohn, and D. C. Hogg. Autonomous learning for a cognitive agent using continuous models and inductive logic programming from audio-visual input. In *Proceedings of the AAAI Workshop on Anchoring Symbols to Sensor Data*, 2004.
- [101] M Markou and S Singh. Novelty detection: A review-part 1: Statistical approaches. *Signal Processing*, 83:12:2481–2497, 2003.



- 
- [102] M Markou and S Singh. Novelty detection: A review-part 2: Neural network based approaches. *Signal Processing*, 83:12:2499–2521, 2003.
- [103] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):pp. 431–441, 1963.
- [104] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. 2009.
- [105] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Conference on Computer Vision & Pattern Recognition*, jun 2009.
- [106] A McNeil. Extreme value theory for risk managers. *Internal Modelling and CAD II*, pages 93–113, 1999.
- [107] Tao Mei, Yu-Fei Ma, He-Qin Zhou, Wei-Ying Ma, and Hong-Jiang Zhang. Sports Video Mining with Mosaic. In *Proceedings of the 11th IEEE International Multimedia Modelling Conference (MMM)*, pages 107–114, January 2005.
- [108] K. Messer, W.J. Christmas, E. Jaser, J. Kittler, B. Levienaise-Obadia, and D. Koubaroulis. A unified approach to the generation of semantic cues for sports video annotation. *Signal Processing*, 83:357–383, 2005. Special issue on Content Based Image and Video Retrieval.
- [109] Donald Michie. Problems of computer-aided concept formation. In *Applications of expert systems*. Addison-Wesley, 1989.
- [110] V. Mihajlovic and M. Petkovic. Automatic Annotation of Formula 1 Races for Content-Based Video Retrieval. Technical Report TR-CTIT-01-41, Centre for Telematics and Information Technology, Computer Science Department, University of Twente, The Netherlands, 2001.
- [111] Tom M. Mitchell. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997.
- [112] H. Miyamori. Automatic annotation of tennis action for content-based retrieval by integrated audio and visual information. In *IEEE Int. Conf. on Image and Video Retrieval*, pages 331–341, 2003.

- 
- [113] T Moeslund, A Hilton, and V Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–127, November 2006.
- [114] Stephen Muggleton. Inductive logic programming. In Stephen Muggleton, editor, *Inductive Logic Programming*, volume 38 of *The APIC Series*, pages 1–27. Academic Press, 1992.
- [115] Stephen Muggleton. Inverse entailment and prolog. *New Generation Computing, Special issue on Inductive Logic Programming*, 13:245–286, 1995.
- [116] Kevin P. Murphy. An introduction to graphical models, 2001. MIT technical report.
- [117] C. Harris N. Owens and C. Stennet. Hawk-Eye Tennis System. In *Proc. VIE 2003 – Intl. Conf. on Visual Information Engineering*, pages 182–185, July 2003.
- [118] A Nairac, T Corbett-Clark, R Ripley, N Townsend, and L Tarassenko. Choosing an appropriate model for novelty detection. In *Proceedings of the 5th IEEE International Conference on Artificial Neural Networks*, pages 227–232, 1997.
- [119] Milind R. Naphade and John R. Smith. On the detection of semantic concepts at trecvid. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pages 660–667, New York, NY, USA, 2004. ACM.
- [120] Chris J. Needham, Paulo E. Santos, Derek R. Magee, Vincent Devin, David C. Hogg, and Anthony G. Cohn. Protocols from perceptual observations. *Artificial Intelligence*, 2005.
- [121] Frank Nielsen. A family of statistical symmetric divergences based on jensen’s inequality. *CoRR*, abs/1009.4004, 2010.
- [122] T Ogata, W Christmas, J Kittler, and S Ishikawa. Tennis stroke detection and classification based on boosted activity detectors and particle filtering. In *Proceedings of the Joint 3rd International Conference on Soft Computing and Intelligent Systems and the 7th International Symposium on Advanced Intelligent Systems*, pages 2035–2040, September 2006.

