Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine

Guangyu Zhu[1], Changsheng Xu[2], Wen Gao[1,3], and Qingming Huang[3]

[1] Harbin Institute of Technology, Harbin, P.R. China
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore
[3] Graduate School of Chinese Academy of Sciences, Beijing, P.R. China
{gyzhu, wgao, qmhuang}@jdl.ac.cn,
xucs@i2r.a-star.edu.sg

Abstract. Motion analysis in broadcast sports video is a challenging problem especially for player action recognition due to the low resolution of players in the frames. In this paper, we present a novel approach to recognize the basic player actions in broadcast tennis video where the player is about 30 pixels tall. Two research challenges, motion representation and action recognition, are addressed. A new motion descriptor, which is a group of histograms based on optical flow, is proposed for motion representation. The optical flow here is treated as spatial pattern of noisy measurement instead of precise pixel displacement. To recognize the action performed by the player, support vector machine is employed to train the classifier where the concatenation of histograms is formed as the input features. Experimental results demonstrate that our method is promising by integrating with the framework of multimodal analysis in sports video.

1 Introduction

The motion or action performed by players in tennis game can reveal the process of the game and the tactics of the players. It is essential for analysis of the matches and desired for sports professionals and longtime fanners for technical coaching assistant and tactics analysis.

Considering the appearance ratio of playfield in one frame shown in Fig. 1, the frames/shots of broadcast tennis video can be divided into two classes: close-up view where the magnitude of player figure is higher, and far-view where the magnitude is lower. In close-up, a player figure is usually 300 pixels tall. It is easy to segment and label human body parts such as the limbs, torso, and head resulting in marking out a stick figure. Existing work [1][2] has achieved good results on action recognition for close-up view. On the other hand, in far-view frame, a player figure might be only 30 pixels tall. The action detail of the player is blurred due to the low figure resolution. In this case, we can only track the player as a blob and extract the spatial translation. It is very difficult to articulate the separate movements of different body parts. Thus we cannot discriminate among too many action categories for the far-view video. To the best of our knowledge, there are few efforts devoted in the research of tennis player action

T.S. Huang et al. (Eds.): HCI/ECCV 2006, LNCS 3979, pp. 89–98, 2006.
© Springer-Verlag Berlin Heidelberg 2006

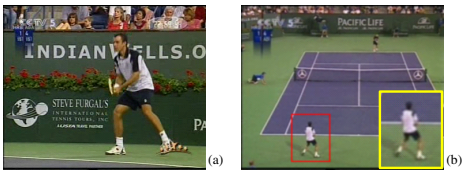

Fig. 1. Two typical frames derived from broadcast tennis video. (a) Close-up scene, (b) Far-view scene, the zoomed picture is the player whose action to be recognized.

recognition in broadcast video. Miyamori and Iisaku [3] developed an automatic annotation system of tennis actions including foreside-swing, backside-swing and over-the-shoulder swing. The analysis is based on silhouette transitions. However, appearance is not necessarily preserved across different sequences and less robust for classification. Compared with action recognition for the videos with high resolution figures, a little work [4][5] is attempting to analyze poor quality, non-stationary camera footage. The approach proposed in [4] modeled the structure of the appearance self-similarity matrix and was able to handle very small objects. Unfortunately, this method was based on periodicity and thus restricted to periodic motion. Efros *et al.* [5] developed a generic approach to recognize actions in "medium field" which is similar to "far-view" defined in this paper. In their paper, a motion descriptor based on optical flow in a spatio-temporal volume was introduced and an associated similarity measure was utilized in a nearest neighbor classification (NNC) framework to perform action categorization. The experimental results on tennis data set are promising. However, the videos they used are non-broadcast which has less challenge for tracking and recognition.

