Various sports video genre categorization methods are proposed recently, mainly focusing on professional sports videos captured for TV broadcasting. This paper aims to categorize sports videos in the wild, captured using mobile phones by people watching a game or practicing a sport. Thus, no assumption is made about video production practices or existence of field lining and equipment. Motivated by distinctiveness of motions in sports activities, we propose a novel motion trajectory descriptor to effectively and efficiently represent a video. Furthermore, temporal analysis of local descriptors is proposed to integrate the categorization decision over time. Experiments on a newly collected dataset of amateur sports videos in the wild demonstrate that our trajectory descriptor is superior for sports videos categorization and temporal analysis improves the categorization accuracy further.

Index Terms—Genre categorization, Activity recognition, Trajectory, Amateur sports video, Temporal analysis.

1. INTRODUCTION

Content-based video categorization has a crucial role in making the ever-increasing amount of digital content accessible. Automatic indexing and genre categorization of sports videos, as a large portion of digital contents, is of significance and enables domain-specific analysis of sports videos [1].

Sports video categorization is a challenging problem. There might be large inter-class similarities due to similarity of movements, coexistence of players and audiences, and commonality between playing fields. In addition, intra-class variations, such as different movements within a single sport video, camera angle variations, and distinct speeds of actions by different people, render the categorization task difficult. While most previous works assume that sports videos are captured for TV broadcasting, and thus, happen in specific sports fields [1–3], this work aims to analyze sports videos in the wild, i.e., amateur videos captured by mobile phones (Fig. 1). These videos have additional challenges due to the field variations, camera motion, and the unskillful capturing.

Many sports activities have a very distinctive set of motions that can be useful to characterize the sport. Our approach is built upon dense trajectories [4] extracted from the optical flow based tracking. However, the simple trajectory descriptor in [4] does not explicitly encode the shape or temporal dynamics of trajectories. It is important to have an efficient and effective trajectory shape descriptor that is robust to camera angle variations and different speeds of actions. Based on these requirements, we propose a novel Orientation based Camera angle Invariant Trajectory descriptor called OCIT, which is both compact and discriminative.

Furthermore, if the temporal ordering or dynamic of the trajectories is not represented, some sports may be confused with each other. Hence, it is beneficial to analyze the trajectories in a temporal framework. In this paper, a temporal analysis (TA) method is proposed to capture local descriptors in overlapping blocks over time and fuse the analysis results from all blocks for making the final categorization decision.

We collect a large dataset of amateur videos captured by users of a leading sports video mobile app. Experiment results on this dataset show that OCIT outperforms displacement-based trajectory descriptor used in [4, 5]. Also, TA of different types of descriptors improves the performances of the Bag of Words (BoW) [6] method using the same descriptors.

In summary, this paper makes three contributions: (i) To the best of our knowledge, this is the first work on categorizing sports videos in the wild; (ii) A novel trajectory descriptor is proposed to capture trajectory information discriminatively and efficiently; (iii) A temporal analysis approach is presented to integrate the categorization of local descriptors over time.

Previous Work Most prior works assume that sports occur in sports arena (thus the existence of specific equipment and field lining), and videos are captured by professional TV broadcast crew. Duan et al. classify TV broadcasting sports videos via motion, color, and shot length features [2]. In [1], the dominant motion and color in each frame is used to classify 3 sports genres. In [7], a hierarchical SVM is used to categorize sports genres by employing temporal and spatial features. Assuming distinct playing field for different sports, histograms of edge direction and intensity are used to categorize 5 sports genres in [8]. In [9] SIFT features of the sample
frames and BoW [6] are used to categorize sports videos.

Sports videos in the wild are more challenging for visual analysis than broadcasting videos. Firstly, the static image context is less discriminative. Secondly, the camera angle variation is enormous and videos are affected by camera motion. Thirdly, multiple activities may appear in a single video. Finally, cluttered backgrounds increase analysis difficulty. Given these challenges, for sports videos in the wild, we prefer motion-based analysis to context-based analysis.

