Abstract—The bag-of-visual-words model has been widely utilized for content based image and video retrieval due to its scalability. In this paper, we extend this model for human action video retrieval. We adopt dense trajectory features which are able to achieve the state-of-the-art performance on action recognition, while most of the existing video retrieval methods utilize descriptors of local interest points. In order to improve similarity measurement between bag-of-visual-words model based representation, we propose to discover and incorporate spatial-temporal correlation (STC) among the trajectories in a given query video. The spatial-temporal correlation consists of spatial proximity and temporal consistence among trajectories, which is capable of strengthening discriminative power among visual words. Note that such query focused spatial-temporal correlation makes our method dynamic for different queries and is able to improve retrieval performance without significantly increasing the size of a visual vocabulary. The experimental results on an action video dataset demonstrate that our proposed method outperforms other similar methods.

I. INTRODUCTION

The popularity of video hosting services brings an explosive growth of user generated and shared video data in recent years. It has been more and more demanding to efficiently and effectively retrieve relevant video content for users from large scale video databases. Action content has been an important aspect of video content. Given a query video, content based action video retrieval aims to search for videos with similar action content. The task is very challenging, because the videos of the same action might have different visual appearances, viewpoint changes, occlusions, and camera motions.

The bag-of-visual-words (BoVW) model [1] has been widely used for content-based image retrieval due to its scalability, which was extended from the bag-of-words (BoW) model for textual information retrieval. A typical BoVW based retrieval process involves the following steps: Firstly, a set of local interest points or regions is detected through interest point detectors, or densely sampled from each image or video frame. Secondly, for each interest point, a high dimensional descriptor is extracted, such as SIFT [2] and SURF [3]. Thirdly, the descriptors are quantized into discrete visual words to form a visual vocabulary, and each image is represented as a bag of unordered visual words, so that textual information retrieval techniques such as inverted indexing can be applied for scalable retrieval. Finally, the similarity between two images are measured by the distance between their BoVW model based feature vectors.

However, such a model neglects the spatial context and correlation among individual interest points, which could compromise the discriminative power of visual words and eventually degrade retrieval performance. In order to rectify this issue, various methods have been proposed to incorporate spatial context of visual words into the BoVW model. In [4], the geometric consistency of visual words was used as a filtering step for similarity measurement. However, the filtering process would affect the efficiency of the retrieval process. In [5], [6], co-occurring pairs of visual words were used to form visual phrase, which leads to the increase of the size of the visual vocabulary. Similarly, in [7], the spatial context of visual words are modelled by a statistical language model. However, these methods eventually increase computational and storage costs of the retrieval process.

Recently, Wang et al. [8] proposed to utilize the BoVW model for content based video retrieval. In particular, spatial-temporal correlation (STC) among visual words was exploited to overcome the limitations of the BoVW model. However, each video was treated as a set of individual frames and each frame was characterized with the SIFT descriptors of local interest points in the frame. As a result, motion information which is essential for action videos is not explicitly taken into account. Meanwhile, in the field of action recognition, features derived from dense trajectories [9] are able to produce the state-of-the-art recognition accuracy, since trajectories obtained through optical flow from densely sampled interest points can effectively characterize global motion of local interest points. Therefore, in this paper, we utilize dense trajectory features to characterize an action video and propose to discover and utilize the spatial-temporal relations among dense trajectories for both efficient and effective action video retrieval.

As shown in Figure 1, trajectories A, B, and C are more important for characterizing waving action then trajectory D. It is noticed that the trajectories A, B, and C are spatially close and temporal consistent, which endorse each other to characterize the action. Therefore, instead of treating each
trajectory equally for similarity measurement, we weigh each trajectory differently in regard to its correlation with other trajectories. In addition, instead of encoding such information in the process of building the visual vocabulary, such spatial-temporal correlation will be exploited with a query video, which will not demand significant extra computational and storage resources. Specifically, for a query video, we define functions to measure the spatial proximity and temporal coherence between two trajectories, and adjust the weights accordingly in similarity computation. In this way, the discriminative trajectories for a query are identified with higher weights in similarity measurement. Note that our proposed retrieval process is still scalable due to the utilization of the bag-of-visual-words model and inverted indexing.