In this paper, we propose a novel motion analysis approach to recognize the player actions in far-view of the broadcast tennis video. Fig. 2 shows the flow diagram. Our method starts by tracking and stabilizing player figure in each frame. In [3][5], the template based and normalized-correlation based tracker were utilized. By our observations, however, these algorithms are not robust enough for long time player tracking in the broadcast video. A sophisticated tracking strategy called SVR particle filter [6], which is an improved particle filter, is employed in our method. Optical flow field is derived as low-level feature with the post-processing of half-wave rectification and Gaussian smoothing. The optical flow field is then divided into slices based on the relationship between locomotory body parts and figure regions. Slice based optical flow histograms, which is a new motion descriptor henceforth abbreviated as S-OFHs, is proposed to provide a compact representation for spatial pattern of noisy optical flow. The concatenation of S-OFHs is fed into support vector machine learning framework to robustly capture the discriminative patterns in S-OFHs space. Two basic actions, left-swing and right-swing, are recognized in our experiments.

The rest of the paper is organized as follows. Section 2 introduces the player tracking and stabilizing module. In section 3, the local motion descriptor, which is slice based optical flow histograms, is proposed. Section 4 describes the

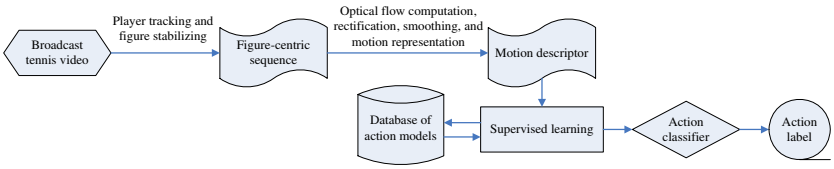

Fig. 2. Flow diagram of action recognition approach

classification mechanism based on support vector machine. Experimental results are presented and analyzed in section 5. Finally, we conclude the paper with future work in section 6.

2 Player Tracking and Stabilization

Our recognition algorithm starts by player tracking and human-centric figure computation. This can be achieved by tracking the player candidate region and then constructing a window in each frame centered at the player region.

The appropriate trackers utilized in our method are required to be consistent, that is, the tracker should be robust enough for the noisy circumstance so as to always map the person in a particular body configuration to approximately the same stabilized image. Existing methods for tracking tennis players are based on template matching [3][7][8] which is similar to the correlation based algorithm in [5]. These trackers are sensitive to the noise such as player deformation and background clutter caused by non-rigid object and low frame resolution, and cannot track player for a long video sequence. This can be exemplified in [7] that the input tennis video was first segmented into chunks of 30 frames and then performed tracking for each chunk separately. A sophisticated tracking strategy called SVR particle filter is employed in our approach. SVR particle filter enhances the performance of classical particle filter with small sample set and is robust enough for the noise in broadcast video. The experimental result is very satisfying. More details about this tracker can be found in [6].

To derive the human-centric figure, the tracking window around the player region is then enlarged by a certain scale in pixel unit and a simple method of computing centroid of the player region is used. The centroid coordinates of the region are defined as follows:

$$m_x = \frac{\sum_{x\in R}\sum_{y\in R} x f(x,y)}{\sum_{x\in R}\sum_{y\in R} f(x,y)} \,. \tag{1}$$

$$m_y = \frac{\sum_{x\in R}\sum_{y\in R} y f(x,y)}{\sum_{x\in R}\sum_{y\in R} f(x,y)} \,. \tag{2}$$

where R is the region occupied by the object on the image plane and $f(x,y)$ the gray level at location (x,y). Then the center of window controlled by tracker is shifted to position (m_x, m_y).

Once the video sequence is stabilized, the motion in broadcast video caused by camera behavior can be treated as being removed. This corresponds to a skillful movement by a camera operator who keeps the moving figure in the center of the view. Any residual motion within the human-centric figure is due to the relative motion of different body parts as limbs, head, torso and racket played with player.

3 Motion Descriptor Computation

As mentioned above, in previous approaches, appearance is not necessarily preserved across different action sequences. Different players may exhibit different postures for the same action and different postures may be recorded in different video even for the same action. Thus the appearance descriptor is not robust and discriminative for action recognition and classification.

