Group Action Recognition in Soccer Videos

Yu Kong[1,2], Xiaoqin Zhang[2], Qingdi Wei[2], Weiming Hu[2], Yunde Jia[1]
[1]*Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P.R. China*
[2]*National Laboratory of Pattern Recognition, Institute of Automation, Beijing, P.R. China*
{*kongyu, jiayunde*}*@bit.edu.cn,* {*xqzhang,qdwei,wmhu*}*@nlpr.ia.ac.cn*

Abstract

Group action recognition in soccer videos is a challenging problem due to the difficulties of group action representation and camera motion estimation. This paper presents a novel approach for recognizing group action with a moving camera. In our approach, ego-motion is estimated by the Kanade-Lucas-Tomasi feature sets on successive frames. The optical flow is then computed on compensated frames. Due to the inaccurate ego-motion estimation, the optical flow can not reflect accurate motion of objects. In this paper, we propose a new motion descriptor which treats the optical flow as spatial patterns and extracts accurate global motion from the noisy optical flow. The Latent-Dynamic Conditional Random Field model is employed to recognize group action. Experimental results show that our approach is promising.

1. Introduction

Understanding group action is essential for many applications such as sport video analysis and video surveillance. Group action analysis in soccer videos is able to make audiences know what happens on the playground, such as which team is attacking and who goals. As to visual surveillance, since current surveillance systems focus on abnormal event detection, group action recognition is an important complement to it.

Most soccer videos exhibit various views from multiple angles and positions. It is the combination of shots that help viewers reconstruct activities as if they are on spot [3]. In this paper, only court views are concerned. During soccer games, two teams attack (see Fig.1), defend or be brought into a stalemate which is a transient state between attacking and defending. Our goal is to recognize their action from videos captured by the moving camera.

(a) frame at time t (b) frame at time $t + 30$

Figure 1. An example for right side attacking.

There are many studies related to ours. Crowd flow stability analysis aiming to detect abnormal incidents is now receiving more considerations. Ali and Shah [1] adopted Lagrangian Particle Dynamics approach to segment high density crowd flows and detect flow instability. The Hidden Markov Models is employed to detect emergency or abnormal event in crowds [2]. However, these approaches merely focus on the stability of flows; the action of crowd is not considered. What's more, flows used in these pieces of work are somewhat accurate since the camera is fixed. In recent years, player activity analysis in sport videos has been deeply investigated. However, individual action is mainly concerned in a vast amount of research; little work deals with group action. Efros et al. [4] introduced a motion descriptor based on the optical flow to recognize individual action in medium view on soccer court. A slice based optical flow histograms (S-OFHs) approach proposed in [11] is to recognize left-swing or right-swing of a player in tennis videos. Tracking is used in their methods.

In this paper, we propose a novel approach to recognize group action with a moving camera. We start with selecting corresponding Kanade-Lucas-Tomasi (KLT) feature sets in two successive frames. These features are used to estimate ego-motion of camera. The optical flow is then computed on frames after ego-motion compensation. Due to the inaccurate ego-motion compensation, the optical flow is noisy and can not reflect accurate motion of moving objects. To overcome this

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

problem, we propose a new motion descriptor which treats the optical flow as spatial patterns and extracts accurate global motion from the noisy optical flow. In the recognition stage, the Latent-Dynamic Conditional Random Field (LDCRF) model [7] is employed to recognize group action.

2. Ego-motion Compensation

The camera we concern is the one which gives field views. This camera always moves to track ball during broadcasting. Therefore, ego-motion compensation is needed to compensate camera motion. Here, camera motion is constrained to pan and tilt.

2.1. KLT Feature

The KLT (Kanade-Lucas-Tomasi) feature tracking algorithm [8, 9] has been a standard technique for feature-based computer vision tasks. It is employed here to select a set of good features from a frame [5].

Assuming F^t is a frame at time t. For a selected window on F^{t-1}, the feature selection algorithm runs on it and selects a feature set f^{t-1}. Feature set f^{t-1} is tracked on F^t and a set of tracked feature f^t is generated. From coordinates of f^{t-1} and f^t, a transformation between two consecutive frames can be derived.

2.2. Ego-motion estimation

Since only pan and tilt are considered, a linear affine model is used to model camera motion.

