Advanced Techniques of Foreground, Background and Object Identification in Video Application: A Survey

Utkarsh Shukla

***Abstract*— In this paper we investigate real-time tracking of a tennis-ball using various image differencing techniques. First, we considered a simple background subtraction method with subsequent ball verification (BS). We then implemented two variants of our initial background subtraction method. The first is an image differencing technique that considers the difference in ball position between the current and previous frames along with a background model that uses a single Gaussian distribution for each pixel. The second is uses a mixture of Gaussians to accurately model the background image. Each of these three techniques constitutes a complete solution to the tennis ball tracking problem. In a detailed evaluation of the techniques in different lighting conditions we found that the mixture of Gaussians model produces the best quality tracking. Our contribution in this paper is the observation that simple background subtraction can outperform more sophisticated techniques on difficult problems, and we provide a detailed evaluation and comparison of the performance of our techniques, including a breakdown of the sources of error.**

***Index Terms*— Foreground, Background, shadow and object detection**

I. INTRODUCTION

While ball tracking systems have been successful in soccer (football) and baseball, ball tracking in tennis matches is less well explored. Applications including computer-assisted refereeing and computer assisted coaching could benefit from real-time tracking of the tennis ball. However, ball tracking in a tennis match poses particular challenges owing to the ball's small size, high speed, and large variation in trajectories. Soccer balls are relatively large and move relatively slowly, while baseballs are slightly larger and (in a baseball pitch) have a much more highly constrained trajectory. Neither object-based techniques nor non-object based techniques are suitable for this application. Such techniques have limited processing speed and sometimes lack the ability to track objects which move significant distances between frames. In this paper, we present an investigation of image processing algorithms aimed at tennis ball tracking. We use background subtraction as the first step in the tracking process; background subtraction generates a number of regions representing changes in the image, all of which are possible ball locations, or ball candidates. We determine which ball candidates to report as tennis balls based on size and shape analysis of the candidate regions. First, we present a basic algorithm which uses an extremely simple, static background model; the results from this were encouraging, so we created two variant methods with different augmentations to the background model. The variants outperformed both the initial method and two existing alternative methods, even in a context where the background varied considerably and background subtraction might be thought unsuitable. Tennis ball tracking systems were reported by Sudhir et al. (Sudhir et al., 1998) and by Pingali et al. (Pingali et al., 2000); neither group of authors made a systematic analysis of their systems' performance in a real-world environment. Also, they did not address ball occlusion or player-ball interaction. Our paper describes the results of our real-world deployment and gives detailed data on error rates and sources of error.

Fig. 1: Representation of Lawn Tennis ground

II. PREVIOUS WORK

Retrieval and summarization of sports footage have received increasing interest in recent years. It **is** expected that the rise of home Digital media will increase the demand for easily browsable content, and hence the need **for** automatic content manipulation. Much of the work has concentrated on using video analysis for example, Sudhir et al. [14] extracted events from tennis using video analysis and the geometry of the court; Chang et.al [15] have used **HMMs** for summarizing Baseball footage. Work is emerging that considers the audio signal for spotting important events [16,17]. In general for sports the audio signal **is** capable of characterizing much shorter duration events than the video signal. In sports like tennis, cricket, badminton it is the short and sharp noise of the ball hitting the racket or bat that defines the basic building block action of the game. Both the audio and video signals therefore contain useful information and this work considers the use of both audio and video features for parsing tennis footage. The basic unit of this game is the serve and subsequent rally.

This work was funded **by EU** RTN MOUMIR **www.moumir.org (HP 118)** and **MUSE-DTV, CASMS,** Enterprise Ireland. or passage of play until the point **is** decided. Tennis summaries therefore generally contain the main court view during each of the main points.