We derive our features on pixel-wise optical flow as it is the most intuitive technique for capturing motion independent of appearance. The key challenge is that the computation of optical flow is not very accurate, particularly on coarse and noisy data such as broadcast video footage. The essence of our approach is to treat optical flow field as spatial pattern of noisy measurements which are aggregated using our motion descriptor instead of precise pixel displacements at points. Within the human-centric figure, the motion is due to the relative movements caused by player's different body parts which are the different regions being mapped into the image plane. These motion characteristics cannot be captured well by global features computed from the whole figure. A simple means of localizing the motion for recognition is to separately pay attention to different regions around the human torso. One way of doing this is to divide the optical flow field into various sub-regions called slices here. The histogram is utilized to represent the spatial distribution for each sub optical flow field in slices.

3.1 Optical Flow Computation and Noise Elimination

Before computing the motion descriptor, the optical flow feature of human-centric figure is derived and the noise in the flow field is eliminated by the algorithm shown in Fig. 3.

Noise in the background of human-centric figure makes significant influence for the computation of optical flow inside the human region. It necessitates background subtraction before computing optical flow. Considering the background of human-centric figure is playfield, an adaptive method of playfield detection [9] is applied in our experiments. After background pixel detection, region growing technique in [10] is performed as the post-processing to connect background pixels into regions, eliminate noises, and smooth boundaries.

We compute optical flow at each figure using Horn-Schunck algorithm [11]. The half-wave rectification and Gaussian smoothing is performed to eliminate the noise in optical flow field. The optical flow magnitudes are first thresholded to reduce the effect of too small and too large motion probably due to noise

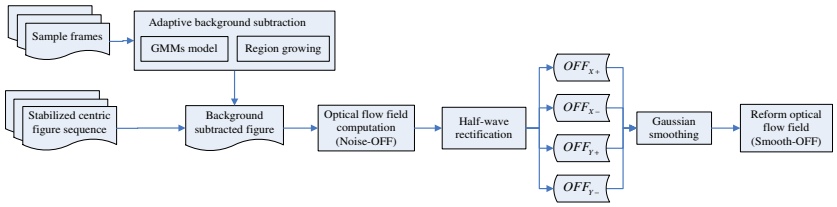

Fig. 3. Optical flow computation and noise elimination

inside the human region. The optical flow vector field OFF is then split into two scalar fields corresponding to the horizontal and vertical components OFF_X and OFF_Y, each of which is then half-wave rectified into four non-negative channels OFF_X^+, OFF_X^-, OFF_Y^+, and OFF_Y^-, so that $OFF_X = OFF_X^+ - OFF_X^-$ and $OFF_Y = OFF_Y^+ - OFF_Y^-$. They are each smoothed with a Gaussian filter to obtain the final channels. The smoothed optical flow field is reformed finally.

3.2 Local Motion Representation

Motion representation with global methods takes the whole image or video sequence into account. Local methods, on the other hand, focus on certain parts of the image or video data.

With the context of action recognition, the motion in the human-centric figure is due to the relative movement of different body parts which are exhibited in the different figure regions. This can be demonstrated by observing the optical flow field computed from the processed human-centric figure (see Fig. 4(a)). For left-swing, the optical flow field in left figure region is much denser than the field in right region. Contrarily, the field in the right region is denser than that in left region for right-swing. By this observation, we adopt a simple but effective region style which is called as slice. The whole optical flow field is split into three slices along the width orientation as show in Fig. 4(b). The height of the slice is equal to the one of figure and the width can be set adaptively in accordance with the object spatial structure. Here, we set even width for each slice.

