Automated Stroke Classification in Tennis

Hitesh Shah[1], Prakash Chokalingam[1], Balamanohar Paluri[1], Nalin Pradeep[1], Balasubramanian Raman[2]

[1] Sarnoff Innovative Technologies, Bangalore - 560025, India
[2] Indian Instititue of Technology, Department of Mathematics, Roorkee - 247667, India
[1]{hshah,cprakash,bpaluri, npradeep}@sarnoff.com
[2]{balarfma}@iitr.ernet.in

Abstract. Stroke recognition in tennis is important for building up statistics of the player and also quickly analyzing the player. It is difficult primarily on account of low resolution, variability in strokes of the same player as well as among players, variations in background, weather and illumination conditions. This paper proposes a technique to automatically classify tennis strokes efficiently under these varying circumstances. We use the geometrical information of the player to classify the strokes. The player is modeled using a color histogram and tracked across the video using histogram back projection. The binarized (segmented) output of the tracker is skeletonized and the gradient information of the skeleton is extracted to form a feature vector. A three class SVM classifier is then used to classify the stroke to be a Forehand, Backhand or Neither. We evaluated the performance of our approach with real world datasets and have obtained promising results. Finally, the proposed approach is real time and can be used with live tennis broadcasts.

Keywords: Skeletonization, Tracking, Optical Flow, Oriented Histogram.

1 Introduction

The increase in sport content archives has boosted the need for automated tools for video content analysis. Large body of literature exist for ball tracking [1], [2], highlight extraction [3], [4], shot classification [5], team behavior classification [6] with applications to sports like Tennis [1], [2], Baseball [4], Basketball [7], Soccer [8], [9], etc.

In this paper, we focus on stroke classification of a player in Tennis game video. We classify a tennis stroke as a forehand, backhand or as a stroke that cannot be classified, representative examples respectively are shown in Figure 1(a,b,c). Such an analysis contributes to better understanding of playing patterns, statistics of the game, and also helps in easy retrieval of game video.

Prior work in stroke classification can be traced to Terrence and Andrew [10] who proposed a technique based on the relative position of the racquet's head to the

player's position and, Miyamori and Iisaku [11] technique used the relative position of the tennis ball to the player's position and also the player's behavior to classify the strokes. However, tracking objects like ball, racquet and head requires high-resolution videos which limits the application of the above approaches.

The proposed technique is based on skeleton features of the player. However, extracting skeletons from low resolution videos require a robust tracking system which can track the player efficiently in different court backgrounds and different illumination conditions. The initialization of the tracker is also crucial. Most of the existing techniques expect the initialization to give a perfect model of the object to be tracked. So any outliers while modeling the tracker will affect the outputs drastically. This has led to the use of hand marked inputs for the tracker, which requires human intervention. In our approach, we counter this problem by using a robust initialization based on optical flow [12]. We have opted the use of a non-parametric tracking approach which is noise resilient and gives us scale and orientation information of the player being tracked.

(a) (b) (c)

Fig. 1. Three Strokes. (a) Forehand, (b) Backhand, (c) Neither

The paper is organized as follows: Section 2 explains the proposed approach, Section 3 details the experimental results on real world data sets, and Section 4 concludes the paper and details the future work.

2 Proposed Approach

The objective of the paper can be formally stated as, given a video taken with a stationary camera behind the player, identify the strokes in each of the frames. The strokes identified by the current system are: Fore hand, Back hand, Neither. The assumption of having a stationary camera is not very restricting as it is the most common camera configuration during the television coverage of tennis.

Figure 2 represents the overall architecture of the proposed system. The main blocks of the system are initialization to identify the player and generate an appropriate model to track the player during the game. Features are then extracted from the tracked player using oriented histogram of player's skeleton. Finally, a Support Vector Machine (SVM) classifier is used to identify the strokes based on the extracted features.