As a relevant topic, activity/action recognition differs in nature to sports video categorization. While the former aims to recognize specific actions separately, in the latter a wide range of activities fall into a single genre, leading to a greater intra-class variance. Also, unlike our dataset, action recognition datasets are usually comprised of short videos that precisely encapsulate the action of interest. Activity recognition works can be categorized in recognition from still images [10–12] and videos [13]. They can also be divided into context [9, 12] or motion based methods [4, 14–16]. In the latter, either space-time features [14,15] or trajectories of motions [10–12] and videos [13]. They can also be divided to context [9, 12] or motion based methods [4, 14–16]. In the latter, either space-time features [14,15] or trajectories of motions are extracted [4, 16–18]. For both, the dense sampling outperforms interest-based sampling [4, 19]. Our work is a new development along the trajectory-based method that by introducing a novel trajectory descriptor and temporal analysis, improves genre categorization of sports videos in the wild.

2. OUR PROPOSED APPROACH

In our method, motion is analyzed by extracting dense trajectories [4]. To robustly analyze videos in the presence of camera motion, frame by frame motion stabilization is first achieved by matching interest points on consecutive frames and applying the RANSAC algorithm [20] to obtain the affine transformation between consecutive frames.

2.1. Dense trajectory and descriptors

As proposed in [4], dense trajectories are extracted at multiple spatial scales. Each point \( p_t = (x_t, y_t) \) at frame \( t \) is tracked to the next frame \( t + 1 \) by performing median filtering in a dense optical flow field \( W = (u_t, v_t) \),

\[
p_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (K * \nabla W)_{(x_t,y_t)},
\]

where \( K \) is the median filtering kernel and \( (\nabla x, \nabla y) \) is the rounded position of \((x_t, y_t)\). Trajectories are started from the sample points on a grid spaced by \( \Delta \) between consecutive frames. \( \Delta \) is the median filtering kernel and \( \Delta \) is used to describe the trajectory shape. We denote the histogram of \( \alpha_t \) and \( \Delta \alpha_t \) as \( h = [h_1, ..., h_N] \) and \( g = [g_1, ..., g_N] \). Since tiny motions result in very small trajectory segments, which are less important in overall shape, the contribution of each trajectory segment is weighted by \( || \Delta p_t || \), for both \( h \) and \( g \). Thus, \( h \) and \( g \) are defined as,

\[
h_n = \sum_{t=1}^{L} \delta_n(\alpha_t)||\Delta p_t||; n \in \{1, ..., N\},
\]

\[
g_m = \sum_{t=1}^{L} \varphi_m(\Delta \alpha_t)||\Delta p_t||; m \in \{1, ..., N\},
\]

where \( \delta_n \) and \( \varphi_m \) are the indicator functions,

\[
\delta_n(\alpha_t) = \begin{cases} 1, & \text{if } \frac{n-1}{N} \frac{\Pi}{2} < \alpha_t < \frac{n}{N} \frac{\Pi}{2} \\ 0, & \text{otherwise} \end{cases}
\]

\[
\varphi_m(\Delta \alpha_t) = \begin{cases} 1, & \text{if } \frac{2(m-1)}{N \Delta} \Pi < \Delta \alpha_t < \frac{2m}{N \Delta} \Pi \\ 0, & \text{otherwise} \end{cases}
\]

By concatenating \( h \) and \( g \), followed by \( L_2 \) normalization, the new trajectory descriptor named Orientation-based Camera angle Invariant Trajectory (OCIT) descriptor, is defined,

\[
\text{OCIT} = \frac{(h, g)}{\sqrt{|| h ||^2 + || g ||^2}}.
\]

Given the descriptors computed from a set of training videos, we perform codebook learning via \( K \)-means clustering and observe the variability of the code words. As shown in Fig. 2 (a), many of the trajectories for s look similar, but in fact they represent different paces of movements. Although the training videos contain many trajectories with considerable curvatures, they are not well captured by the code words...
2.3. Temporal analysis of videos

Since temporal segmentation is normally not available for sports videos in the wild, it is possible to encounter cases that only a small part of the video contains representative motions. Thus, it is critical to analyze the video in short time segments and properly fuse them to make the final decision based on the most informative segments. Many works split the videos temporally to capture semantics of actions. Some works find the most informative segments. Many works split the videos temporally to capture semantics of actions. Some works find the most informative segments.