Histogram based methods are widely used for spatial recognition. The advantage of histogram based representation is that it provides much information using very compact description if the dimensionality of the histogram is low. Motivated by the kernel density estimation for color distribution [12], a group of slice based optical flow histograms (S-OFHs) are derived. First we define $b(\mathbf{p}) \in \{1, \ldots, m\}$ as as the bin index of histogram associated with the optical flow vector $\mathbf{f}$ at location $\mathbf{p}$. For each position $\mathbf{q}$ inside optical flow field OFF, considering a grid region $R(\mathbf{q})$ centered at $\mathbf{q}$, the probability of the bin $u = 1, \ldots, m$ in the histogram of OFF is then computed as

$$h_u = C \sum_{\mathbf{q} \in OFF} \sum_{\mathbf{p} \in R(\mathbf{q})} k(\|\mathbf{p} - \mathbf{q}\|) \delta[b(\mathbf{p}) - u] \,. \tag{3}$$

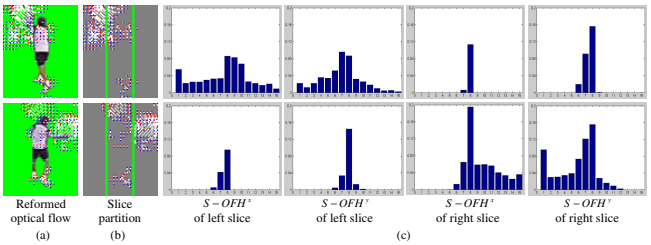

Fig. 4. Slice partition and slice based optical flow histograms (S-OFHs) of left and right swing

where δ is the Kronecker delta function, C is the normalization constant ensuring $\sum_{u=1}^{m} h_u = 1$, k is a convex and monotonic decreasing kernel profile which assigns a smaller weight to the locations that are farther from the center $\mathbf{q}$.

Given an optical flow field OFF_i for the figure F_i in human-centric figure sequence, $i = 1, \ldots, N$, where N is the total figure number, $OFF_{i,j}$ is the sub optical flow field in the jth slice, $j = 1, \ldots, L$ and here $L = 3$, the $S - OFH_{i,j}$ is then defined as follows accroding to Eq. 3:

$$h_u^{i,j} = C \sum_{\mathbf{q} \in OFF_{i,j}} \sum_{\mathbf{p} \in R(\mathbf{q})} k(\|\mathbf{p} - \mathbf{q}\|) \delta[b(\mathbf{p}) - u] \ . \tag{4}$$

Thus, for each $OFF_{i,j}$, two S-OFHs annotated as $S - OFH_{i,j}^x$ and $S - OFH_{i,j}^y$ are constructed for x and y orientation, respectively.

In our approach, left and right slices are selected for computing the S-OFHs excluding middle slice. Four S-OFHs for one figure are ultimately utilized as the motion descriptor. Fig. 4(c) shows the S-OFHs for corresponding action. We can see that S-OFHs can effectively capture the discriminative features for different actions in spatial space.

4 Action Classification

Various supervised learning algorithms can be employed to train an action pattern recognizer. Support vector machine (SVM) [13] is used in our approach. SVM has been successfully applied to a wide range of pattern recognition and classification problems. Compared with artificial neural networks (ANNs), SVM is faster, more interpretable, and deterministic. Moreover, SVM is a classification approach which has been gaining popularity due to its ability to correctly classify unseen data as opposed to methods as nearest neighbor classification (NNC). The advantages of SVM over other methods consist of: 1) providing better prediction on unseen test data, 2) providing a unique optimal solution for a training problem, and 3) containing fewer parameters compared with other methods.

The concatenation of four S-OFHs for each optical flow field in one figure is fed as feature vector into support vector machine. The radial basis function

(RBF) $K(x, y) = \exp(-\lambda\|x - y\|)$ is utilized to map training vectors into a high dimensional feature space for classification.

5 Experiments

This section shows our experimental results on recognizing two basic actions, which are left-swing and right-swing, performed by the player near the camera in far-view scene within broadcast tennis video. By our observations, these two actions occupy about 90% behavior occurred in tennis games.