Affine transform T^t_{t-1} is defined as

$$\begin{pmatrix} x_t \\ y_t \\ 1 \end{pmatrix} = T^t_{t-1} \begin{pmatrix} x_{t-1} \\ y_{t-1} \\ 1 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \\ 1 \end{pmatrix}.$$

T^t_{t-1} is derived by minimizing SSD (sum of squared differences). However, some features may be associated with moving objects. Thus, these features should be eliminated since they lead to inaccurate estimation.

The motion of feature i on the current frame is

$$m_i = \mid f^t_i - T^t_{t-1} f^{t-1}_i \mid .$$

If $m_i > \delta$, feature i should be eliminated and T^t_{t-1} should be re-computed using the subset of f^t_{i-1} and f^t_i from which all outliers have been eliminated.

3. Computing Motion Descriptor

A group action is a behavior acted by several individuals. These individuals must act for the same goal. In this sense, they act as a whole and their group action is meaningful. We model group motion as global motion which are achieved by computing optical flow using the Lucas-Kanade algorithm [6].

(a) original image (b) optical flow of the region

Figure 2. Static objects have nonzero velocities on the optical flow field computed on compensated frames.

The key issue is how to effectively extract accurate global motion from the noisy optical flow. Although ego-motion is estimated, compensation is not precise because of noises. This leads to the inaccuracy of optical flow field. On this kind of optical flow field, static objects may have non-zero velocities (see Fig.2) and velocities of moving objects are incorrect. Compared with the optical flow computed on frames captured by a fixed camera or on a spatial-temporal volume, optical flow computed on compensated frames is more inaccurate.

To overcome the above problem, we treat the noisy optical flow as spatial patterns, and propose a robust motion descriptor for group action recognition.

3.1. Global motion representation

Given a optical flow field OFF, we treat it as vectors or complex numbers, i.e., $OFF = \boldsymbol{x} + \boldsymbol{y}i$. Its corresponding polar coordinate representation is $OFF\_PC = (\boldsymbol{\rho}, \boldsymbol{\theta})$, where $\boldsymbol{\rho}$ is the magnitude and $\boldsymbol{\theta}$ is the phase angle. Motivated by [4], the magnitude and the phase angle are treated as two separated channels of the optical flow.

The centroid of magnitude channel is defined as the average of all normalized ρ in OFF_PC:

$$C_\rho = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \exp(-\frac{\rho_i^2}{2\sigma^2})\right) \quad (1)$$

where N is the number of pixels, and σ is the standard deviation of Gaussian kernel used in normalization of ρ.

The centroid of phase angle channel is computed as

$$C_\theta = ANGLE\left(\sum_{j=1}^{N} \frac{x_j + y_j i}{\rho_j}\right), \quad (2)$$

where $x_j + y_j i$ is the j^{th} element on OFF, and the function $ANGLE$ is to compute the phase angle of a vector. This definition is more reasonable than the average of all angles. For example (see Fig.3), the angle of *vector 1* is 15° and that of *vector 2* is 345°. Normalized vectors are shown as short vectors. Our method is to obtain the angle of *vector 3* (0°) rather than that of *vector 4* (180°)

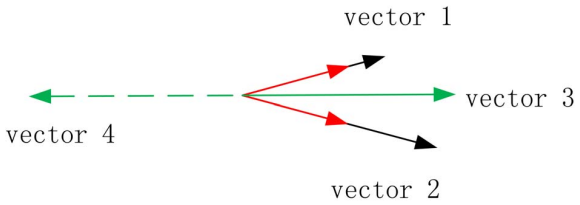

Figure 3. An example for computing the centroid of phase angle.

which is the average of 15° and 345°. From a geometrical point of view, 0° is much closer to 15° and 345° than 180°.

A complex number $x + yi$ on OFF can be seen as a two-dimensional point (x, y). Thus, all points on OFF form a motion area which is the spatial patterns of optical flow field. This motion area is similar to a figure's silhouette which is an important cue in determining underlying motion of a walking figure [10]. Motivated by this, the motion area containing spatial patterns of the optical flow is used to extract accurate global motion.