Utkarsh Shukla, Lecturer, Department of Computer Science, Shri Ramdevi Ramdayal Tripathi Mahila polytechnic, Kanpur, India

This article presents a mechanism for extracting each rally by identifying the court view and by building a mechanism that 'listens' for the sound of the racket hitting the ball (referred to as a racket hit in this paper). The video sequence is first segmented into shots using the common histogram analysis technique. The task addressed by this paper is to identify each shot that is a passage of play shot containing the court view. The classification can be achieved by noting that the relevant shots contain both a full court view and a noise of the ball hitting the rackets.

A number of object detection and tracking techniques have been developed in the last two decades for tracking humans (Rano et al., 2004) and cars (Stauffer and Grimson, 1999). More recently, researchers have examined computer vision techniques for tracking sporting events (Han et al., 2002; Assfalg et al., 2002; Sudhir et al., 1998). Here we review the related literature on computer vision based object detection and tracking techniques.

Viola et al. (Viola and Jones, 2001) introduced classifier cascades for object recognition. They trained a set of weak classifiers on a set of very simple features, one classifier per feature; the classifiers are used in sequence to detect the presence of the target object, and since the weak classifiers are able to reject most non-target objects quickly, the majority of the computational effort is spent on difficult cases.

Lienhart and Maydt (Lienhart and Maydt, 2002) extended this work by proposing a richer set of features (Haar-like features, including edge, line, and centersurround features) and showing a lower false positive rate than was achieved by the simple feature set of Viola et al (Viola and Jones, 2001).

Stauffer and Grimson (Stauffer and Grimson, 1999) proposed a background model in which each pixel is a mixture of Gaussian distributions; pixels which fit into some existing distribution are considered background, while pixels which lie outside all distributions are considered foreground. The method allows the distributions to adapt to new samples, so that only parts of the image which change faster than a set learning rate are still considered foreground, and portions which change more slowly are incorporated into the background.

Ren et al. (Ren et al., 2004) devised K-ZONE, a system for tracking baseball pitches. They used a mixture of Gaussians for background discrimination; their method uses trajectory information to reject some ball candidates. They report good results for their context, but the trajectory of the baseballs is considerably constrained compared to the variation we can expect in a tennis match.

D’Orazi et al. (D’Orazio et al., 2002) propose a system for tracking soccer balls using a modified Hough transform. They use the parametric representation of a circle to transform the image and determine points which are on the soccer ball. They show that the circular Hough transform is effective in detecting the soccer ball. However, their algorithm requires considerable processing to be viable as a real time ball tracking technique.

In (Sudhir et al., 1998) the authors perform an automatic analysis of tennis video to facilitate content based retrieval. They generate an image model for the tennis court-lines based on the knowledge of the dimensions and connectivity of a tennis court and typical geometry used when capturing a tennis video. They use this model to track the tennis players over a sequence of images. In (Pingali et al., 2000) the authors use multiple cameras to track the 3D trajectory of the ball using stereo matching algorithms. A multi-thread approach is taken to track the ball using motion, intensity and shape. However, they do not give enough details of their implementation to compare their approach with ours.

Throughout this paper we use various image processing techniques, including median filtering and shape feature extraction (Shapiro and Stockman, 2001). The median filter is used to reduce noise in the image while shape features, including aspect ratio, compactness, and roughness, are used to check if a region’s properties resemble a ball or not.