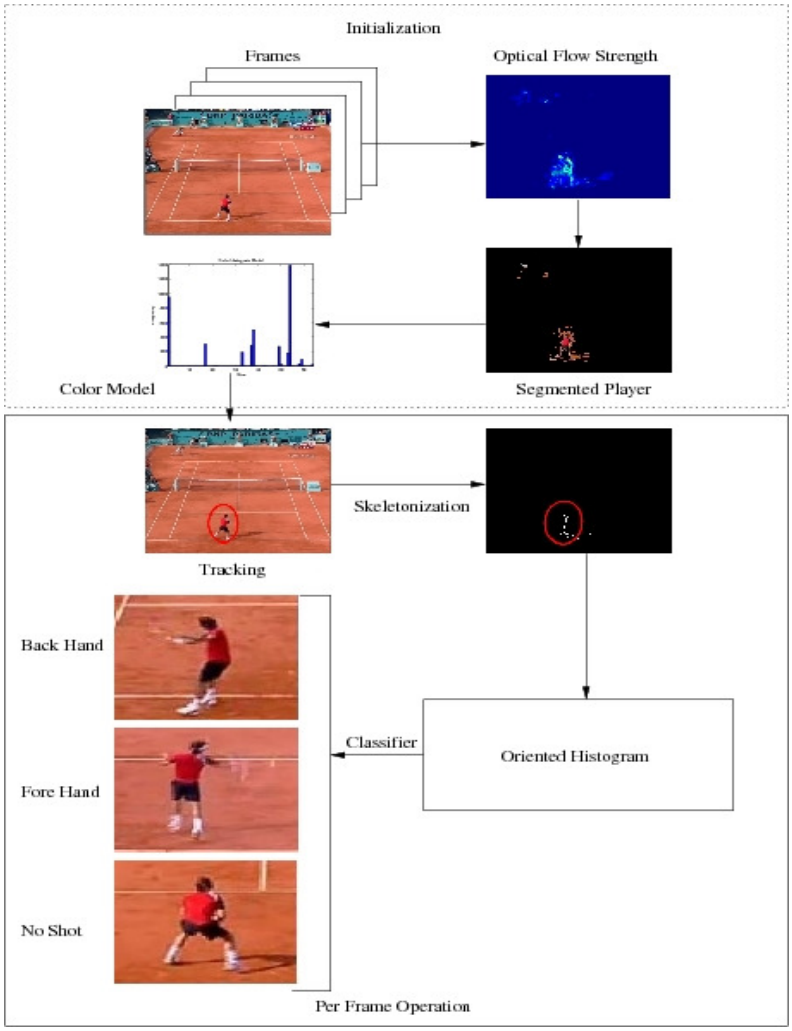


Fig. 2 An Overview of the proposed system

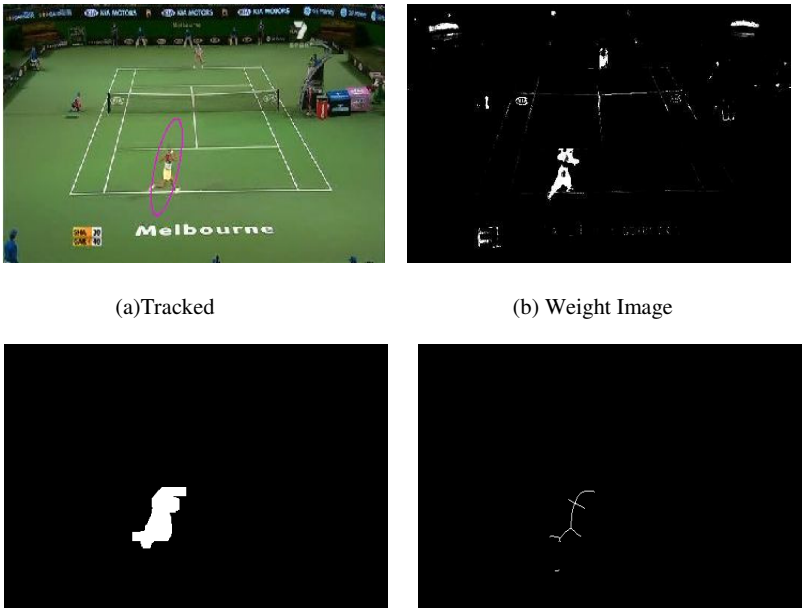
2.1 Initialization

Given a video we need to locate the player. On account of the variability in various features such as clothes color, court color, illumination, non rigid deformation and size of players, we have opted for using motion as a cue to segment the player.

Optical flow is calculated between consecutive frames initially, as proposed in [12] which introduces regularization constraints for spatial coherence along with data fitting to obtain robust pixel motion. The magnitude of the optical flow vector at each pixel describes the displacement of the corresponding pixel. As the two players and

the ball will be the only moving entities in the video, the pixels where they are imaged will have high magnitude of motion (figure 2 shows the magnitude of the optical flow). The image frame is segmented using global adaptive thresholding. In the thresholded image, the largest blob is contributed due to the motion of the player near the camera. The largest blob is located and used to segment the player in the frame.

