Action recognition using restricted dense trajectories

To cite this article: Qinghui Li et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 322 062021

View the article online for updates and enhancements.
Action recognition using restricted dense trajectories

Qinghui Li*, Aihua Li, Zhigao Cui
502 Faculty, Xi'an Institute of High Technology, Xi'an, China

*Corresponding author e-mail: lqhui1212@126.com

Abstract. This paper presents an action recognition algorithm using restricted dense trajectories (RDT). In feature extraction step, restricted dense trajectories are obtained by tracking the refined points in optical flow field, which remove most of meaningless trajectories while preserve the discriminative power. Then we extract a new set of descriptors to capture the appearance and motion information of trajectories. For encoding step, we improve VLAD by assigning each descriptor to their K nearest words and employing these words as basis vectors to linearly approximate the descriptor under the minimum squared error criterion. Experimental results show the proposed algorithm obtains state-of-the-art results.

1. Introduction
Action recognition is a very active area in the fields of computer vision[1], owing to its broad range of applications, such as video surveillance, human-computer interaction, motion analysis and content-based video retrieval. Even though a large number of approaches have been proposed, it is still very challenging.

Building an efficient and discriminative representation for action is the key step for action recognition system. Wang et al. [2] tracked the densely sampled points using optical flow fields in each scale separately and computed local descriptors within the trajectory-aligned volume. Furthermore, they proposed improved dense trajectories (IDT) [3], which provided the state-of-the-art results on various datasets. However, a large number of feature points fall into the background area that is irrelevant to the action resulting in a substantial amount of redundant trajectories. Peng et al. [4] used a motion boundary based dense sampling strategy. Mukher et al. [5] introduced a 3D representation of IDT. In our work, we utilize an expanded human rectangular box to refine the dense points. Then, relative trajectory and spatio-temporal co-occurrence descriptors are extracted based on the refined dense trajectories, called restricted dense trajectories (RDT).

In recent years, VLAD [6] is a widely used encoding method. But it is difficult for VLAD to describe the distribution information of descriptors. Picard et al. [7] accumulated the second order tensor product and proposed VLAT. Peng et al. [8] computed two kinds of high-order statistics as complementary information. Unlike their work, we assign each descriptor to their K nearest visual words and employ these words as basis vectors to linearly approximate them under the minimum squared error criterion. Obtaining the linearly coefficients, we use them as weights to compute VLAD on the nearest multiple words.
2. Restricted dense trajectory descriptor extraction

2.1. Restricted dense trajectories

We first use deformable part based model to detect human in frames. Then the expanded human rectangular box is cropped by expanding the human rectangular box with 2 times of the width and 1.5 times of the height, as shown in Figure 1. Given the expanded human rectangular box, feature points within it are retained and others are removed. Then we track the remaining feature points to obtain restricted dense trajectories.

\[ RDT \]

\[ \mathbf{R} = \left[ (P_1, P_{i+1}, \ldots, P_{n-1}), (P'_1, P'_{i+1}, \ldots, P'_{n-1}) \right] \]

(1)

2.2. Relative trajectory descriptor

We extract the geometric center \( P_i \) of the expanded human rectangular box and track it to get its trajectory. Given the trajectories of \( P_i \) and feature points, the relative trajectory descriptor is defined as Equation (1).

2.3. Spatio-temporal co-occurrence descriptors

We extract our spatio-temporal co-occurrence descriptors in the cuboids align with each RDT. The size of cuboids is \( N \times N \times L \) pixels.
For the image block with the size WxH, the co-occurrence matrix is defined as Equation (2) when the offset is \((a, b)\)

\[
C(i, j) = \sum_{s=1}^{S} \sum_{t=1}^{T} \left\{ \frac{M(i+a, j+b) + M(i+a, j+b)}{2} \right\}, \text{if } O(i, j) = s \text{ and } O(i+a, j+b) = t; \quad \text{otherwise} \tag{2}
\]

where \(s\) and \(t\) are quantized orientation labels of gradient or optical flow, \(M(i, j)\) and \(M(i, j)\) are respectively the magnitude and orientation of gradient or optical flow at position \((i, j)\).

Spatial co-occurrence HOG (SCoHOG) counts gradient orientations of pixel pairs at a fixed distance from the single frame, and uses the resulting co-occurrence matrix as image representation. SCoHOF counts the optical flow orientations of pixel pairs, and SCoMBH counts the optical flow gradient orientations of pixel pairs. We use two offsets (i.e., \((2, 0)\) and \((0, 2)\)) and quantize the orientation into 8 bins with an additional bin for HOF.

Temporal co-occurrence HOG (TCoHOG) counts gradient orientations of pixel pairs over time align with the trajectory. The calculation methods of TCoHOF and TCoMBH are similar to the corresponding spatial feature, except that their co-occurrence units are extracted in temporal dimension. TCoHOF can convey the changes of motion orientation while TCoMBH reflect the alteration in the flow field. The offset of temporal co-occurrence units is set to \(\Delta t = 2\).

