A PINHOLE CAMERA MODELING OF MOTION VECTOR FIELD FOR TENNIS VIDEO ANALYSIS[*]

Peng Wang, *Rui Cai*, *Bin Li*, *and Shi-Qiang Yang*

Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China

ABSTRACT

Motion Vector Field (MVF) represents motion characteristics in video sequences, and has been widely used and been proved to be effective in sports video analysis. However, in tennis video analysis, MVF-based methods are seldom utilized for the reasons that (i) the players' true motion is not accurately represented by the extracted motion vector, due to the deformation caused by the diagonal shooting of the camera, and (ii) the motion vector's magnitude is not prominent enough and is prone to be disturbed by noises. In this paper, a pinhole camera modeling of motion vector field is proposed to revise the deformed motion vector. In this modeling, a foreground object mask is adopted and global motion compensation is incorporated as the pre-processing steps. Evaluation of the proposed modeling using four hours of tennis videos shows very encouraging results.

1. INTRODUCTION

As one of the most salient visual characteristics, motion is crucial for content-based sport video analysis. The motion vector extracted from compressed video bit-stream reflects the displacement of a macro block, and the collective set of all motion vectors in a video frame is called Motion Vector Field (MVF). Most of current motion features used in sports video analysis are based on MVF. Duan et al. [1] give a comprehensive summarization of MVF-based mid-level representation as well as corresponding implementations on various sports games. Also based on MVF, Ma [2] calculates the motion energy spectrum for video retrieval, and a motion energy redistribution function is proposed in [3] for motion events recognition. It is shown that the MVF-based analysis methods have the advantage of efficient computation and effectual performance for most generic applications.

However, MVF-based methods are seldom utilized in tennis video analysis. Conventional methods mainly focus on detecting and tracking of players or balls in the image sequence [4] [5], as well as incorporating with human gesture and behavior analysis [6]. Although these computer vision related methods may provide more elaborate annotation of tennis game, they result in complicated implementation, inflexible utilization and nontrivial limitation. With our investigation, two main reasons baffle the MVF utilization in tennis video. Firstly, because the shooting camera is diagonal but not perpendicular to the court plane, MVF can not correctly represent the players' true motion, especially the motion in the vertical direction. The magnitude of motion vector is reduced and the orientation is distorted, and the deformation is particularly evident for the player at the top half court. Secondly, the court with homologous color may introduce random noises when estimating motion vectors. Furthermore, the players' motion is not prominent enough, and thus the estimated players' motion vectors are often unreliable. To utilize MVF in tennis analysis, two issues must be resolved: (i) revise the motion vectors according to players' actual motion, and (ii) reduce the estimation noises.

In this paper, a *Pinhole camera Modeling of Motion Vector Field* (PMoMVF) for tennis video analysis is proposed to revise the original motion vectors. To reduce the noises introduced by motion estimation or slight camera panning, foreground object mask and global motion compensation are incorporated as preprocessing steps. In order to verify the proposed pinhole modeling, classification of players' basic actions in tennis video is carried out by generating the temporal motion curves, which have been successfully used in previous work [3]. Experimental results on recorded tennis videos demonstrate the effectiveness and efficiency of the proposed modeling.

The rest of this paper is organized as follows. Section 2 will present the motion vector transformation by utilizing the pinhole camera model. Section 3 will describe how to improve the reliability of the transformation by using foreground object mask and global motion compensation. In Section 4, the application of the PMoMVF for classifying tennis players' basic actions will be introduced. Experiments and discussion will be given in Section 5. Finally, Section 6 presents the conclusion and future works.

2. MOTION VECTOR TRANSFORMATION

Diagonal shooting camera in tennis game is usually placed right above the vertical symmetrical axis of the tennis court, thus the court in rectangle becomes an isosceles trapezoid, as shown in Fig. 1. The players' movements in tennis video are also distorted and the motion vector estimated from video sequence cannot correctly reflect the true motion. As illustrated in the left part of Fig. 1, a motion vector can be denoted as the displacement from a given point p_1 in current frame to its corresponding point q_1 in the next frame. If the player is watched moving from p_1 to q_1 in video frame, the real movement in tennis court should be from p_2

[*] This work was supported by the National High Technology Development 863 Program of China, and the National Grand Fundamental Research 973 Program of China

0-7803-8554-3/04/$20.00 ©2004 IEEE.

to q_2. Not only the motion's magnitude is reduced in video frame, but also the orientation is distorted. The distortion of vertical motion is especially significant, and there is always $y_1<y_2$ in Fig.1. Such deformations make it difficult to analyze the players' true motion, for instance, we can hardly tell whether the player is taking the net or not, directly depending on the vertical projection of the motion vector. However, this task would become feasible if the motion vector can be revised according to the true motion in tennis court plane.