- 
- [123] Stephen O'Hara and Bruce A. Draper. Introduction to the bag of features paradigm for image classification and retrieval. *CoRR*, abs/1101.3354, 2011.
- [124] H. P. M. Jimison, D. Weinshall, A. Zweig, F. W. Ohl, and H. Hermansky. Detection and identification of rare events in cognitive and engineering systems. Technical report, Idiap, 2008.
- [125] Tom Pajdla, Michal Havlena, and Jan Heller. Learning from incongruence. In Daphna Weinshall, Jörn Anemüller, and Luc Gool, editors, *Detection and Identification of Rare Audiovisual Cues*, volume 384 of *Studies in Computational Intelligence*, pages 119–127. Springer Berlin Heidelberg, 2012.
- [126] S. H. K. Parthasarathi, B. Motlicek., and H. Hermansky. Exploiting contextual information for speech/non-speech detection. In *LNCS - TSD*, volume 5246, pages 451–459, 2008.
- [127] A Patcha and J-M Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Networks*, 51:12:3448–3470, 2007.
- [128] A. Patron-Perez, M. Marszałek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. 2010.
- [129] M. Petkovic, W. Jonker, and Z. Zivkovic. Recognizing strokes in tennis videos using Hidden Markov Models. In *Proceedings of Intl. Conf. on Visualization, Imaging and Image Processing, Marbella, Spain*, Sep 2001.
- [130] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from tv formula 1 programs. volume 1, pages 817–820 vol.1, 2002.
- [131] Ronald Poppe. A survey on vision-based human action recognition. 28(6):976–990, June 2010.
- [132] W.H. Press. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*. Fortran Numerical Recipes. University Press, 1992.

- 
- [133] L. R. Rabiner and B. H. Juang. An introduction to Hidden Markov Models. *IEEE Signal Processing Magazine*, 61(3):4–16, June 1986.
- [134] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.
- [135] G Ratsch, S Mika, B Scholkopf, and K-R Muller. Constructing boosting algorithms from svms: An application to one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:9:1184–1199, 2002.
- [136] Niall Rea, Rozenn Dahyot, and Anil Kokaram. Modelling High Level Structure in Sports with Motion Driven HMMs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, volume 3, pages 621–624, May 2004.
- [137] Niall Rea, Rozenn Dahyot, and Anil Kokaram. Classification and Representation of Semantic Content in Broadcast Tennis Videos. In *IEEE International Conference on Image Processing (ICIP 2005)*, September 2005.
- [138] Ian Reid and Andrew Zisserman. Goal-directed Video Metrology. In *Proceedings of the 4th European Conference on Computer Vision (ECCV’96)*, volume 2, pages 647–658, 1996.
- [139] I.D. Reid and K. Connor. Multiview segmentation and tracking of dynamic occluding layers. *Image and Vision Computing*, 28(6):1022–1030, June 2009.
- [140] C Riese. Transformer fault detection and diagnosis using rulemaster by radian (technical report. In *Radian Corporation. Applications of Machine Learning 18*, 1984.
- [141] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
- [142] S Roberts. Novelty detection using extreme value statistics. *Proceedings of IEE - Vision, Image and Signal processing*, 146:124–129, 1999.
- [143] S Roberts and L Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computing*, 6:2:270–284, 1994.

- 
- [144] M-C Roh, W Christmas, J Kittler, and S-W Lee. Robust player gesture spotting and recognition in low-resolution sports video. In A Leonardis, H Bischof, and A Pinz, editors, *Proceedings of 9th European Conference on Computer Vision - Part IV*, pages 347–358. Springer, May 2006.
  - [145] R. Rosales and S. Sclaroff. Inferring Body Pose without Tracking Body Parts. In *Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 714–720, Jun 2000.
  - [146] V Roth. Kernel fisher discriminants for outlier detection. *Neural Computation*, 18:4:942–960, 2006.
  - [147] P J Rousseeuw and A M Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., 1987.
  - [148] P Salembier, L Torres, F Meyer, and Chuang Gu Chuang Gu. Region-based video coding using mathematical morphology, 1995.
  - [149] V Saligrama, J Konrad, and P-M Jodoin. Video anomaly identification. *IEEE Signal Processing Magazine*, 27:18–33, 2010.
  - [150] Lawrence K. Saul and Michael I. Jordan. Boltzmann Chains and Hidden Markov Models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 435–442. The MIT Press, 1995.
  - [151] B Schalkopf, J C Platt, J C Shawe-Taylor, A J Smola, and R C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:7:1443–1471, 2001.
  - [152] Jean Serra. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., Orlando, FL, USA, 1983.
  - [153] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
  - [154] Affan Shaukat, David Windridge, Erik Hollnagel, and Luigi Macchi. Adaptive, perception-action-based cognitive modelling of human driving behaviour using