The segmented player is used to calculate an appearance model to track the player in the subsequent frames. The appearance model is a distribution of colors represented by a histogram **m**, which is compared with a histogram of colors **d** observed within the current frame. The sample weight at each pixel with color **i** is then set to $\sqrt{\frac{m_i}{d_i}}$ a value related to the Bhattacharyya coefficient of histogram similarity. Obtained weight image (figure 3(b)) is then used with a Expectation Maximization (EM) algorithm, an extension of mean-shift algorithm, as proposed by [13]. The EM-like algorithm not only estimates the position of the player but also estimates the covariance matrix that describes the shape of the player. This helps in obtaining an appropriately oriented and scaled ellipse around the player in each frame. (see figure 3(a)). The orientation and the scale of the player is of significant importance in classifying the stroke as explained in the next section. Moreover, the weight image is also used in the feature extraction phase as shall be explained in the next section.



(a)Tracked (b) Weight Image

(c) Blob (d) Skeleton

Fig. 3 Outputs at various stages of the algorithm

2.2 Feature Extraction

Study of tennis videos suggested that body shape of the players can be used as a discriminative feature for identifying the strokes. However, naive approaches like template matching, contour matching cannot be applied directly due to the subtle differences in the players, their orientation while playing strokes, and the scale of the player in the video. To address the above problems, we have proposed the use of oriented histogram of the skeletonized binary images of the player.

The weight image, calculated during the tracking phase, inside the tracked ellipse is thresholded to obtain a binary image of the body of the player (figure 3(c)). We use the technique proposed by [14] to obtain skeleton of the player (figure 3(d)). The algorithm computes curve skeletons from binary images by computing a potential field over the entire object. The potential at any object pixel (x, y) is defined as

$$\phi = \sum_{i=0}^{N} \frac{1}{r_i^m} \tag{1}$$

where $r_i = \left[\left(x - x_i \right)^2 + \left(y - y_i \right)^2 \right]^{\frac{1}{2}}$ and **N** is the number of points on the boundary of the object. The gradient of the potential field at each object pixel is then calculated to form a vector field. This vector field is then analyzed to extract the seed points which start from high curvature values, high divergence values and saddle points.

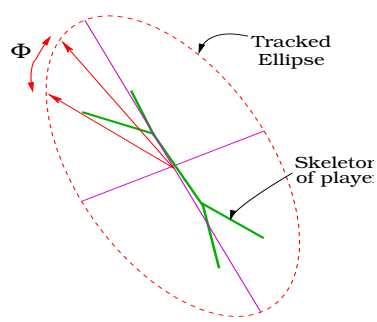


Fig. 4 Tracked Ellipse along with the skeleton

The gradient vector of the potential field starting from the seed point converges at critical points which forms the skeleton of the object. The skeletonization algorithm is very efficient in generating smooth skeleton and is robust to noise, which is one of our major requirements.

The skeletonization helps us overcome the differences in the body sizes of the players but not the rotation and scale factors. For rotation invariance we use the oriented ellipse obtained during the tracking phase to align the players to common reference axis (minor axis of the ellipse).

Finally, to achieve scale invariance, we take a oriented histogram of the skeleton inside the tracked ellipse. The oriented histogram is calculated by dividing the ellipse

into **n** sectors of equal angle (ϕ) as shown in figure 4 and calculating the number of skeleton points in each of the sectors. We normalize the histogram by dividing each bin value by the sum of all bin values, this is important to have scale invariance. The sample histograms of a Forehand, Backhand and Neither strokes are shown in figure 5. The histogram is used as the feature vector and a trained SVM classifier is used to classify it.

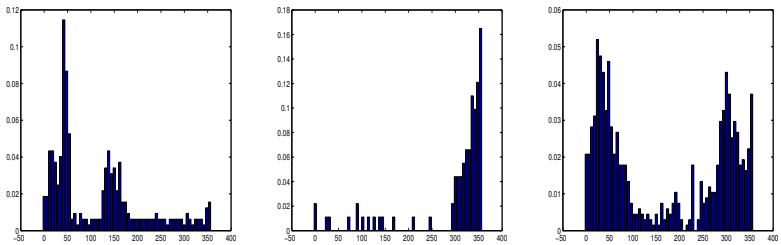


Fig. 5 Histograms. (a) Forehand (b) Backhand (c) No Stroke

2.3 Classification

For classification we have tried with classifiers like correlation, Bhattacharya coefficient and nearest neighbour techniques along with SVM [15]. We have observed SVM classifier to outperform the rest. SVM's performance can be attributed to its ability of mapping data to a high dimensional space where classes are linearly separable by hyper planes.