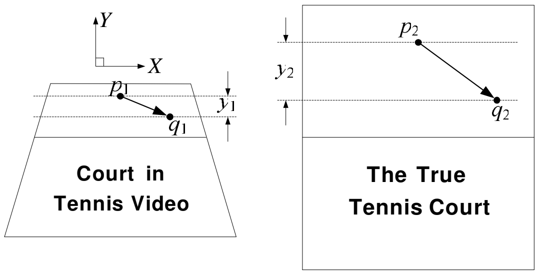


Fig. 1. Illustration of the motion vector deformation

In fact, for point p_1 and q_1, if the corresponding point p_2 and q_2 can be correctly located in tennis court plane, the transformation of motion vector is achieved. Thus the essential problem is that for any given point in video frame, how to find the corresponding point in the tennis court plane. In order to perform this task, a pinhole camera modeling is employed in this paper. As illustrated in the top left part of Fig. 2, for a pinhole camera, there is

$$L/l = u/f \tag{1}$$

where u and f denote the object distance and the camera focus, L and l are the lengths of object and image respectively. Supposing the horizontal distance between the camera and the bottom baseline of tennis court is d, and the height of camera from ground is h, with Eq. (1), there are

$$\begin{aligned} W/w_1 &= \sqrt{d^2+h^2}\Big/f \\ W/w_2 &= \sqrt{(H+d)^2+h^2}\Big/f \\ W/w_3 &= \sqrt{(2H+d)^2+h^2}\Big/f \end{aligned} \tag{2}$$

Here W and H denote the width and half height of the true tennis court [9], and w_1, w_2, w_3 respectively represent the lengths of the bottom baseline, net line and top baseline in the image plane, as shown in Fig. 2.

For any given point p' in the trapezoidal court in image plane, the line passing through p' and being parallel with the baselines is segmented by p' and the two court sidelines into two parts, whose lengths are denoted as w_{x1} and w_{x2} respectively. The position of p' is uniquely represented by w_{x1} and w_{x2}. Supposing p is the corresponding point of p' in the true tennis court plane, and p is uniquely represented by x_1, x_2, y, which denote the distances between p and the sidelines and bottom baseline respectively, as illustrated in Fig. 2. The relations between (x_1, x_2, y) and (w_{x1}, w_{x2}) are

$$\begin{cases} W/(w_{x1}+w_{x2}) = \sqrt{(y+d)^2+h^2}\Big/f \\ w_{x1}/w_{x2} = x_1/x_2 \quad \text{and} \quad x_1+x_2=W \end{cases} \tag{3}$$

With Eq. (2), parameters d and h can be solved, thus for a given point in video frame, the position of the corresponding point in the true tennis court plane can be directly calculated through Eq. (3). With the point transformation functions, the two end points of motion vector are transformed to the true tennis court plane first, and then the new motion vector is calculated by taking the difference between two transformed end points.

In most of the video shots of a tennis game, the shooting camera is usually appropriately placed and our assumption is approximately justified. With the robust line detection algorithm proposed in [7], the exact position of the trapezoidal court in tennis video, including the lengths of the borders and the coordinates of the corners, can be obtained through averaging the line detection results in several beginning frames of the game shot. When the position information of the trapezoidal tennis court is obtained, all motion vectors in *Player Active Area* are transformed to the true tennis court plane in experiments. The *Player Active Area* is defined as a larger isosceles trapezoid covering the tennis court in the image plane, as the trapezoid in dash-dot line shown in Fig. 2.