- 
- control, gaze and signal inputs. In *Proceedings of Brain Inspired Systems 2010 (BICS 2010)*, 2010.
- [155] AlanF. Smeaton, Paul Over, and Wessel Kraaij. High-level feature detection from video in trecvid: A 5-year retrospective of achievements. In Ajay Divakaran, editor, *Multimedia Content Analysis*, Signals and Communication Technology, pages 1–24. Springer US, 2009.
- [156] P. Smith and G. Buechler. A branching algorithm for discriminating and tracking multiple objects. *Automatic Control, IEEE Transactions on*, 20(1):101 – 104, feb 1975.
- [157] S. M. Smith and J. M. Brady. SUSAN - a new approach to low level image processing. *Int. Journal of Computer Vision*, 23(1):45–78, May 1997.
- [158] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 399–402, New York, NY, USA, 2005. ACM.
- [159] X Song, M Wu, C Jermaine, and S Ranka. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19: 5:631–645, 2007.
- [160] A Soule, K Salamatian, and N Taft. Combining filtering and statistical methods for anomaly detection. In *IMC '05: Proceedings of the 5th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, New York, NY, USA, 2005.
- [161] Thad Starner, Joshua Weaver, and Alex Pentland. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.
- [162] C Stefano, C Sansone, and M Vento. To reject or not to reject: that is the question - an answer in case of neural classifiers. *IEEE Transactions on Systems, Man and Cybernetics Part C*, 30:1:84–94, 2000.

- 
- [163] I Steinwart, D Hush, and C Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.
- [164] P Sun and S Chawla. Slom: a new measure for local spatial outliers. *Knowledge and Information Systems*, 9:4:412–429, 2006.
- [165] D Tax and R Duin. Data domain description using support vectors. In M Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks*, pages 251–256, 1999.
- [166] D Tax and R Duin. Support vector data description. *Pattern Recognition Letters*, 20:11-13:1191–1199, 1999.
- [167] D M J Tax. *One-class classification; concept-learning in the absence of counter-examples*. Ph.D. thesis, Delft University of Technology, 2001.
- [168] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.
- [169] Tuan Hue Thia, Li Chengd, Jian Zhanga, Li Wange, and Shinichi Satohf. Structured learning of local features for human action classification and localization. 30(1):1–14, January 2012.
- [170] R F Thompson and W A Spencer. Habituation: A model phenomenon for the study of neuronal substrates of behaviour. *Psychological Review*, 73:1:16–43, 1966.
- [171] M. Tien, Y. Wang, and C. Chou. Event detection in tennis matches based on video data mining. In *IEEE Int. Conf. on Multimedia and Expo*, pages 1477–1480, 2008.
- [172] Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. Integrating Highlights for More Complete Sports Video Summarization. *IEEE Multimedia*, 11(4):22–37, October-December 2004.
- [173] T. Tommasi and B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In *Proc. BMVC*, 2009.