Assuming $\xi_{i,j}$ is the j^{th} element of channel $i, i = \rho, \theta$. $\xi_{\rho,j}$ should be normalized as ρ in (1). The relation between $\xi_{i,j}$ and the centroid of channel i is given by

$$R_{\xi_{i,j}} = \exp\left(-\frac{(\xi_{i,j} - C_i)^2}{2\sigma_i^2}\right) \tag{3}$$

where C_i is the centroid of channel $i, i = \rho, \theta$, and σ_i is a parameter of channel i. In this way, the motion area is represented as relations of two separated channels.

Let $S_{k,i}$ denotes the k^{th} interval for relations of channel i. Then the percentage of relations of channel i in the k^{th} interval is

$$M_{k,i} = \frac{\sum_{j, R_{\xi_{i,j}} \in S_{k,i}}^{N} R_{\xi_{i,j}}}{\sum_{j}^{N} R_{\xi_{i,j}}}. \tag{4}$$

M is a histogram which describes the ratio of the summation of relations in $S_{k,i}$ to the total relations of channel i. Our motion descriptor is composed of M.

Assuming OFF is the optical flow field computed on imprecisely compensated frames and OFF^* is the one computed on precisely compensated frames. Due to the imprecise compensation, C_ρ and C_θ may be not equal to them of OFF^*. Nevertheless, since ego-motion is modeled as a linear affine model, the magnitudes and angles of ξ associated with static objects have almost the same errors with C_ρ and C_θ respectively. Consequently, if ξ is associated with a static object, its R of OFF is almost equal to it of OFF^*. As to ξ associated with a moving object, its R is in proportion to it of OFF^*. Based on the assumption that the moving objects are in the minority in global motion, M is a approximation to that of OFF^*. Therefore, our motion descriptor composed of these statistics performs reliably with the noisy data and is a discriminative feature for group action.

4. Action Recognition

The Latent-Dynamic Conditional Random Field (LDCRF) model [7] is used to classify group action. The LDCRF model aims to learn a mapping between observations $\boldsymbol{x}$ and labels $\boldsymbol{y}$ via the hidden variables $\boldsymbol{h}$ which is used to model the sub-structure of a class sequence and learn dynamics between class labels.

Latent conditional model is defined as

$$P(\boldsymbol{y}|\boldsymbol{x}, \theta) = \sum_{\boldsymbol{h}} P(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{x}, \theta) P(\boldsymbol{h}|\boldsymbol{x}, \theta),$$

where θ are the parameters of the model.

If h_j is not in the set of Hidden States H_{y_j}, $P(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{x}, \theta)$ is set to 0. Thus, the model can be written as

$$P(\boldsymbol{y}|\boldsymbol{x}, \theta) = \sum_{\boldsymbol{h}: \forall h_j \in H_{y_j}} P(\boldsymbol{h}|\boldsymbol{x}, \theta).$$

Conditional Random Fields formulation is used to define $P(\boldsymbol{h}|\boldsymbol{x}, \theta)$:

$$P(\boldsymbol{h}|\boldsymbol{x}, \theta) = \frac{1}{Z(\boldsymbol{x}, \theta)} \exp\left(\sum_{k} \theta_k \cdot F_k(\boldsymbol{h}, \boldsymbol{x})\right),$$

where the partition function Z is defined as $Z(\boldsymbol{x}, \theta) = \sum_h \exp\left[\sum_k \theta_k \cdot F_k(\boldsymbol{h}, \boldsymbol{x})\right]$. F_k is defined as $F_k(\boldsymbol{h}, \boldsymbol{x}) = \sum_{j=1}^{m} f_k(h_{j-1}, h_j, \boldsymbol{x}, j)$.

At learning stage, parameters θ is learned using the follow objective function:

$$L(\theta) = \sum_{i=1}^{N} \log P(\boldsymbol{y}_i|\boldsymbol{x}_i, \theta) - \frac{1}{2\sigma^2}\|\theta\|^2.$$

Here, we assumed that θ is subject to a Gaussian prior with variance σ^2, i.e. $P(\theta) \sim \exp(\frac{1}{2\sigma^2}\|\theta\|^2)$.

5. Experimental Results

The dataset we used for experiments is composed of 126 (about 19000 frames) real soccer video clips containing three action categories. Each video clip only belongs to one action class. 30 clips are used for training and the rest are for testing. Each frame is filtered using the Median Filter to regress noises before computing the optical flow.

