A framework for automatic sports video annotation with anomaly detection and transfer learning

T. E. de Campos, Aftab Khan, Fei Yan, Nazli FarajiDavar
David Windridge, Josef Kittler, William Christmas*
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, GU2 7XH
acasva@list.surrey.ac.uk

Abstract

This paper describes a system that can automatically annotate videos and illustrates its application to tennis games. A unified apparatus is proposed, cast in a Bayesian reasoning framework. This is supported by a cognitive memory architecture that allows the system to store raw video data at the lowest cognitive level and its semantic annotation with increasing levels of abstraction up to determining the score of a game. Also embedded in the system is a set of mechanisms to detect anomalies caused by a change of domain in the input data. Once an anomaly is detected, transfer learning methods are triggered to adapt the knowledge to new domains, such as new sport modalities. We also present a generic framework for rule induction that is crucial in the context of an adaptive annotation system.

1 Introduction

This paper summarises the work that is being carried out as part of a project on Adaptive Cognition for Automated Sports Video Annotation. The goal of this project is to perform reasoning at all cognitive levels in the problem of sports video annotation problem: from pixel-wise segmentation to high-level game scoring. In particular, our goal is to avoid the pitfalls of purpose-built systems and allow for the creation of systems of arbitrary complexity and hierarchy using a standard building block for any level of the reasoning process. As a long term ambition, this should constitute the basis for a self-adaptive system which is able to detect when the domain has changed, trigger semi-supervised learning and transfer knowledge from the previously learnt background to a new domain (e.g., from tennis to badminton).

2 A cognitive memory system for tennis video annotation

We propose the use of a unified apparatus for reasoning in context which is cast in the Bayesian framework of evidential reasoning. This apparatus models the interaction between objects and events in a generic way and it is used as the main building block for any level of the reasoning process.

The inherent spatial relation of objects in a video sequence is modelled as an attributed relational graph (ARG) where the set of vertices consists of objects detected and tracked in the video sequence. The edges describe relations between each object and its neighbours. The vertices and edges are associated with sets of unary and binary features.

The problem of video interpretation is then formulated as one of finding the most probable set of labels for all objects, given the measurement information conveyed by the attributed relational

*http://cvssp.org/acasva/

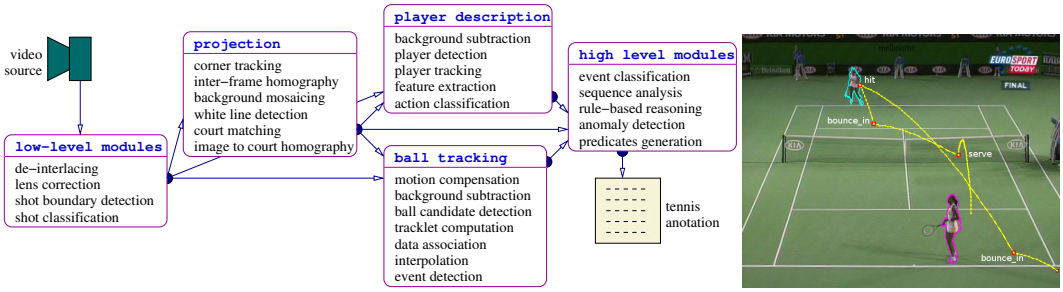

Figure 1: A summarised sketch of the architecture of the cognitive memory system for tennis annotation, grouping modules by their general purpose and cognitive level. The arrows indicate the main interactions between modules which is modelled as memory requirements. The panel on the right hand side shows the final tracking and event labelling result of a single camera shot superimposed on the final frame.

graphs at all frames up to the current frame as well as the interpretation of all the objects in the previous frames. The interpretation problem is essentially equivalent to the optimisation of the joint probability of all measurements and label assignments over the duration of the sequence. This optimisation can be carried out by taking advantage of the temporal continuity present in video sequences. By using the Markovian assumption and assuming we wish to simply demonstrate the step-by-step evolution of the labelling process, we show in [7] that the optimisation problem can greatly be simplified.