III. Object detection/Recognition

The ultimate objective of a completely automated object tracking system is probably event recognition. Despite the fact that it is very critical and valuable to recognize an activity, it is difficult to characterize the motion type that is interesting and significant inside the sports context. Thus, there are numerous studies addressing diverse events types. Polana and Nelson [1] figure the optical flow fields between successive frames and sum up the vector magnitudes in regions of object to gain high dimensional component vectors that are utilized for recognition. Exercises are grouped by using the closest neighbour algorithm. To discover simple movement characteristics again attempted and in [2] proposes a “star” skeletonization strategy. The items are recognized by using background subtraction and then their boundaries are removed and a skeleton is created. The authors claim that skeletonization gives vital motion signs like posture of body and cyclic motion of skeleton segments, which thusly are used in finding human activities like walking or running [2] Rather than making analysis of simplistic object motions, patterns of activity in time might also be observed. A state-based learning architecture was proposed in [3] with coupled hidden Markov models (CHMM), to model behaviors of object and communications between them. Object motion was represented by Johnson, *et al.* [4] using flow vectors, which include spatial location and instantaneous object velocity. Then, the trajectories are built as a grouping of flow vectors and a competitive learning network is adapted to model the probability density functions of flow vector sequences. In the similar way, [5] produce probabilistic models to describe the normal motion in the scene. The flow vectors are further quantized to get a prototype representation and trajectories are converted into prototype vector sequences. Thereafter, these sequences are evaluated using the probabilistic trajectory models. A codebook of prototype representations was produced in [6] from input representations (*x,y, vx, vy, size of object, binary mask*) using on-line *Vector Quantization* (VQ). At that point, a co-occurrence matrix is characterized over the prototypes in the codebook and a hierarchical classifier is produced by making use of co-occurrence data. Lee, *et al.* [7] likewise work with prototype vectors and its objective is of the classification of both local and global trajectory points. *Support Vector Machines* are used by them for the detection of local point abnormality while the classification of global trajectories (sequences of vectors) is done by using HMMs. As a last step, a rule-based system consolidates local and global information to make the decision on the abnormality of the motion pattern [7].

Recognition of the persons’ entering the scene is another essential part of a object tracking framework. Most recent studies on individual identification exhibit the popularity of architectures that are based on biometrics (distinctive personal components). Face and gait are the prime biometric features that can be seen inside passive object tracking context [6].

We have a long history of research on face recognition and there are many studies on face tracking, face recognition, face detection, extraction of facial features [9][10][11]. Gait based recognition has acquired more attention in recent couple of years. The above mentioned studies can be grouped into three principle categories: physical feature based method model-based methods, statistical methods. Anatomical models are used by model based methods for analyzing gait of a person.

To construct the models, parameters such as joint trajectories or angular speeds are used. In statistical strategies, moment features of object regions are used for recognizing peoples. At last, physical feature based models utilizes the geometric structural properties of human body to recognize the movement pattern of a person. Some of these are cadence, height and stride length. Detailed discussion on the gait-based recognition studies could be found in [6].

IV. VIDEO FEATURES AT FRAME LEVEL

This section presents respectively the visual and audio features used, and proposes expressions for their likelihoods. The likelihoods express the probabilities that a frame represents a main court view and that the audio represents racket hits.

Frame level visual feature

The second moment of the though transform of the edges is computed for each image [18]. This measure, noted xu, is used to detect frames showing **a** main view of the court where its value remains constant. This moment feature is low when there is strong scene geometry since the Hough space will contain compact clusters representing major lines in the image. As the large view of the court is dominated by the physical, rectangular court structure the feature works well to discriminate it without the need to resort to any 3D information (as used in [14]).

Frame level audio feature

Since the racket hit is a shon sound between **10** to 20 ms long, we have chosen to compute the spectrogram of the audio track using a 40 ms window (duration of a frame in the video). The power spectrum of this Fourier transform, normalized by its energy, is then computed for each window and corresponds to our audio features.

Eigenspace representation. K audio features corresponding to racket hits are collected. **A** Principal Component Analysis (PCA) is then performed over this training database. *J* eigenvectors corresponding to the *J* highest eigenvalues are retained to span the eigenspace *F*.

Distance from the feature space. A common way to measure the similarity of an unknown observation xa with the training cloud, is to compute the distance between x. and the eigenspace *F*. This ***Distance Fmm Feature Space*** (DFFS) is defined as *[19]:*

$$\mathrm{dffs}(\mathbf{x}_a) = \|\mathbf{x}_a - \mu_a - \mathrm{U}^{\mathbf{T}}(\mathbf{x}_a - \mu_a)\|$$