The three classes we concern are "left side attacking" (LA), "stalemate" (ST) and "left side defending" (LD) (see Fig.3). "Attacking" means that the attacking group keeps the ball; its rough moving direction is toward their rival. "Defending" is the action that the group sets formation for defending and is not controlling the ball. As to "stalemate", it is a transient state between attacking and defending. In the "stalemate" action, each group strives to intercept the ball or one group passes ball patiently to keep pace of the game.

(a) left side attacking (b) stalemate (c) left side defending

Figure 4. Three actions for recognition.

5.1. Results

Comparison experiments are conducted using the optical flow histograms and the S-OFHs [11]. Ego-motion compensation is served as a preprocessing step for the three kinds of descriptors. Experimental results are shown in Table 1.

In the experiment of our motion descriptor, recognition rate of "LD" and "ST" are somewhat lower than that of "LA". It is because that, in "LD" clips, the camera moves quite fast to track the ball. Consequently, camera motion estimation is excessively inaccurate and leads to a coarse approximation of the accurate optical flow. As for "ST" clips, the misclassification is due to the big error of the approximation. This occurs when the assumption that moving objects are in the minority of global motion is broken.

There are 40 clips out of 96 misclassified in histograms experiment. Since histograms merely focus on counting, they are sensitive to noises and unable to obtain correct values from the noisy optical flow.

In the results of S-OFHs, recognition rate of "left side attacking" is 90.63% while the other two are all below 40%. In our test, about 41% clips of "LD" and "ST" are misclassified as "LA". This draws to a conclusion that S-OFHs fails to extract discriminative features from the noisy data, so a majority of clips are classified as "LA". Therefore, S-OFHs does not have a good discriminant quality in group action recognition.

Since spatial relations of the accurate optical flow are reserved and approximations of the accurate optical flow statistics can be obtained in further step, our motion descriptor is effective in extracting accurate global motion from the noisy optical flow. In addition, the statistics have different characteristics in different group actions. Therefore, these statistics are discriminative features for global motion classification and our motion descriptor is a fine model for group action recognition.

6. Conclusion

This work is devoted to group action analysis. A novel approach is proposed to recognize group action with a moving camera. The corresponding KLT feature sets on successive frames are used to tackle the problem of ego-motion compensation. In our approach, group action is modeled as global motion achieved by computing the optical flow. In order to handle the noisy data caused by inaccurate ego-motion compensation, a new motion descriptor which treats the optical flow field as spatial patterns is proposed to represent global motion. The LDCRF model is employed for recognition. Experimental results show that our approach is promising and our motion descriptor is robust.

Table 1. Experimental Result

	Recognition Rate (%)		
	LA	LD	ST
Our Motion Descriptor	90.63	83.87	84.85
Optical Flow Histograms	62.50	67.74	45.45
S-OFHs	90.63	29.03	36.36

Acknowledgement

This work is partly supported by NSFC (Grant No. 60672040, 60705003) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

References

[1] S. Ali and M. Shah. A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis. In *CVPR*, 2007.

[2] E. L. Andrade, S. Blunsden, and R. B. Fisher. Hidden Markov Models for Optical Flow Analysis in Crowds. In *ICPR*, vol 1, pages 460-463, 2006.

[3] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin. A Unified Framework for Semantic Shot Classification in Sports Video. *IEEE Transactions on Multimedia*, 7(6):1066–1083, 2005.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Mal. Recognizing Action at a Distance. In *ICCV*, vol.2, pages 726-733, 2003.

[5] B. Jung and G. S. Sukhatme. Real-time Motion Tracking from a Mobile Robot. Technical report, University of Southern California, 2005.

[6] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Imaging Understanding Workshop*, pages 121-130, 1981.

[7] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-Dynamic Discriminative Models for Continuous Gesture Recognition. In *CVPR*, 2007.

[8] J. Shi and C. Tomasi. Good Features to Track. In *CVPR*, pages 593-600, 1994.

[9] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, CMU, April 1991.

[10] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Trans. PAMI*, 25(12):1505–1518, December 2003.

[11] G. Zhu, C. Xu, Q. Huang, and W. Gao. Action Recognition in Broadcast Tennis Video. In *ICPR*, vol 1, pages 251-254, 2006.