To enable the implementation of such apparatus for a problem with multiple cognitive levels with multiple types of scope in time and space, it is necessary to adopt an appropriated architecture. For that, we use a hierarchical memory mechanism which allows the system to store raw video data at the lowest level and its semantic annotation, with increasing levels of abstraction at the higher levels. This architecture is summarised in Figure 1 and it was designed with inspiration from cognitive systems. Each module has its own memory and operates at its own cognitive level, with a different requirement in terms of context in space and time. Each module has been implemented as a thread and they can all run in parallel. The system operates by invoking modules depending on the memory item that has been requested.

For instance, if the user requests the event classification result at a time t and the *event classification* module has 'forgotten' the estimated event at t (or has not computed it yet), the module *event classification* is invoked. In order to classify an event, this module requires that the ball position at frame t is known, so it invokes the memory from *ball tracking* modules. It also requires that the ball position is mapped onto a meaningful court position, which is a piece of memory information that is be produced by the *projection* modules. In addition, it requests the estimated player action of the nearest player to the ball, and this information is produced by the *action classification* module. Memory requests are propagated based on these dependencies until the system obtains all the required information and generates the requested output.

In [7], we give full details of this architecture and a description of all the modules. Next section presents our recent advances on event detection and classification. This is followed by sections describing methods that aim to make the system adaptable and autonomous.

3 Structured output learning for event detection and classification

In a nutshell, structured output learning (SOL) jointly embeds input-output pairs $(\mathbf{x}_i, \mathbf{y}_i)$ into a feature space, and applies linear classifiers in the feature space. In the case of hinge loss, a max-margin hyperplane is sought, and the resulting learning machine can be thought of as structured SVM (S-SVM). We formulate the problem of event classification in court games as one of learning a mapping from features to structured labels, and employ S-SVM to achieve a max-margin solution [8]. The features used for S-SVM are ball trajectories generated from the ball tracking module of our annotation system. We have compared closely the more popular generative approach based on the hidden

Markov model (HMM) with our discriminative approach on both artificial games and two real world tennis games, and observed a consistent advantage of our method.

4 Anomaly detection

As a prerequisite to transfer learning and cognitive bootstrapping, we include anomaly detection capabilities in the proposed framework. In [6] we introduce the concept of *domain anomaly* as distinct from the conventional notion of anomaly used in the literature and provide a taxonomy of domain anomaly events, which include outlier, noise, distribution drift, novelty detection (object, object primitive), rare events, and unexpected events. One of the mechanisms helping to pinpoint the nature of anomaly is based on detecting incongruence between contextual and noncontextual sensor(y) data interpretation.

We proposed to incorporate anomaly detection mechanisms in each of the groups of modules in the framework of Figure 1. At *low level*, the problem of detecting if a camera shot relates to tennis or not is cast as an outlier detection problem, i.e., if the global descriptors of motion, texture and colour do not conform with the model of a usual tennis game, it is classified as an outlier (or 'non-play shot' in tennis). In the *projection* modules, if the line detectors confidently detect typical tennis court lines, but together these lines do not conform with a court model, the method indicates that this is an anomalous shot (e.g. trying to run a system trained for tennis on a badminton match). In the *player description* modules if the person detector is confident but the number of people detected does not match the expectation, this is flagged as an anomaly (e.g. if a system is trained with singles and applied on double games). In the *ball tracking* modules, anomaly is detected when the blob classifier confidently detect a ball, but together, their trajectory does not match an expected ball track. In the *high level* modules, if the system is confident on all the mid-level results but together the data does not conform with the rules of tennis, an anomaly is detected. We evaluated the latter in [2].