Likelihood of having a Racket hit. Assuming **a** uniform distribution over the eigenspace *F*, the likelihood of having a racket hit can be approximated **[19, 20]** using the likelihood of the reconstruction error:

$$\mathcal{P}(\mathbf{x}_a|\mathrm{Racket\ hit}) \propto \exp\left[\frac{-(\mathrm{dffs}(\mathbf{x}_a))^2}{2\,\sigma_a^2}\right]$$

V. SHADOW REMOVAL

At the time of the objects segmentation from the background, moving cast shadows are constantly misclassified, as a moving object part. As the shadow causes an important intensity change on the surface it is cast upon, hence this result is expected. On the other hand, the segmentation of the moving objects that are desired should not contain shadows. An algorithm is applied on the change mask to uproot them [12-13]. The aim behind the method is as follows: On casting a shadow upon a surface, the intensity value reduces in a significant way, while normalized color value does not change much.

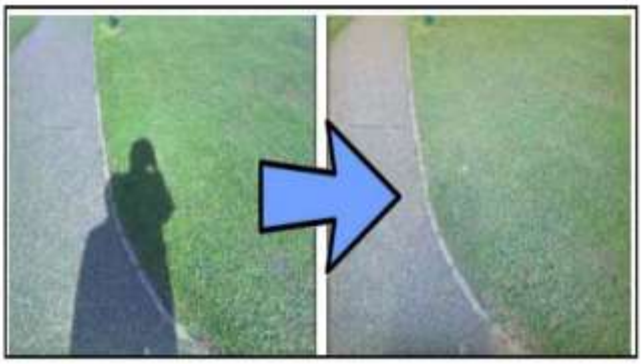

Fig. 2: Shadow Removal from an image

VI. FRAME DIFFERENCE

Under this heading calculation shows the frame difference between the current frame and previous frame, which is to be stored in to the frame buffer. It can be presented as,

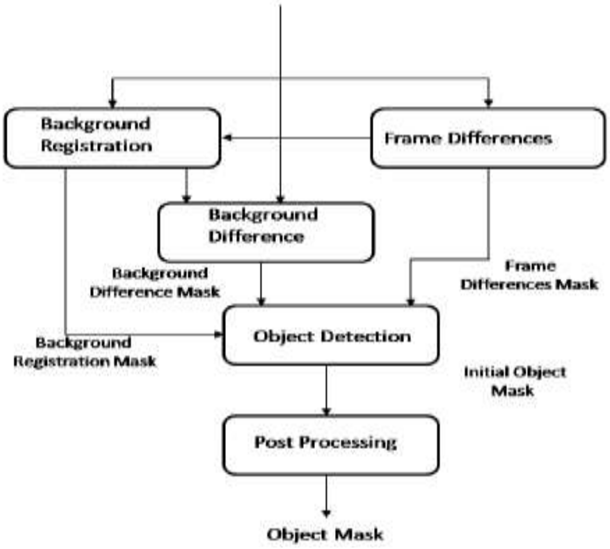

Fig. 3: Block diagram of the baseline mode

$$F_D(x,y,t) = \left| I(x,y,t) - I(x,y,t-1) \right|$$

$$F_{DM}(x,y,t) = \begin{cases} 1 & if \quad F_D \geq T_h \\ 0 & \text{otherwise} \end{cases}$$

Above equation, represent frame data, *FD* shows frame difference and *FDM* represents Frame *Difference Mask*. Here one thing is important that the pixels which belong to the *FDM* are in the category of moving pixels. Here one thing is noticed that the parameters which are required in this case are set in advanced.

VII. Conclusion

This paper presents a detailed method that how video can be used in finding out of minute details in still frames which can be obtained from videos. This paper discuses the baseline model for detecting foreground, shadow and object from sequence of frames. Simulation results are presented by considering a lawn tennis ground. The considered model correctly detects object form a frame. The result obtained in the paper are early results and set directions for the development of a system which can be used for lawn tennis coaching, player and ball tracking. This work provides a methodology about how a mathematical can be used in players tracking in lawn tennis round.

References

[1] J. Choi, Y. Yoo, and J. Choi, “Adaptive shadow estimator for removing shadow of moving object,” Volume 3 Issue 11, November 2014.