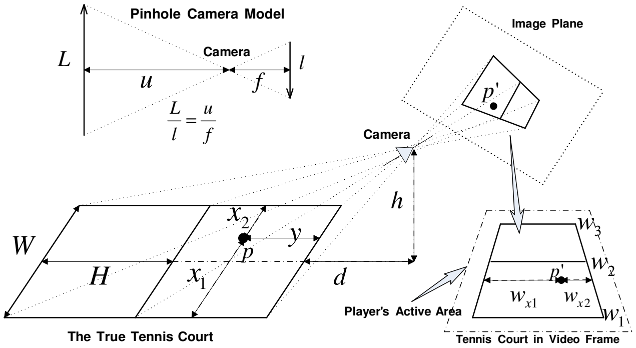


Fig. 2. Pinhole camera modeling based transformation

3. FOREGROUND OBJECT MASK

In tennis game, the court with homologous color often causes random noises in motion vector estimation when using the block-matching algorithm. And the players in tennis video are relatively small in proportion to the whole image frame, thus these random noises can not be neglected since they will interfere the estimation of the players' motion vectors in practice. A straightforward solution to obtain more reliable players' movement is to take into account of motion vectors in players' areas, which usually are foreground objects in tennis video, and discard all others. Besides, the shooting camera sometimes may slightly pan with the moving players, and will introduce additional global motion component to the estimated motion vectors. The proposed camera modeling can still hold valid if we can compensate the global motion component to characterize the players' true movements more accurately.

Actually, technologies of foreground object segmentation and global motion compensation have been widely used in object-based video coding, and can be used to resolve the mentioned problems. To locate players' areas and compensate global motion component, the fast algorithm proposed in [8] is adopted in this paper. The global motion between two images is represented by the six parameter affine model, whose parameters can be estimated using the Gauss-Newton or the Levenberg-Marquadet iterative calculation. In each iteration step, pixels of foreground object will be excluded by using the residual-block based outlier rejection, as well as the global motion is re-estimated. Finally the *Foreground Object Mask* (FOM) and

Global Motion Component (GMC) can be obtained simultaneously. Fig. 3 illustrates an example, where the two players' areas in the left original image could be coarsely located in the right image with FOM. The threshold controlling the outlier rejection is experientially set to 10% in experiments. More details could be found in [8].

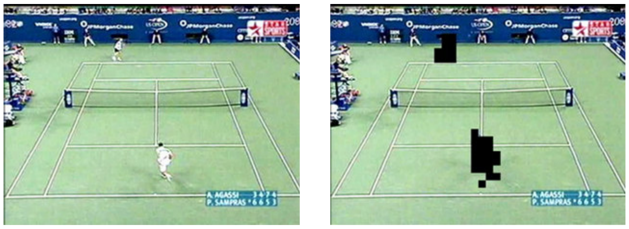
The obtained FOM has the same dimension with that of the original MVF, and can be denoted as

$$FOM = \left(f_{i,j}\right) \tag{4}$$

where $f_{i,j}$ is either 0 or 1. Value 0 denotes the macro-block (*i,j*) is estimated as the background and value 1 denotes it is part of the foreground object. Similarly, the GMC is denoted as

$$GMC = \left(g_{i,j}\right) \tag{5}$$

where $g_{i,j}$ denotes the estimated global motion component for macro-block (*i,j*) to be compensated.



(a) (b)

Fig. 3. Original image (a) with foreground object mask (b)

4. CLASSIFICATION OF PLAYER'S BASIC ACTIONS

In this section, the classification of players' basic actions based on the proposed PMoMVF is introduced. Firstly, the flowchart of PMoMVF is illustrated in Fig.4. For one game related shot (G-shot), which can be extracted from the tennis video by using color based selection algorithm [4], the original MVF is extracted as input of the system. As Fig.4 shows, the PMoMVF consists of two stages. In the ***P*** (Pre-processing) stage, FOM and GMC are obtained through global motion estimation and performed on the original MVF as

$$P(MVF) = (MVF \oplus GMC) \otimes FOM \tag{6}$$

where $\oplus$ indicates the matrix addition and $\otimes$ indicates the entry-by-entry product of matrix. In the ***T*** (Transformation) stage, with the court line detected in G-shot, corresponding transformation functions are set up following the methods proposed in Section 2.

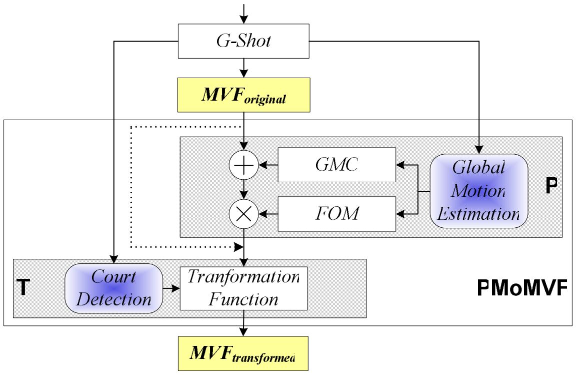


Fig. 4. Illustration of the flowchart of PMoMVF

Two player's basic actions are classified in this paper: net game and baseline game. Net game is that the player moves into the forecourt and toward the net to hit volleys. Baseline game is that the player hits the ball from near the baseline, against the net game. In this paper, the MVF-based method [3] for sports event classification is utilized to classify player's basic actions. As described in [3], the energy redistribution function is implemented on the input MVF first, and then convolution with weight templates is performed as filtering for certain motion pattern. In experiments, the horizontal and vertical motion filters are employed to generate two temporal motion curves. Then these curves as features are used to classify player's actions by Hidden Markov Models.