For TA of motions, each video volume is divided to non-overlapping temporal cells of 1-second length. Histograms of different descriptors based on the trajectories are calculated for each cell. As shown in Fig. 3, the histograms of \( N_c \) consecutive cells are then concatenated and \( L_2 \) normalized to form the feature representation of one block \( b^k \), where \( k \) is the index of the block. Thus, the feature dimension of each block, \( d \), is \( N_c \) times the feature dimension of each cell. Blocks slide over cells, with \( \frac{100(N_c-1)}{N_c} \) percentage of overlapping between consecutive blocks. Now, in the collection of block features corresponding to a single video, at least one block represents the most informative \( N_c \)-second segment of the video. Since TA increases the dimension of video representation by a factor of \( N_c \), we reduce the dimension of \( b^k \) via PCA such that 95% of the variance is retained. If the number of cells in a video is less than \( N_c \), the cells are concatenated and zero padded to form the block.

For a \( C \)-category categorization problem, a classifier \( f : \)

\[ R^d \rightarrow R^C \]

is trained over \( d \)-dim block feature \( b^k \) and outputs a \( L_1 \) normalized \( C \)-dim score vector representing the probability of the block belonging to each of the \( C \) categories. Given a test video \( i \), \( M \) cells and \( M - N_c + 1 \) blocks are generated. Experiments show that for blocks where the trajectories are not representative of a specific sports genre, the scores are more randomly distributed over a larger number of categories, and hence the maximum score is relatively low. By denoting the feature representation of block \( k \) of video \( i \) as \( b^i_k \), and its scores as \( x^i_k (k = 1, \ldots, M - N_c + 1) \), a weighted fusion is used to compute the final score vector of video \( i \), denoted as \( x^i_k \) (both \( x^i_k \) and \( x^i_k \) are \( C \)-dim vectors). The weight of each block is the likelihood of the maximum score of the block given a correct categorization, denoted as \( p^i_c (\cdot) \) and estimated by a Gaussian distribution during training. Thus,

\[
f(b^i_k) = x^i_k = (x^i_{k,1}, \ldots, x^i_{k,C}) \text{ s.t. } \sum_{c=1}^{C} x^i_{k,c} = 1, \quad (6)
\]

\[
x^i_k = \sum_{k=1}^{M-N_c+1} p^i_c (\max_{c} x^i_{k,c}) x^i_{k,c} \cdot (7)
\]

The final sports category of video \( i \), \( y^i_i \), is the category with the maximum value in \( x^i_c \),

\[
y^i_i = \arg \max_{c} (x^i_{i,1}, \ldots, x^i_{i,C}). \quad (8)
\]

3. EXPERIMENTAL RESULTS

Dataset We collected a dataset of 1,047 videos from 15 sports categories captured by amateur users via a mobile phone app. In each category, 50 videos are used for training and 15 – 25 videos for testing. The average, max., and min. video length is 35s, 242s, and 1s respectively. Videos are not temporally segmented, so the most informative segment may appear at any part of the video. Our dataset is favorably comparable in size with UCF Sports [25] and Olympic Sports [26] datasets (both are professional sports videos captured by professional TV crew), where each of the three datasets has 15, 9 and 16 categories, and \( \sim 70 \), \( \sim 20 \) and 50 videos per category respectively. We will make our dataset publicly available.