The rest of the paper is organized as follows. In Section II, we review the related works on content based video retrieval in terms of video content representation and spatial-temporal context in videos. In Section III, we introduce our dense trajectory based action retrieval method in terms of dense trajectory descriptors, video representation, and STC based similarity measurement. In Section IV, we present our experimental results and discussions. In Section V, we conclude the paper with discussions on our future work.

II. RELATED WORK

There have been many studies on content based image and video retrieval, as reviewed in [10][11]. In this section, we mainly focus on the studies based on the BoVW model.

A. Video Content Representation

Followed by the success in image domain, interest point detectors have been adopted to videos. To detect interest points in a video, a straight-forward approach is to apply a 2D detector, such as scale invariant feature transform (SIFT) [2], on individual frames. As an extension, some features have been devised to detect interest points in space-time volume of the video, such as space-time interest point (STIP) [12] or 3D-SIFT [13].

Recent research suggests that trajectory features are able to effectively characterize action content of videos [9]. Trajectories are typically formed by matching or tracking interest points. For example, in [14], trajectories are formed by matching SIFT descriptors between two consecutive frames. To improve the matching quality, a unique-match constraint is imposed among descriptors, and the matches that are too far apart are discarded. In [15], KLT tracker [16] is used to track interest points. The velocity of each point are quantized to form the “velocity history”, which is used as the feature representation. In dense trajectory method [9], trajectories are formed by tracking densely sampled points using optical flow fields. To improve the tracking quality, points in homogeneous areas without any structure are not tracked, and trajectories with no motion information or sudden large motion are removed. In this paper, we utilize dense trajectory to ensure the quality and quantity of the representation features of videos.

B. Spatial-Temporal Context in Videos

There exist various approaches to exploit the spatial and temporal context among local features of a video. In this section, we organize the existing studies into three categories by the way how they incorporate the extracted spatial/temporal information to the video retrieval process. In the first category, the spatial/temporal information can be used as verification constraints to filter out inconsistent matches of local features. For example, in the weak geometry consistency approach [4], the angle and scale parameters for the set of matching local features are verified. In [17], the RANSAC algorithm is used to estimate the transformation correspondences between images and filter out the “outliers”. This type of approaches are generally computationally expensive, as filtering needs to be applied to all candidate results.

In the second category, the spatial/temporal information can be used to form language models. For example, [6], [5] forms visual phrases by mining co-occurring pairs of visual words. This poses a challenge on storage as the size of visual vocabulary is greatly increased. In [14], Markov chain model is used to describe the inter and intra trajectory relationship, and in [7], a visual language model is built, which achieves comparable performance with 2D multi-resolution hidden Markov models. These methods generally increase computational and/or storage costs of the retrieval process.

In the last category, the spatial/temporal information can be used to adjust the weights the query words. In [8], the spatial-temporal correlations (STC) between local features of individual frames are used to reformatulate the query visual words. Compared to the above approaches, the method is more efficient, as only the STC of the query video is extracted. However, it may not be suitable for action videos since only local images features (i.e., SIFT descriptors) were utilized. Motivated by this work, in this paper, we utilize dense trajectory features to explicitly characterize action videos and devise STC among dense trajectories instead of individual keypoints in video frames.

III. DENSE TRAJECTORY BASED VIDEO RETRIEVAL

A. Dense Trajectory Descriptors

We utilize the method proposed in [9] to obtain dense trajectories and their descriptors. Given a video frame, feature
points are densely sampled in multiple scales and then tracked by optical flows to form trajectories. To ensure dense coverage, a new feature point is added into the tracking process whenever a large enough empty area is found without any tracking point. For each trajectory, the following descriptors are extracted as a trajectory descriptor: trajectory path, HOG (histogram of oriented gradients), HOF (histogram of optical flow), and MBH (motion boundary histogram).

Define a trajectory of length \( L \) (set to 15 in [9]) as \( \{P_1, P_1+1, \ldots, P_{i+L} \} \), the shape of a trajectory is defined as a series of displacements \( \{\Delta P_1, \ldots, \Delta P_{L+L-1} \} \), where \( \Delta P_i = P_{i+1} - P_i \) is the displacement between two consecutive frames. The resulting vector is \( L \times f \) where \( f \)

\[
T_{Disp} = \frac{\{\Delta P_1, \ldots, \Delta P_{L+L-1} \}}{\sum_{t=1}^{L+L-1} |\Delta P_t|}
\]

The HOG, HOF and MBH descriptors are obtained by first dividing the space-time volume of the trajectory into cells and computing the HOG, HOF, and MBH histograms for each cell, respectively. Finally, the trajectory displacement vector, HOG, HOF, and MBH descriptors are concatenated to form a 426 dimensional descriptor to characterize each trajectory.