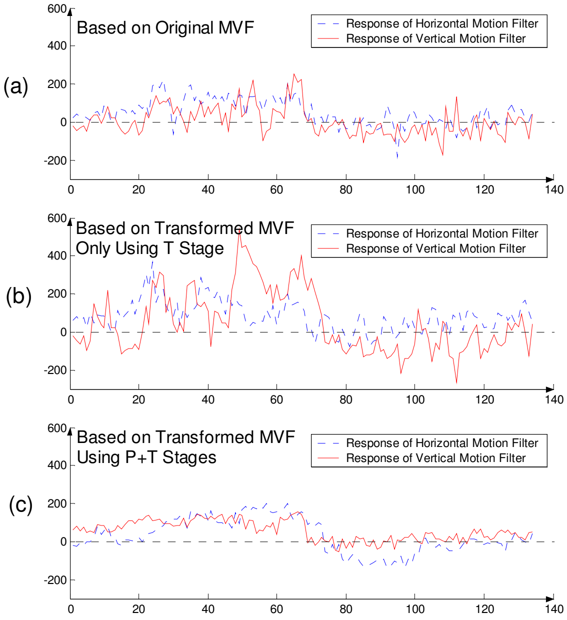


Fig. 5. Comparison of temporal motion curves of vertical and horizontal motion filters based on different MVFs

With the detected net line, the obtained MVF is divided into two parts, for classifying the basic actions of players in top half court and bottom half court respectively. In experiments, the two temporal motion curves are calculated for comparison based on (i) the original MVF, (ii) the transformed MVF only using ***T*** stage (following the dot-line arrow in Fig. 4) and (iii) the transformed MVF with the whole PMoMVF. Fig. 5 shows an example of the two temporal motion curves of net game in the top half court. The *X* axis denotes the frame number and the *Y* axis denotes the calculated motion response value. Positive value on vertical motion curve means movement to bottom in the image plane, and movement to right for the horizontal motion curve. From frame 1 to 80, the player runs to take the net from the left end of the baseline toward the right end of net, then from frame 81 to 134, the player walks back from the net to the right end of the baseline. As shown in Fig. 5 (a), both curves are quite noisy, and the vertical motion curve is too irregular to characterize the net game. In Fig. 5 (b), the responses of horizontal and vertical motion filters are both enlarged, and the segment representing the net approach is more evident, however, the noises are still prominent. Fig. 5 (c) gives the best result relatively, with more clear motion trend and less noise.

5. EXPERIMENTS

Four hours recorded live tennis videos are used in experiments to validate the performance of the proposed PMoMVF in tennis video analysis. The experimental video data are collected from the matches of A. Agassi and P. Sampras at US Open 2002 (*Video*$_1$), and R. Federer and M. Philippoussis at Wimbledon 2003 (*Video*$_2$). As ground truth, the G-shot containing net game segment is labeled with *Net game Shot* (NS), otherwise it is labeled with *Baseline game Shot* (BS), for the two players respectively. The detail information of the selected video data is listed as follows.

Table 1. Information of the experimental video data

Video	#Shot	#G-shot	Top half		Bottom half	
			#NS	*#BS*	*#NS*	*#BS*
Video$_1$	881	316	58	258	230	86
Video$_2$	624	271	100	171	114	157
∑	1505	587	158	429	344	243

For player in certain half court, two HMMs for the net-game segment and baseline-game segment in G-shot are respectively built, and they are then circularly connected to construct a higher-level HMM, which represents the transition between net-game and baseline-game within a G-shot. The transition probabilities to the two sub-HMMs are both set to 0.5. Half of the experimental data are selected randomly as training data, and each G-shot in training set is further divided into net-game segments and baseline-game segments. In recognition, all G-shots with net-game segment detected are considered as NS, and others are BS.