[2] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, “Foreground object detection from videos containing complex background,” in Proc. Eleventh ACM Int. Conf. Multimedia, Nov. 2003, pp. 2–10.

[3] Z. Zivkovic, “Improved adaptive Gausian mixture model for background subtraction,” in Proc. IEEE Int. Conf Pattern Recognition, Aug.2004, pp. 28–31.

[4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time tracking of the human body,” IEEE Trans. Patt. Anal. Mach. Intell., vol. 19, no. 7, pp. 780–785, Jul. 1997.

[5] E. Salvador, A. Cavallaro, and T. Ebrahimi. Cast shadow segmentation using invariant color features. CVIU, 95(2):238–259, August 2004.

[6] O. Schreer, I. Feldmann, U. Goelz, and P. Kauff. Fast and robust shadow detection in videoconference applications. In Proc. IEEE VIPromCom, pages 371–375, 2002.

[7] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handlinge. Image and Vision Computing, 24(11):1233–1243, 2006.

[8] J. McHugh, J. Konrad, V. Saligrama, and P.Jodoin, “Foreground-adap- tive background subtraction,” IEEE Signal Process.Letters, vol. 16, no.5, pp. 390–393, May 2009.

[9] M. Vanrell, F. Lumbreras, A. Pujol, R. Baldrich, J. Llados, and J.J. Villanueva. Colour normalisation based on background information. In Proceedings of IEEE International Conference on Image Processing, volume 1, pages 874–877, 2001.

[10] J. J. Veenman, M. J. T. Reinders, and E. Backer. Resolving motion correspondence for densely moving points. IEEE Transactions On Pattern Analysis And Machine Intelligence, 23(1):54–72, 2001.

[11] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. Pattern Recognition, 36(3):585–601, 2003.

[12] P. Spagnolo, T.D Orazio, M. Leo, and A. Distante. Moving object segmentation by background subtraction and temporal analysis. Image and Vision Computing, 24(5):411–423, May 2006.

[13] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for realtime tracking. In IEEE CVPR’99, volume 1, pages 22–29, Ft. Collins, CO, USA, 1999.

[14] F. Sudhir, J.C.M. Lee, and A.K. Jain, “Automatic classifi cation of tennis video for high-level content-based retrieval.” In ***pmceedings of Inremarional workshop on Contenr-Based Access of Image and Video Darabases (CAIVD’98).*** 1998.

[15] P. Chang, M. Han, and Y. Gong, “Extract highlights from baseball game video with hidden markov models,’’ ***inpmceedings*** of ***International Conference on Image Processing (ICIP),*** Rochester,NY, September 2002.

[16] D. Zhang and D. Ellis, “Detecting sound events in basketball video archive,” Tech. Rep., Electrical Engineering Depmment of Columbia university, 2001

[17] Erwin M. Bakker and Michael ***S.*** Lew, “Semantic video retrieval using audio analysis,” in ***proceedings*** of ***Inrermtional Conference on Image and Video Retrieval (CIVR),*** London, UK, July 2002, pp. 271-217.

[18] **H.** Denman, N. Rea, and A. Kokaram, “Content based analysis for video from snooker broadcasts,” in ***proceedings*** of ***Inremrioml Conference on Image and Video Rerrieval (CIVR),*** London.UK, July 2002.

[19] **B.** Moghaddam and A. Pentland, ”Probabilistic visual leaming for object recognition,” ***IEEE Tramactions on Parrern Analyris and Machine Intelligence,*** vol. 19, no. 7, **pp. 6 9 6** 710, Juillet 1997.

[20] D.J.C. MacKay, “Probable network and plausible predictions - a review of practical bayesian methods for supervised neural networks:’ ***Nemork,*** pp. 469-505, 1995.

Utkarsh Shukla, Lecturer, Department of Computer Science, Shri Ramdevi Ramdayal Tripathi Mahila polytechnic, Kanpur, India