The test data used in experiments are derived from the video recorded from live broadcast television for one of the matches of Pacific Life Open 2004 in Indian Wells between Agassi and Hrbaty. The video is compressed in MPEG-2 standard with the frame resolution of 352×288. 5 sequences are extracted from the whole video which the total number of frames is 6035 including 1099 for left-swing, 1071 for right-swing. Others are non-swing frames which consist of close-up scenes of players, views of spectators and so on. 56 left-swing actions and 49 right-swing actions are involved in the test sequences.

Two experiments are implemented: one is for action recognition on swing frames; the other is for recognition on swing action clips. To qualitatively gauge the performance, the values of Recall (R) and Precision (P) are calculated. The Accuracy (A) metric is employed to evaluate the holistic performance. Additional 750 swing frames are utilized to train the two-action SVM model.

5.1 Recognition on Frames

First we perform the experiment for classifying 2170 swing frames into two action categories by pre-defined action model. Table 1 summarizes the experimental results which are promising. The Accuracy of the holist is 87.10%. The reason of incorrect recognition is that the player is deformable object of which the limbs make free movement during the action displaying. This will disturb the regular optical flow distribution to make the S-OFHs misreport the motion characteristics in the figure.

Table 1. Experimental results of recognition on frames

	# frame	Recall(%)	Precision(%)	Accuracy(%)
Left-swing	1099	84.08	89.80	
Right-swing	1071	90.20	84.66	
Total	2170			87.10

5.2 Recognition on Action Clips

Based on frame recognition and voting strategy, 56 left-swing and 49 right-swing actions are classified into two categories so as to recognize each action type displayed in the video. This experiment is performed within the framework of

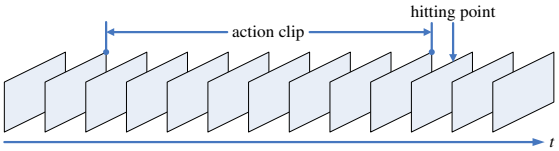

Fig. 5. Location of action clip in the video

multimodal analysis by integrating our action recognition approach with audio-assisted technique in sports video.

First, we employ the audio modalities based method in [14] to detect the hitting ball so as to locate the action clip in the video. As shown in Fig. 5, the frame corresponding to the occurrence of hitting ball is called hitting point. Then the group of frames in the adjacent window before hitting point is selected as action clip. The window length is empirically set to 25 frames in the experiment.

Given f_i which is the ith frame in video clip V, the corresponding human-centric figure obtained by our approach is hc_i. The vote that f_i contributes to V is defined as:

$$Vote(f_i) = \begin{cases} 1 & \text{if } Reg(hc_i) = left - swing \\ -1 & \text{if } Reg(hc_i) = right - swing \end{cases}. \tag{5}$$

where function $Reg(\bullet)$ refers to our action recognition approach. The final recognized action category is determined as

$$Category(V) = \begin{cases} left - swing & \text{if } \sum\limits_{f_i \in V} Vote(f_i) \geq 0 \\ right - swing & \text{if } \sum\limits_{f_i \in V} Vote(f_i) < 0 \end{cases}. \tag{6}$$

Because here are two categories, the equation is just assigned for left-swing so as to avoid the occurrence of marginal classification.

Table 2. Experimental results of recognition on action clips

	# clip	Recall(%)	Precision(%)	Accuracy(%)
Left-swing	56	87.50	90.74	
Right-swing	49	89.80	86.27	
Total	105			88.57

Table 2 shows the experimental results which the Accuracy for all the action clips is 88.57%. Compared with the results on frames in Table 1, it is more satisfactory because the classification is aggregated over the temporal domain by voting strategy. Fig. 6 illustrates some representative frames from the video for the actions recognized by our approach accurately.