We trained a SVM classifier which classifies the feature vector (histogram) into one of the three classes: Forehand, Backhand and Neither. The SVM with an Gaussian Radial Basis Function (RBF) kernel was trained. The values of the parameters sigma σ and c were set to 0.1 and 10 respectively. We trained the SVM classifier with 500 histograms of each class.

3 Experimental Results

We have evaluated the performance of the proposed approach on a variety of videos. The dataset videos consisted of tennis games on grass, clay, artificial grass courts as well as had both male and female players. The player is initially segmented based on optical flow. The parameters used in optical flow are the search window size, gaussian blurring window size. We experimentally found that optical flow outputs were best for a search window size of 5 and gaussian tap size of 7. We also used a look up table for eigen calculations to speed up the computation.

The player is then modeled using a color histogram. We used 64 bins, 4 bins for each of the Red, Green and Blue colors. By increasing the number of bins we can have a rich representation of the object to be tracked but it is computationally expensive. Moreover, the players to be tracked are homogeneous in nature. Thus 64 bins were found sufficient for accurate tracking of the player. The histogram is used

to generate the weight image as explained in the earlier section. EM-like algorithm is used to fit an appropriate ellipse. Mean indicates the position and variance indicates the scale and orientation of the ellipse used in tracking. In our experiments, we found that the EM algorithm takes 6 iterations to converge.

The weight image obtained is thresholded using a global adaptive threshold to obtain a player blob inside the ellipse. The threshold was taken as 80% of the value of the maximum weight value. A Closing operation is then performed with a circular structured element of radius 5 units. This is done so that the skeleton outputs are robust to variations across videos.

The oriented histogram accounting for each point in the skeleton with respect to the minor axis of the tracked ellipse is computed. The histogram obtained varies from being generic to specific with increase in the number of bins respectively. After thorough experimental evaluation the number of bins were fixed to 60, i.e. it maps 6 degrees into one bin. The trained svm classifier [16] classifies the 60 length normalized feature vector (histogram values).

The system performs at a speed of 20 fps, on a Pentium 4, 3.4 GHz processor. The speed of the method is directly dependent on speeds of the tracker, skeletonization and svm classification. The tracker speed depends on the size of the object to be tracked, which is same for most of the videos. The same applies to the skeleton computation. The speed of the svm classifier is dependent on the size of the feature vector to be classified which is fixed to 60. Hence, the system performs at the same fps for most of the tennis videos.

Table 1 shows the recognition accuracies of the proposed approach on 150 different games played in different tennis courts across different players. Each game was clipped to segments of 10 minute duration. For generating the ground truth, each of the frames in the segment was manually labeled according to the shot. Our experiments showed high recognition accuracies for low resolution real world datasets under different conditions as shown in figures 6, 7, and 8.

Table 1: Recognition Accuracies in Percent

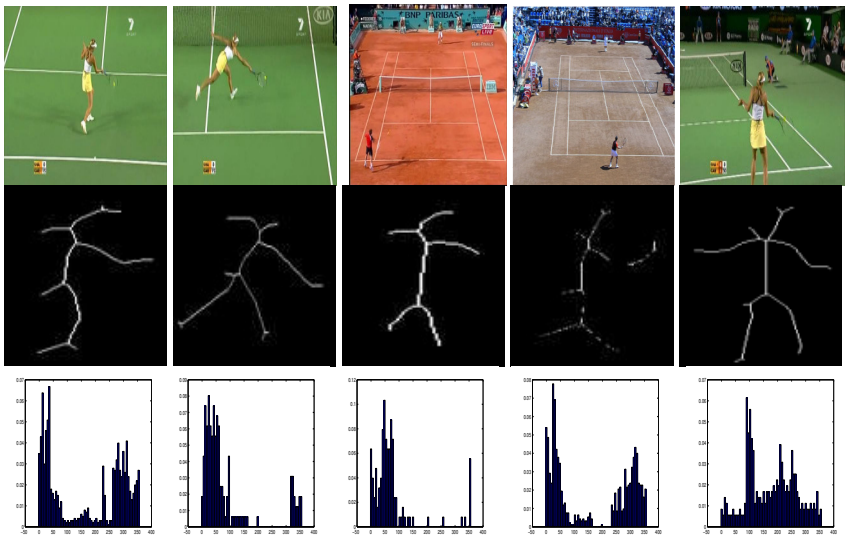
Stroke	ForeHand	BackHand	No Stroke
ForeHand	97.24	0.44	2.32
BackHand	0.56	96.43	3.02
No Stroke	1.18	0.79	98.02

