Anomaly Detection and Knowledge Transfer in Automatic Sports Video Annotation

I Almajai, F Yan, T de Campos, A Khan, W Christmas, D Windridge and J Kittler

Abstract A key question 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 [10] 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. Our ultimate aim, building on existing annotation technology, is to construct an adaptive system for court-based sport video annotation.

1 Introduction

Artificial cognitive systems 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 [8]). The anomaly detection problem 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

CVSSP, University of Surrey, Guildford GU2 7XH, UK, e-mail: d.windridge@surrey.ac.uk
This is the preprint of a paper published in Studies in Computational Intelligence, Volume 384, 2012, Pages 109-117, DOI: 10.1007/978-3-642-24034-8_9, Editors: D. Weinshall, J. Anemuller and L. van Gool. ©2012 Springer-Verlag Berlin Heidelberg.

anomalies in a robust manner, maximising 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 paper, 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 [10, 15] which distinguishes outliers from anomalies via the disparity between a generalised 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 characterised as being between strongly constrained (contextual) and weakly constrained (non-contextual) classifiers [2]. A similar approach can be used for model updating and acquisition within the context of tracking [13] and for the simultaneous learning of motion and appearance [14]. Such tracking systems explicitly address the *loss-of-lock* problem that occurs without model updating.

In this paper 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 [5, 11], 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 3 we discuss the problem anomaly detection mechanism. We describe some experiments to validate the ideas in Section 4, incorporating the results into the anomaly-adaptation/rule-update stage in the immediately following section. We conclude the paper in Section 6.

2 Weak Classifiers

2.1 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 [11] 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 analysing 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.

Fig. 1 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 that the high level interpretation module may misinterpret the event sequences. A set of continuous-density left-to-right first-order HMMs, $\Lambda = \{\lambda_1, \ldots, \lambda_k, \ldots, \lambda_K\}$ are used to anal-

yse the ball trajectory dynamics and recognise events regionally within the tracked ball trajectory, based on [1], but using the detected ball motion changes to localise 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: $\mathbf{o}_t = \{\dot{\mathbf{x}}_t, \ddot{\mathbf{x}}_t\}$. To classify an event at a time t, a number of observations $\mathbf{O}_t = \mathbf{o}_{t-W}, \mathbf{o}_{t-W+1}, ..., \mathbf{o}_{t+W}$ are considered within a window of size $2W+1$. Each HMM is characterised by three probability measures: the state transition probability distribution matrix A, the observation probability distribution B and the initial state distribution π, defined for a set of N states $S = (s_1, s_2, ..., s_N)$, and ball information observation sequence $\mathbf{O}_t$. Each state s_j is represented by a number, M_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 [12]. 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 [12]. The event is then classified by computing $\hat{k} = \arg\max_k (P(\mathbf{O}_t|\lambda_k))$.

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 recognised 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. [11]. This happens often when there is a long time gap between two recognised events. To recover from such suspected false negatives, an exhaustive search is used to find other likely events in such gaps.

2.2 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 [5]. To reduce the number 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. Fig. 2 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 centre 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

Fig. 2 Sample images and detected players performing each primitive action of tennis.

frames. This value was determined using the validation set of the KTH dataset. As a spatio-temporal descriptor, we chose the 3DHOG (histogram of oriented gradients) method of Klaser et al. [6]. This method gave state-of-the-art results in recent benchmarks of Wang et al. [9] 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 [3], 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 keypoint 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) at the beginning of a play shot, we assume that detected *hits* are actually *serves*; (iii) at later moments, serves are no longer enabled, i.e., if a serve is detected later in a play shot, the action is classified as *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 (normalised distance to the decision boundary).

2.3 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 neighbouring regions. This can be modelled by a convolution between the probability function $\hat{P}(bounce_{in}|\mathbf{x}_t)$, 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 σ_{ball}.

$$P(bounce_{in}|\mathbf{x}_t) = \int_{\psi} \hat{P}(bounce_{in}|\psi)p(\mathbf{x}_t - \psi)d\psi \quad (1)$$

Finally, the probability of $bounce_{out}$ is given by $P(bounce_{out}|\mathbf{x}_t) = 1 - P(bounce_{in}|\mathbf{x}_t)$.