For comparison, experiments are performed based on the original MVF, the transformed MVF by only using the T stage, and the transformed MVF with the whole PMoMVF respectively, and results are listed in Table 2. When using the original MVF, the vertical motion response between net-game and baseline-game can not be effectively distinguished, as shown in Fig. 5 (a), and many of the baseline games are misclassified into net games. Furthermore, performance for player in top half court is greatly lower than that of the player in bottom half, for that the deformation of motion vector in top half is more evident. With the transformed MVF by only using ***T*** stage, performances are improved notably, especially for the net-game classification of the player in top half court, whose precision and recall ratios are both doubled. Finally, the transformed MVF with the whole PMoMVF gives more encouraging results, which could be accepted in practical applications.

Table 2. Experimental results based on the original MVF, the transformed MVF by only using the *T* stage, and the transformed MVF with the whole PMoMVF

MVF	Shot	Top Half Court		Bottom Half Court	
		Pre.(%)	Rec.(%)	Pre.(%)	Rec.(%)
Original	*NS*	30.91	37.78	64.55	67.78
	BS	70.53	63.81	47.75	44.17
Using ***T*** Stage	*NS*	61.11	73.33	85.06	82.22
	BS	87.50	80.00	74.60	78.33
Using PMoMVF	*NS*	80.00	66.67	90.85	82.78
	BS	86.67	92.86	77.21	87.50

Sometimes the block-based estimation of original MVF has unavoidable mistakes and errors, under which the misclassified results are unable to be corrected even by the PMoMVF. However, experimental results indicate that in most conditions of tennis videos, the PMoMVF can properly revise the deformed MVF as well as reduce the noises, thus enable those MVF-based methods to be feasible in tennis video analysis.

6. CONCLUSION

A *Pinhole camera Modeling of Motion Vector Field* for content based tennis video analysis is proposed in this paper. In this modeling, the original deformed motion vectors are revised according to players' true motion, and the foreground object mask and global motion compensation are incorporated as preprocessing steps for noise reduction. Experiments on classification of players' basic actions show very encouraging results. The future works include: (i) make some steps more robust in setting up the model, such as the location of court lines, and (ii) try more applications in tennis analysis with the PMoMVF.

Acknowledgements We would like to express our appreciation to Prof. Jeng-Neng Hwang of University of Washington, who gave us valuable advices to this paper.

7. REFERENCES

[1] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu, "A Mid-level Representation Framework for Semantic Sports Video Analysis", *Proc. of the 11th ACM International Conference on Multimedia*, pp.33-44, Berkeley, CA, USA, Nov. 2-8, 2003.

[2] Y.-F. Ma and H.-J. Zhang, "A New Perceived Motion based Shot Content Representation", *Proc. of IEEE International Conference on Image Processing*, Thessaloniki, Greece, Vol.3, pp.426-429, Oct. 7-10, 2001.

[3] G. Xu, Y.-F. Ma, H.-J. Zhang, and S.-Q. Yang, "Motion based Event Recognition Using HMM", *Proc. of the 16th International Conference on Pattern Recognition*, Quebec, Canada, Vol.2, pp.831-834, Aug. 11-15, 2002.

[4] G. Sudhir, John C.M. Lee, and Anil K. Jain, "Automatic Classification of Tennis Video for High-Level Content-Based Retrieval", *Proc. of 1998 International Workshop on Content-Based Access of Image and Video Databases*, pp.81-90, Bombay, India, Jan. 03-03, 1998.

[5] G.S. Pingali, Y. Jean, and I. Carlbom, "Real Time Tracking for Enhanced Tennis Broadcasts", *Proc. of 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.260-265, Santa Barbara, CA, USA, Jun. 23-25, 1998.

[6] H. Miyamori, "Improving Accuracy in Behaviour Identification for Content-based Retrieval by Using Audio and Video Information", *Proc. of the 16th International Conference on Pattern Recognition*, Vol.2, pp.826-830, Quebec, Canada, Aug. 11-15, 2002.

[7] H.-J. Di, L. Wang, and G.-Y. Xu, "A Three-step Technique of Robust Line Detection with Modified Hough Transform", *Proc. of 3rd SPIE Symposium on Multi-spectral Image Processing and Pattern Recognition*, pp.835-838, Beijing, China, Oct. 20-22, 2003.

[8] Y.-W. He, B. Feng, S.-Q. Yang, and Y.-Z. Zhong, "Fast Global Motion Estimation for Global Motion Compensation Coding", *Proc. of IEEE International Symposium on Circuits and System*, Vol.2, pp.233-236, May 6-9, 2001.

[9] http://www.hickoksports.com/glossary/gtennis.shtml