Implementation Details We use \( K \)-means clustering to learn a BoW codebook [6] and the implementation in [5] to calculate dense trajectories. To prevent the trajectories of the background or audiences from dominating the trajectories of the players, in all BoW representations the bins with values

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{TA of video blocks, each composed of \( N_c \) cells.}
\end{figure}

\[\text{http://www.cse.msu.edu/~liuxm/sportsVideo}\]
larger than $\mu + u\sigma$ are clipped to $\mu + u\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation of the values in all bins and $u$ is a clipping parameter. This is similar to the clipping normalized histograms in SIFT by a threshold of 0.2 to ensure robustness to illuminations [27]. To compute the trajectories and descriptors, we set $W = 15$, $L = 30$, $N = 10$, $N_\Delta = 5$, $N_c = 5$, and $u = 3$. Default parameters as in [4] are used for trajectory aligned HOG and MBH descriptors. RBF kernel SVM is used for classification and the parameters are tuned via 5-fold cross validation. In all experiments, the number of code words in BoW is 100, 50, 100 and 100 for $s$, OCIT, HOG and MBH respectively. The categorization accuracy, the fraction of correctly categorized videos, is used as the metric.

**Results of Accuracy** As shown in Tab. 1, OCIT outperforms $s$ by itself (45% vs. 38%) and by combining with other descriptors, in both BoW and TA. Considering the compactness of OCIT, this is an impressive result. For all combinations of descriptors, TA outperforms BoW. However, as features get richer through combination of descriptors, TA deals with higher dimensionality and results in less improvement due to the curse of dimensionality. Performance of TA for HOG+OCIT is better than BoW for richer combination of MBH+HOG. Note that OCIT is substantially more efficient to compute than MBH. The top performance of 69.1% is achieved by fusing MBH, HOG and OCIT in the TA approach. To compare with prior work, we implement a state-of-the-art content-based sports categorization method [9], which uses BoW on SIFT descriptors, and receive an accuracy of 42.9%. This demonstrates that for categorization of sports videos in the wild, motion-based method is more powerful. Figure 4 shows the environment diversity in our data and provide probable reason for the poor result of [9].

Figure 5 illustrates a temporal analysis result of a 31-second video of Hockey category with the label 11. TA extracts 27 blocks for this video. The label assigned to each block is shown in this figure, with the size of circle representing the weight assigned to each block as in Eqn. 7. The first 10 seconds of the video are more representative of Hockey and are correctly labeled as 11 with larger weights. Therefore, in spite of all the ambiguities in the later part of the video, this video is correctly labeled as Hockey by the temporal analysis.

Figure 6 shows the accuracy of categorization for each category using the TA scheme. Performances are especially low for Pole vault and Baseball. All descriptors confuse Pole vault mainly with Weightlifting. For Baseball, $s$, OCIT, HOG, and MBH mainly confuse this category with Golf, Swimming, Golf, and Volleyball respectively. For categories like Golf and Bowling that have very distinct and limited set of movements, the categorization performance is very good.

**Results of Efficiency** Since most of the computation time is spent on optical flow calculation, using BoW or TA has negligible computational cost in the test phase. For example, while trajectory and descriptors computation takes $\sim 1.225$s for a 35s-video, BoW and TA for the MBH descriptor take average of 0.002$s$ and 0.018$s$ respectively. Thus, the categorization time using BoW and TA methods is almost the same, and they are on par with other trajectory-based methods [5, 17].

**4. CONCLUSIONS**

This paper proposes a sports video genre categorization method. We introduce a compact and efficient orientation-based trajectory shape descriptor that is invariant to camera angle variations. A temporal analysis method is introduced to integrate the decisions of local descriptors over time. Superior performance is observed on a dataset of amateur sports videos in the wild when compared to baseline methods.

We plan to extend this approach to a large sports video dataset and human activity datasets. Also, integrating trajectory descriptors with methods in body pose estimation [28, 29] can be another future direction.
5. REFERENCES