B. BoVW Based Video Representation and Similarity Measurement

Following the BoVW model, the trajectory descriptors of all the videos are quantized by \( k \)-means algorithm to produce a trajectory vocabulary with trajectory words \( \{w_1, w_2, \ldots, w_K \} \), where \( K \) is the size of the vocabulary. As a result, each frame can be represented with the histogram of the trajectory words. Through the inverted indexing technique, a query will be compared with videos in a video database sharing trajectory words \( w_i \) with the query.

For a query video \( v_q = \{f_q\} \) and a database video \( v_d = \{f_d\} \), where \( f_q \) and \( f_d \) are a query frame and a database video frame represented with a \( K \) dimensional feature vector, respectively. Due to the redundancy in video content, downsampling of frame rate is often utilized to reduce computational complexity. In our experiments, one frame was chosen among every 5 frames. Similar to the BoW model in textual information retrieval, the similarity \( \text{sim}(f_q, f_d) \) between two frames \( (f_q \) and \( f_d) \) represented in histograms of trajectory words is measured with cosine function:

\[
\text{sim}(f_q, f_d) \approx \frac{\sum_{i=1}^{K} \text{score}(w_i)}{|f_q|^2 \times |f_d|^2},
\]

\[
\text{score}(w_i) = f_q(w_i) \times f_d(w_i),
\]

where \( |f| \) is the \( L_2 \)-norm of vector \( f \), and \( f(w_i) \) is the term frequency (TF) of the matched trajectory word between \( f_q \) and \( f_d \).

Then the video clip similarity measurement proposed by [18] is adopted to measure the similarity between two videos, where the highest score among all frame combinations between the two videos is taken as the final similarity:

\[
\text{sim}(v_q, v_d) = \max_{f_q \in v_q, f_d \in v_d} \text{sim}(f_q, f_d).
\]

Note that other more advanced similarity measurements can be adopted and explored in our work which is built on top of the inverted indexing technique, such as incorporating location information of individual trajectories.

C. STC based Similarity Measurement

As shown in Eqn. 3, the trajectory words are treated independently without taking the correlation among individual trajectories into account. As shown in Figure 1, correlation between two trajectories reflects a certain level of endorsement of each other towards their discriminative power in measuring similarity between two frames. Therefore, Eqn. 3 can be rewritten as follows to take the spatial-temporal correlation (STC) among trajectory words into account:

\[
\text{score}'(w_i) = \frac{1}{K} \sum_{j=1}^{K} f_q(w_i) \times st(i, j) \times f_d(w_j),
\]

where \( st(i, j) \) measures the STC between the \( i^{th} \) and \( j^{th} \) trajectory words.

Since the trajectories with strong spatial-temporal correlation generally come from the same moving objects or regions, most existing methods resort to object tracking and segmentation techniques which are computationally expensive with limited reliability and flexibility. Therefore, in this work, we decompose spatial-temporal correlation into two components: spatial proximity \( s_{i,j} \) and temporal coherency \( t_{i,j} \) between the \( i^{th} \) and \( j^{th} \) trajectories.

As indicated in Figure 1, the spatial correlation between a pair of trajectory words should be inversely proportional to their physical distance. In order to simplify the calculation of physical distance between trajectories, the position of each trajectory is represented as the mean \( (x, y) \) position of feature points along the trajectory, though other advanced proximity functions will be explored in our future study.

The spatial proximity between a pair of trajectory words, is approximated by the inverse of Euclidean Distance [19]. As shown in Figure 1, trajectory word A and B are closer in proximity compared with A and D, and need to be assigned with higher discriminative power.