Figures 6, 7, and 8 shows the frames, their corresponding cropped skeleton images and oriented histograms of ‘forehand’, ‘backhand’ and ‘no stroke’ respectively. The example figures have been taken to illustrate the variations in scale, orientation, tennis courts and position of the players. Under these challenging variations, the skeletons and oriented histograms of the tracked player are classified with good accuracy as in Table 1. For instance, in Figure 6(b), the player is classified correctly even while stretching for the ball, oriented downwards. Also figures 7(a), 7(d) are classified similarly with different orientations. Figures 6(c), 6(d), 7(a), 7(b) show the forehand and backhand classification succeed for cameras placed far off thereby not getting affected to the player scales. In figure 7(e), in doubles match, the tracker helped in getting accurate skeletons of the player and was classified as ‘backhand’. Similarly, it can be observed for ‘No Stroke’ as shown in figure 8. Figure 6(e) is one of the failure

cases, we believe that the failure can be attributed to the fact that the player has both the hands stretched, hence classifying it as 'No Stroke'.

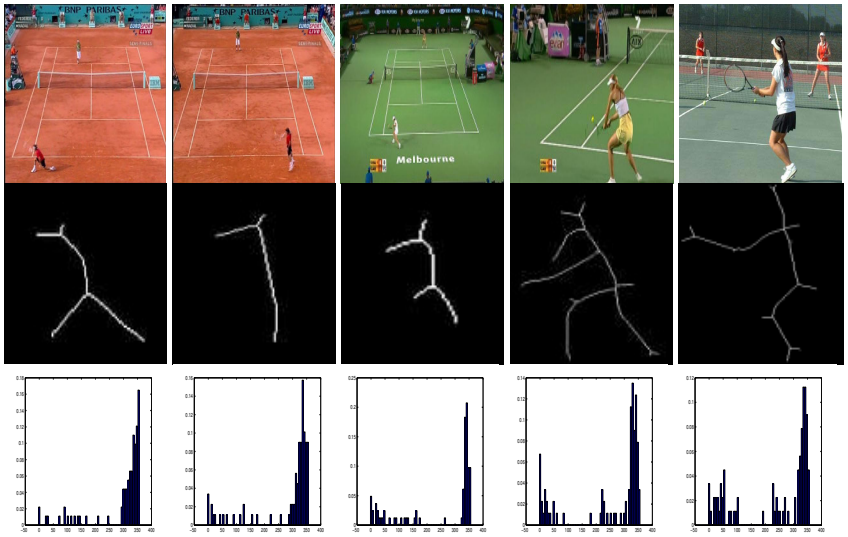
4 Conclusion

An automated stroke classification system for real world tennis videos has been proposed. Motion vectors are used to automatically model the player, structural and geometrical information of the player tracked across the video are used as features. The features are invariant to intra and inter class variations of players strokes. A trained SVM classifier is used to recognize the strokes of the player. In future a possible application of the proposed system is to automatically annotate a player with metadata to generate statistics and also aid in video search capabilities. Recognizing other tennis strokes such as slice, volley, serve using more semantic actions from the videos will be focused on.