- 
- [174] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3081–3088, june 2010.
- [175] Tatiana Tommasi and Barbara Caputo. Towards a quantitative measure of rareness. In Daphna Weinshall, Jörn Anemüller, and Luc Gool, editors, *Detection and Identification of Rare Audiovisual Cues*, volume 384 of *Studies in Computational Intelligence*, pages 129–136. Springer Berlin Heidelberg, 2012.
- [176] Vasanth Tovinkere and Richard J. Qian. Detecting Semantic Events in Soccer Games: Towards a Complete Solution. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo (ICME'01)*, pages 1040–1043, 2001.
- [177] Pavan K. Turaga, Rama Chellappa, V. S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488, 2008.
- [178] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- [179] G C Vasconcelos, M C Fairhurst, and D L Bisset. Investigating feedforward neural networks with respect to the rejection of spurious patterns. *Pattern Recognition Letters*, 16:2:207–212, 1995.
- [180] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.
- [181] Fei Wang, Yu-Fei Ma, Hong-Jiang Zhang, and Jin-Tao Li. A Generic Framework for Semantic Sports Video Analysis Using Dynamic Bayesian Networks. In *Proceedings of the 11th International Multimedia Modelling Conference, 2005 (MMM)*, pages 115–122, January 2005.
- [182] H. Wang, M. M. Ullah, A. Käser, I. Laptev, and C. Schmid. Evaluation of local



- 
- spatio-temporal features for action recognition. In *20th British Machine Vision Conference*, 2009.
- [183] Jinjun Wang, Changsheng Xu, Engsiong Chng, Xinguo Yu, and Qi Tian. Event Detection based on non-broadcast Video. In *Proceedings of the 2004 IEEE International Conference on Image Processing (ICIP 2004)*, volume 3, pages 1637–1640, October 2004.
- [184] D. Weinshall, H. Hermansky, A. Zweig, J. Luo, H. Jimison, F. Ohl, and M. Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In *Advances in Neural Information Processing Systems (NIPS)*, Dec 2009.
- [185] T.A. Welch. A technique for high-performance data compression. *Computer*, 17(6):8–19, june 1984.
- [186] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.
- [187] Tao Xiang and Shaogang Gong. Video behavior profiling for anomaly detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5):893–908, may 2008.
- [188] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with domain knowledge and Hidden Markov models. *Pattern Recognition Letters*, 25(7):767–775, 2004.
- [189] Eric P. Xing and Kyung ah Sohn. Hidden markov dirichlet process: Modeling genetic recombination in open ancestral space. In *Bayesian Analysis*. MIT Press, 2007.
- [190] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, Yong Rui, and Thomas S. Huang. *A Unified Framework for Video Summarization, Browsing and Retrieval: with Applications to Consumer and Surveillance Video*. Academic Press, 21 December 2005.

- 
- [191] Peng Xu, Lexing Xie, Shih-Fu Chang, Ajay Divakaran, Anthony Vetro, and Huifang Sun. Algorithms and system for Segmentation and Structure Analysis in Soccer Video. In *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo (ICME)*, pages 721–724, 2001.
- [192] F. Yan, W. J. Christmas, and J. Kittler. Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences. *Trans. PAMI*, 2008.
- [193] B. Yao and L. Fei-Fei. Recognizing human actions in still images by modeling the mutual context of objects and human poses. 2012. in press.
- [194] N Ye. A markov chain model of temporal behavior for anomaly detection. In *Proceedings of the 5th Annual IEEE Information Assurance Workshop*, 2004.
- [195] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book Version 3.0*. Cambridge University Press, 2000.
- [196] Xinguo Yu, Hon Wai Leong, Joo-Hwee Lim, Qi Tian, and Zhenyan Jiang. Team possession analysis for broadcast soccer video based on ball trajectory. volume 3, pages 1811–1815 vol.3, Dec. 2003.
- [197] J Zhang and H Wang. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and Information Systems*, 10:3:333–355, 2006.
- [198] K Zhang, S F Shi, H Gao, and J Li. Unsupervised outlier detection in sensor networks using aggregation tree. In *International Conference on Advanced Data Mining and Applications*, pages 158–169, 2007.
- [199] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. pages 431–440, 2006.
- [200] K. Zimmermann, T. Svoboda, and J. Matas. Adaptive parameter optimization for real-time tracking. In *Workshop on Non-rigid Registration and Tracking through Learning (in proc. ICCV)*, 2007.

- 
- [201] K. Zimmermann, T. Svoboda, and J. Matas. Simultaneous learning of motion and appearance. In *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis, in conjunction with ECCV*, Marseille, 2008.
- [202] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *IEEE 11th International Conference on Computer Vision*, 2007.