As indicated in [20], the spatial layout of moving objects can be modeled by a Gaussian Mixture Model, the spatial proximity between a pair of trajectory words in a video frame can be formulated as follows:

\[
s_{i,j} = \sum_{a=1}^{N_f} \sum_{b=1}^{N_f} e^{-kd_{a,b}^2},
\]

where \( N_f(w_i) \) is the number of instances of \( w_i \) in frame \( f_q \), and \( d_{a,b} \) is the Euclidean Distance between the \( a^{th} \) instance of the \( i^{th} \) trajectory word and the \( b^{th} \) instance of the \( j^{th} \) trajectory word. The parameter \( k \) is analogous to the inverse of variance in Gaussian Mixture Model, which controls the decreasing speed of spatial correlation. The larger the \( k \) is, the more words are considered spatial correlated. In our experiments, \( k \) is empirically set to 1.
Temporal coherence can be measured by the similarity between the shapes of two trajectories which requires warping the samples along the trajectories. In this work, we simplify the motion trend as the mean trajectory displacement direction. As a result, the difference between the displacement directions of two trajectories can be used as an indication of temporal coherency.

Similarly, temporal coherence can be modeled as follows:

$$t_{i,j} = \sum_{a=1}^{N_{I}} \sum_{b=1}^{N_{I}} e^{-\gamma \| \Delta m_{a,b} \|^2} ,$$

(7)

where $\Delta m_{a,b}$ denotes the relative motion difference between the $a^{th}$ instance of the $i^{th}$ trajectory word and the $b^{th}$ instance of the $j^{th}$ trajectory word in the frame. The relative motion is computed by $\Delta m_{a,b} = m_a - m_b$, where $m$ denotes the mean motion vector (displacement direction) of the trajectory word. Similar to $k$, the $\gamma$, which controls the decreasing of temporal correlation along with the expansion of the relative motion, is set to 1.

The overall spatial-temporal correlation can be obtained by combining $s_{i,j}$ and $t_{i,j}$. In this work, a simple additive fusion is adopted as follows:

$$ST_{i,j} = \sum_{l=1}^{L} s_{i,j}^l = \sum_{l=1}^{L} (s_{i,j}^l + t_{i,j}^l) ,$$

(8)

where $L$ is the number of frames in $v_q$, and the superscribe $l$ denotes the $s_{i,j}$ and the $t_{i,j}$ measurement computed from the $l^{th}$ frame.

Finally, a STC matrix is computed by Eqn. 8 for each pair of trajectory words in the query video. In practice, since the frame vector $f$ is sparse, in case the trajectory $w_i$ does not exist in $f$, weight 0 is used as the default weight. Therefore, the STC matrix is also sparse.

IV. EXPERIMENTS AND DISCUSSIONS

A. Experimental Settings

We used the ICPR’04 dataset\(^1\) to evaluate our proposed method. The dataset was collected from 25 subjects who performed 6 actions in 4 scenarios. The actions in this video collection are: boxing, handclapping, handwaving, jogging, running, walking. Figure 2 gives a set of sample images from the video collection. As a result, there were $25 \times 6 \times 4 = 600$ video files in the dataset. Note that only 599 video files were used in our experiments since one video file was missing from the website\(^2\).

We compare our proposed method with two other popular local features, SIFT (Scale Invariant Feature Transform) [2] descriptors and STIP (Spatio-Temporal Interest Point) [12] descriptors, since SIFT descriptors were utilized in [8] and STIP descriptors have been widely used for action recognition.

\(^1\)http://www.nada.kth.se/cvap/actions/
\(^2\)http://www.nada.kth.se/cvap/actions/00sequences.txt
\(^3\)http://www.nada.kth.se/cvap/actions/actions.gif

The motion vector of a SIFT or STIP descriptor is computed by a pair of matched key point in sequential frames.

To build the code book, $k$-means algorithm was used to cluster the feature descriptors. We set $k = 8000$ which is the point where increasing the $k$ value will not significantly decrease the cost of the model in our experiments, and ran the clustering algorithm 3 times to reduce impact of initialization problem of the $k$-means algorithm. The traditional BoVW model is used as the baseline, which is referred to as without STC. Precision-recall and Mean Average Precision (MAP) were used as evaluation metrics.

B. Experimental Results

In this experiment, we use each video from the collection as the query to evaluate the retrieval performance across different methods.

As shown in Figure 3, our proposed method denoted as DT-stc (Dense Trajectory-spatial temporal correlation) performs best among all the other methods in comparison. It is also noticed that descriptors (e.g., dense trajectories descriptors and STIP descriptors) designed for action videos perform better than those designed for static visual content (i.e., images), such as SIFT descriptors. Interestingly, spatial temporal correlation clearly improves retrieval precision for dense trajectory features and STIP descriptors, however having little improvement on SIFT features, since SIFT features are not designed for characterizing action videos.