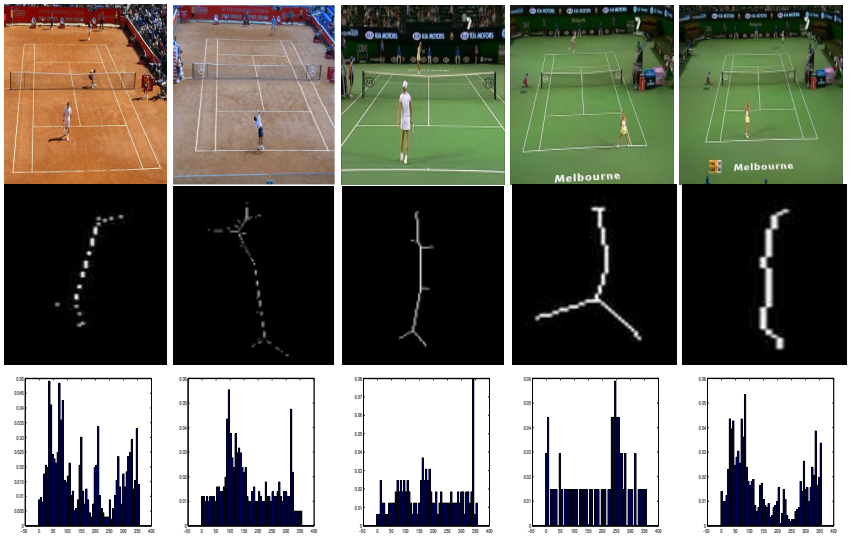
(a) **(b)** **(c)** **(d)** **(e)**

Fig. 6 'Forehand' Strokes (a,b,c,d) and (e) 'No stroke' with Zoomed Skeleton images and Histograms



(a) **(b)** **(c)** **(d)** **(e)**

Fig. 7 'Backhand' Strokes with Zoomed Skeleton images and Histograms



(a) **(b)** **(c)** **(d)** **(e)**

Fig. 8 'No Strokes' with Zoomed Skeleton images and Histograms

References

1. Fei Yan, Alexey Kostin, William Christmas, and Josef Kittler, "A novel data association algorithm for object tracking in clutter with application to tennis video analysis," in CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2006, pp. 634–641, IEEE Computer Society.
2. G. S. Pingali, Y. Jean, and I. Carlbom, "Real-time tracking for enhanced tennis broadcasts," in CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 1998, p. 260, IEEE Computer Society.
3. A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," Multimedia, IEEE Transactions on, vol. 7, no. 6, pp. 1114–1122, 2005.
4. Yong Rui, Anoop Gupta, and Alex Acero, "Automatically extracting highlights for tv baseball programs," in MULTIMEDIA '00:Proceedings of the eighth ACM international conference on Multimedia, New York, NY, USA, 2000, pp. 105–115, ACM Press.
5. Ling-Yu Duan, Min Xu, Xiao-Dong Yu, and Qi Tian, "A unified framework for semantic shot classification in sports videos," in MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia, New York, NY, USA, 2002, pp. 419–420, ACM Press.
6. Christian Thurau, Thomas Hettenhausen, and Christian Bauckhage, "Classification of team behaviors in sports video games," in ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Washington, DC, USA, 2006, pp. 1188–1191, IEEE Computer Society.
7. Wensheng Zhou, Asha Vellaikal, and C. C. Jay Kuo, "Rule-based video classification system for basketball video indexing," in MULTIMEDIA '00: Proceedings of the 2000 ACM workshops on Multimedia, New York, NY, USA, 2000, pp. 213–216, ACM Press.
8. A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," in Proceedings of IEEE Transactions Image Processing, 2003, pp. 796– 807.
9. Sachiko Iwase and Hideo Saito, "Parallel tracking of all soccer players by integrating detected positions in multiple view images," in ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 4, Washington, DC, USA, 2004, pp.751–754, IEEE Computer Society.
10. T. Bloom and P. Bradley, "Player tracking and stroke recognition in tennis video," in Proceedings of the WDIC, 2003, pp. 93–97.
11. H. Miyamori and S. ichi Iisaku, "Video annotation for content-based retrieval using human behaviour analysis and domain knowlegde," in Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 320–325.
12. M.J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in ICCV '93:, 1993, pp. 231–236.
13. Z.Zivkovic and B. Krose, "An em-like algorithm for color-histogram-based object tracking," in Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2004
14. Yatharth Saraf, R. Balasubramanian and K. Swaminathan**, "**A classical approach for thinning of binary images using divergence of the potential field", *International Journal of Computer Mathematics*, Vol. 82**,** No.6, pp. 673-684, 2005.
15. V Vapnik, "The nature of statistical learning theory", Springer, New York, 1995.
16. O Chapelle and V N Vapnik, "Model selection for support vector machines," in Advances in Neural Information Processing Systems 12, MIT press, pp 230-236, 2000.