5 Transductive transfer learning

Once a domain anomaly has been detected, a process of domain adaptation is triggered and the system starts to process incoming data to learn how to transfer existing knowledge to the new domain. We proposed and evaluated techniques for transductive transfer learning using feature space transformations. In a classification problem, these methods use an initial estimate of the class posterior $P_{\Lambda_{src}}(y|\mathbf{X}_i^{trg})$ for each of the unlabelled samples $\mathbf{x}_i^{trg}$ from the target domain. To obtain those initial probabilities, we use the source domain samples for training $\mathbf{x}_i^{src}$. Based on these estimates, we compute transformations $G(x^i_{j\,src})$ that reduces the difference between the distributions of samples in source and target domain. We used two types of linear transformations that are applied independently to each feature j and each class y. In [3], these methods were evaluated for the problem of player action classification, where each match constitutes a domain. The benefits were clear when transferring between tennis and badminton and it was also evident between pairs of tennis matches (shot at different locations, with different players).

6 Multi-level Chinese takeaway process and label-based processes for rule induction

A generalized high-level module for the framework of Figure 1 is required for an adaptive annotation system. For this purpose, four novel methods are proposed in [5] for generating hierarchical Hidden Markov Models (hHMMs) [4] for rule induction. The first method, Cartesian Product Label-Based Hierarchical Bottom-Up Clustering (CLHBC), employs the latent structure in the labels used to annotate videos. These labels are thus employed to build hierarchical structures based on various Cartesian Product-based combinations, such that an hHMM of common repeated event structures is established. The second method proposed, the Multi-Level Chinese Takeaway Process (MLCTP), is based on the classical Chinese Restaurant Process [1] with, analogically, tables replaced by takeaways that may be re-visited within different cities representing levels in the rule hierarchy. This is a stochastic process, with many hierarchies generated among which the highest likelihood stochastic rule structure is inferred.

We also propose two hybrid methods in [5], the MLCTP with recursive Baum-Welch estimated hidden state transitions and the MLCTP-CLHBC, that leverage the stochasticity of MLCTP (whereby various hierarchical structures are produced), in conjunction with the label sequence resulting in a composite approach to hHMM inference. All of these methods finally generate finite intermediate-depth hHMMs 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.

Comparative prediction results for all of the proposed methods reveal that all of the hierarchical methods perform better relative to the flat Markov Model (with the most optimal method being the MLCTP-CLHBC hybrid). Rule induction framework can also be employed to address the problem of transferring knowledge from one domain to another via analysing 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. minimising the need for re-training.

7 Summary

This paper presented an overview of activities of a project on Adaptive Cognition for Automated Sports Video Annotation, which aims to annotate court sport videos at all cognitive levels in an adaptable way. For that, we proposed a generic framework and memory architecture and indicated how to include adaptation capabilities by detecting anomalies, performing domain adaptation and inducing rules.

Acknowledgements

We are grateful for the support of the PASCAL2 network of excellence (EC) and EPSRC-UK through grant EP/F069421/1 (ACASVA).

References

[1] 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.

[2] I. Almajai, F. Yan, T. deCampos, A. Khan, W. Christmas, D. Windridge, and J. Kittler. Anomaly detection and knowledge transfer in automatic sports video annotation. *Studies in Computational Intelligence*, 384, Detection and Identification of Rare Audiovisual Cues:109–117, 2012.

[3] N FarajiDavar, T E deCampos, D Windridge, J Kittler, and W Christmas. Domain adaptation in the context of sport video action recognition. In *Domain Adaptation Workshop, in conjunction with NIPS*, 2011.

[4] Shai Fine, Yoram Singer, and Naftali Tishby. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32(1):41–62, 1998.

[5] 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.

[6] J. Kittler, W. Christmas, T. de Campos, D. Windridge, and F. Yan. Domain anomaly detection in machine perception: A framework and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. Under review.

[7] 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 Cybernetics*, 2013. Under review.

[8] F. Yan, J. Kittler, K. Mikolajczyk, and D. Windridge. Automatic annotation of court games with structured output learning. In *International Conference on Pattern Recognition*, 2012.