The retrieval performance for each action class is presented in Table I. Although STIP features perform comparable to dense trajectory features without STC, after STC is taken into account, dense trajectory features perform better than STIP, which could be contributed by the fact that dense trajectories are more suitable for discovering spatial temporal correlation among visual words.

We also show one retrieval result in Table II for a boxing action video with camera zooming. We can observe that

![Fig. 3. Comparison of different retrieval methods in terms of precision-recall curves.](image-url)
the results obtained without using STC have more irrelevant actions like running and handclapping, while our method taking STC into account is able to return more relevant action videos.

By further analyzing each action class, it is observed that action jogging and running performs relatively worse than the other classes. One reason would be that these two actions are ambiguous when observing from afar which is the case in video recording. The other reason could be that trajectory is not good at distinguishing the speed of movements.

V. Conclusions

In this paper, we present a novel similarity measurement for content based action video retrieval based on the scalable bag-of-visual-words model. Our proposed method utilizes one of the state-of-the-art action features derived from dense trajectories to represent each action video. In order to improve retrieval performance, we propose to discover and incorporate the spatial-temporal correlation among trajectories in a query video. As a result, our proposed retrieval method is able to achieve improved retrieval precision under the scalable retrieval framework without significantly increasing the complexity of the whole retrieval system. Since the correlation is mined from queries, our ranking process does not rely on obtaining prior knowledge of video databases. In our future work, we will further extend this method to large scale action video datasets, which could help video annotation and action recognition. We will also investigate advanced algorithms to mine more complex spatial-temporal patterns among trajectories and other representations such as Fisher Vectors [21] and Multi-skIp Feature Stacking (MIFS) [22] so that the retrieval performance of some action classes such as jogging and running which are difficult to differentiated in this work can be further improved.

Acknowledgements

This work is partially supported by ARC grants and Microsoft Azure Research Grant.

Table I

<table>
<thead>
<tr>
<th></th>
<th>DT</th>
<th>DT-stc</th>
<th>STIP</th>
<th>STIP-stc</th>
<th>SIFT</th>
<th>SIFT-stc</th>
</tr>
</thead>
<tbody>
<tr>
<td>boxing</td>
<td>48.08</td>
<td>55.11</td>
<td>56.14</td>
<td>80.77</td>
<td>9.65</td>
<td>21.61</td>
</tr>
<tr>
<td>handclapping</td>
<td>59.47</td>
<td>56.18</td>
<td>50.38</td>
<td>13.63</td>
<td>7.64</td>
<td>21.95</td>
</tr>
<tr>
<td>handwaving</td>
<td>54.79</td>
<td>73.10</td>
<td>44.09</td>
<td>21.43</td>
<td>7.59</td>
<td>20.59</td>
</tr>
<tr>
<td>jogging</td>
<td>31.25</td>
<td>36.43</td>
<td>32.77</td>
<td>29.26</td>
<td>32.11</td>
<td>22.92</td>
</tr>
<tr>
<td>running</td>
<td>29.27</td>
<td>43.82</td>
<td>31.56</td>
<td>32.39</td>
<td>28.89</td>
<td>24.57</td>
</tr>
<tr>
<td>walking</td>
<td>33.50</td>
<td>48.05</td>
<td>38.29</td>
<td>30.95</td>
<td>40.90</td>
<td>22.22</td>
</tr>
<tr>
<td>Average</td>
<td>42.73</td>
<td><strong>52.12</strong></td>
<td>42.21</td>
<td>34.74</td>
<td>21.13</td>
<td>22.31</td>
</tr>
</tbody>
</table>

Fig. 2. ICPR’04 video collection samples of 6 actions

REFERENCES

<table>
<thead>
<tr>
<th>Query Example (boxing)</th>
<th>Sample Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1.png" alt="Sample Frame" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First 10 Retrieval Results without using STC</th>
<th>First 10 Retrieval Results using STC</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.png" alt="First 10 Retrieval Results without using STC" /></td>
<td><img src="image3.png" alt="First 10 Retrieval Results using STC" /></td>
</tr>
</tbody>
</table>