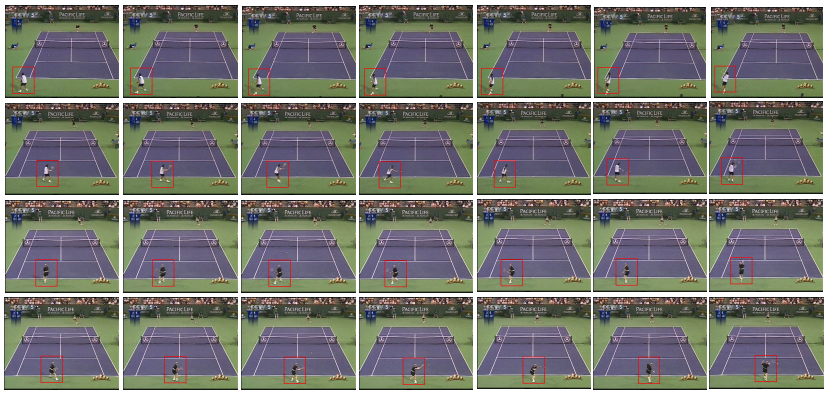

Fig. 6. Experimental results of action recognition for left-swing and right-swing

6 Conclusion and Future Work

An action recognition approach is presented in this paper for motion analysis of tennis player in broadcast video. A novel motion descriptor, which is a group of histograms abbreviated as S-OFHs, is proposed based on smoothing and aggregated optical flow measurements. The noisy optical flow is treated as a spatial pattern of noisy measurements instead of precise pixel displacements. S-OFHs are derived as spatial pattern representation. To recognize the action being performed by a human figure in one frame, support vector machine is employed to train the classifier where the concatenation of S-OFHs is formed as the input feature. The experiments demonstrate that our method is outperformed and the results are promising.

More effective slice partition and elaborate S-OFHs description will be considered in future work so as to include more semantic tennis actions in the recognition framework. Player trajectory is another useful information for semantic analysis and understanding of the game. The integration of action recognition and trajectory tracking will be paid more attention for the research of sport video analysis and enrichment such as event detection, tactic analysis, and 3-D scene reconstruction.

Acknowledgements

This work is supported by Beijing Natural Science Foundation: 4063041.

References

1. Rao, C., Shah, M.: View-invariance in action recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (2001) 316–322
2. Yacoob, Y., Black, M.J.: Parameterized modeling and recognition of activities. In: IEEE International Conference on Computer Vision. (1998) 120–127

3. Miyamori, H., Iisaku, S.: Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In: IEEE International Conference on Automatic Face and Gesture Recognition. (2000) 320–325
4. Cutler, R., Davis, L.S.: Robust real-time periodic motion detection, analysis, and applications. IEEE Transaction on Pattern Analysis and Machine Intelligence **22**(8) (2000) 781–796
5. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision. (2003) 726–733
6. Zhu, G., Liang, D., Liu, Y., Huang, Q., Gao, W.: Improving particle filter with support vector regression for efficient visual tracking. In: IEEE International Conference on Image Processing. Volume 2. (2005) 422–425
7. Sudhir, G., Lee, J.C.M., Jain, A.K.: Automatic classification of tennis video for high-level content-based retrieval. In: IEEE International Workshop on Content-Based Access of Image and Video Databases. (1998) 81–90
8. Pingali, G.S., Jean, Y., Carlbom, I.: Real time tracking for enhanced tennis broadcasts. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (1998) 260–265
9. Jiang, S., Ye, Q., Gao, W., Huang, T.: A new method to segment playfield and its applications in match analysis in sports video. In: ACM Multimedia. (2004) 292–295
10. Ye, Q., Gao, W., Zeng, W.: Color image segmentation using density-based clustering. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume 3. (2003) 345–348
11. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17**(1-3) (1981) 185–203
12. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transaction on Pattern Analysis and Machine Intelligence **25**(5) (2003) 564–577
13. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)
14. Xu, M., Duan, L.Y., Xu, C.S., Tian, Q.: A fusion scheme of visual and auditory modalities for event detection in sports video. In: IEEE International Conference on Acoustics, Seech, and Signal Processing. Volume 3. (2003) 189–192