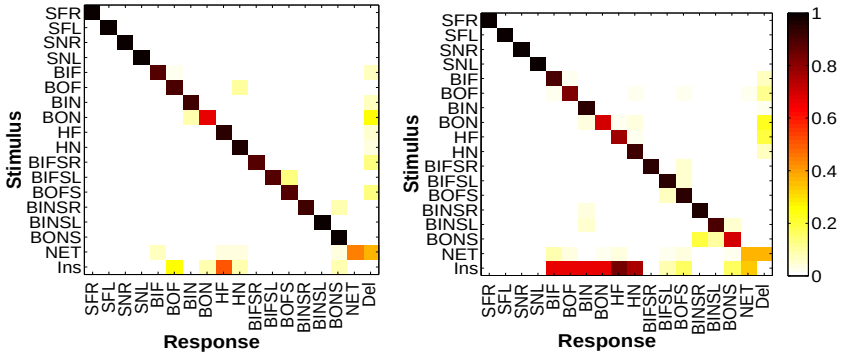
2.4 Combining evidence

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 [7]). 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.

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 [7]. Here, each sequence of events starting with a serve is analysed 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 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. 1)). 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



(a) Validation dataset (singles) (b) Test dataset (doubles)

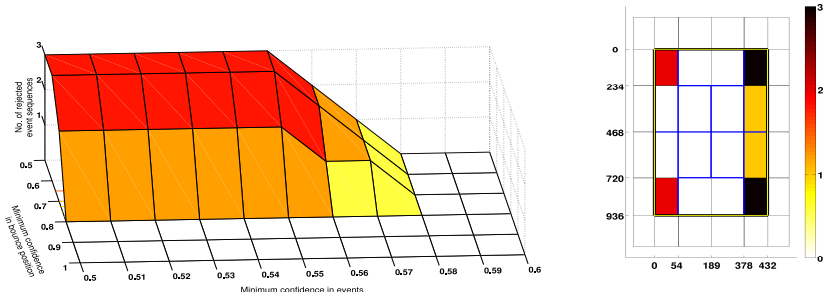
Fig. 3 Confusion matrices of recognised events, K. We use the same labels as [1].

play area does not appear to exceed four events. This is consequently our basis for classification of context.

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 for validation. The test data is composed of 163 play shot 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 parameterisation of these HMMs was optimised on the validation data. A window size of 7 observations is selected ($W = 3$ in Sec. 2.1). An accuracy of 88.73% event recognition was reached on the validation data, see figure 3(a). 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 (Fig. 4(a)). 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 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



(a) Confidence of weak classifiers vs. no. rejected events (singles) (b) Anomalies (doubles)

Fig. 4 (a) Number of event sequences of the validation set that contain errors with varied confidence thresholds and (b) Detected anomaly triggering events of the test set.

using the parameters optimised on the singles data (Fig. 3(b)). 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.

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(b). 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 [4]. However, in the absence of false positives, and with 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 Fig. 4(b), where rectangles are defined by $\{$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 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’).

6 Conclusions

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 in 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; more stochastic methods would be required were this not the case.

References

1. 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. In *Proc Int Conf on Image Processing, Hong Kong, September 26-29*, 2010. In press.
2. L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernock. Combination of strongly and weakly constrained recognizers for reliable detection of OOVs. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4081–4084, 2008.
3. T. de Campos, M. Barnard, K. Mikolajczyk, Josef Kittler, F. Yan, W. Christmas, and D. Windridge. An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In *Proc 10th IEEE Workshop on Applications of Computer Vision, Kona, Hawaii.*, January 5-6 2011.
4. A. Khan, D. Windridge, T. de Campos, J. Kittler, and W. Christmas. Lattice-based anomaly rectification for sport video annotation. In *Proc. ICPR*, 2010.
5. 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.
6. 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.
7. I. Kolonias. *Cognitive Vision Systems for Video Understanding and Retrieval*. PhD thesis, University of Surrey, 2007.
8. T. Tommasi and B. Caputo. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In *British Machine Vision Conference*, 2009.
9. 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.

10. 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.
11. F. Yan, W. Christmas, and J. Kittler. Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 2008.
12. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book Version 3.0*. Cambridge University Press, 2000.
13. 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.
14. 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*, Marseilles, 2008. In conjunction with ECCV.
15. A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *IEEE 11th International Conference on Computer Vision*, 2007.