3. Improved VLAD encoding method

For descriptor \(x_i\), its \(K\) nearest visual words in the dictionary \(D\) are found to construct a sub-dictionary defined as \(D_i = [d_{x_1}, d_{x_2}, \ldots, d_{x_K}]\). These \(K\) words are employed as basis vectors to linearly approximate \(x_i\) under the minimum squared error criterion. Obtaining the linearly coefficients, we use them as weights to compute VLAD on the nearest \(K\) words. The steps of our improved VLAD are as follows:

**Algorithm: improved VLAD (IVLAD)**

**Input:** \(X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D \times N}\), \(D = [d_1, d_2, \ldots, d_M] \in \mathbb{R}^{D \times M}\), \(K, \alpha\)

**Output:** \(V = [v_1, v_2, \ldots, v_M] \in \mathbb{R}^{D \times M}\)

1: \(\text{for } m = 1 \text{ to } M \text{ do}\)
2: \(v_m = 0\)
3: \(\text{end for}\)
4: \(\text{for } i = 1 \text{ to } N \text{ do}\)
5: \(\text{Find the } K \text{ nearest words of } x_i \text{ as the local dictionary } D_i = [d_{x_1}, d_{x_2}, \ldots, d_{x_K}]\)
6: \(c = \arg \min_c \left\| x_i - \sum_{k=1}^{K} c_k d_{x_k} \right\|\)
7: \(\text{for } k = 1 \text{ to } K \text{ do}\)
8: \(v_{c_k} = v_{c_k} + c_k \times (x, d_{x_k})\)
9: \(\text{end for}\)
10: \(\text{end for}\)
11: \(\text{for } m = 1 \text{ to } M \text{ do}\)
12: \(v_m := \text{sgn}(v_m) \|v_m\|\)
13: \(\text{end for}\)
14: \(V := V\)

---

**Fig. 3** Schematic diagram of spatio-temporal co-occurrence feature descriptor extracting
4. Experimental Results

4.1. Datasets
We used three popular datasets in experiments: KTH, YouTube and HMDB51. KTH [9] consists of 2,391 video samples. YouTube [10] contains 11 categories of 1,168 video samples. And, HMDB51 [11] comprise 51 action categories.

4.2. Comparison to IDT descriptor
In this section, we compare RDT descriptors to original trajectories descriptors using the standard BoVW framework, as shown in Table 1. Dictionaries are constructed for each type of descriptor separately, and the number of visual words in every dictionary is fixed to 4000. Relative trajectory has improved the performance on the three datasets compared with the trajectory shape descriptor. The improvement on YouTube and HMDB51 are 3.5% and 4.3%, which indicates the validity of the trajectory spatial position. Since introduced spatio-temporary co-occurrence information, our two combinations of descriptors outperform the original combination of descriptors on the three datasets, which demonstrates they can encode the deep level information of appearance and motion.

| Descriptor                     | KTH       | YouTube  | HMDB51   |
|-------------------------------|-----------|----------|----------|
| Trajectory Shape (Baseline)   | 90.2%     | 67.2%    | 32.4%    |
| HOG+HOF+MBH (Baseline)        | 93.9%     | 83.4%    | 45.8%    |
| Relative Trajectory           | 90.9%     | 70.7%    | 36.7%    |
| SCoHOG+SCoHOF+SCoMBH          | 93.7%     | 80.9%    | 49.8%    |
| TCoHOG+TCoHOF+TCoMBH          | 94.8%     | 82.7%    | 54.8%    |
| SCo+TCo                       | 96.3%     | 84.6%    | 57.6%    |
| All combined                  | 97.2%     | 86.3%    | 60.9%    |

4.3. Comparison to other encoding methods
Table 2 presents the comparison of our IVLAD with FV and VLAD. We employ two feature extraction schemes for each encoding method respectively, which are the original IDT and our RDT. The size of dictionary for FV and IVLAD are set to 256, which is employed in many documents. For VLAD, We evaluate the size of dictionary with 256 and 512 visual words in order to make the comparison more fair.

Experimental results verify the importance of distribution information for enhancing action recognition. On the other hand, we also can see that our restricted dense trajectory descriptors yield better results than dense trajectory descriptor on three datasets. The best results, integrating RDT with IVLAD, are 98.1%, 89.9% and 70.5%, achieving the state-of-the-art results.

| Method | KTH       | YouTube  | HMDB51   |
|--------|-----------|----------|----------|
|        | IDT      | RDT      | IDT      | RDT      | IDT      | RDT      |
| VLAD 256 | 95.7% | 96.9% | 82.4% | 86.8% | 56.9% | 62.1%  |
| VLAD 512 | 96.6% | 97.5% | 84.6% | 87.5% | 58.4% | 63.7%  |
| FV     | 96.8% | 97.4% | 85.3% | 87.6% | 60.5% | 64.1%  |
| IVLAD  | 97.3% | 98.1% | 87.2% | 89.9% | 63.6% | 70.5%  |

5. Conclusion
Our restricted dense trajectories remove most of the meaningless trajectories and capture the spatial-temporal contextual information, which provide an efficient and discriminative representation for action. Regarding the encoding method, we introduce an improved VLAD to encode the distribution information of descriptors. The experimental results demonstrate the effectiveness of our approach.
References

[1] S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition: A survey. Image & Vision Computing, 2017, 60(4): 4-21.

[2] H. Wang, A. Klaser, S. Codella, and C. L. Liu. Action recognition by dense trajectories. IEEE Conference on Computer Vision and Pattern Recognition, 2011: 3169-3176.

[3] H. Wang, and C. Schmid. Action recognition with improved trajectories. International Conference on Computer Vision, 2013: 3551–3558.

[4] X. Peng, Y. Qiao, and Q. Peng. Motion boundary based sampling and 3D co-occurrence descriptors for action recognition. Image & Vision Computing, 2014, 32(9):616-628.

[5] S. Mukherjee, and K. K. Singh. Human action and event recognition using a novel descriptor based on improved dense trajectories. Multimedia Tools & Applications, 2017(6):1-18.

[6] H. Jégou, M. Douze, and C. Schmid. Aggregating local descriptors into a compact image representation. IEEE Conference on Computer Vision and Pattern Recognition, 2010: 3304-3311.

[7] D. Picard and P. H. Gosselin. Improving image similarity with vectors of locally aggregated tensors. International Conference on Image Processing, 2011: 669–672.

[8] X. Peng, L. Wang, Y. Qiao, and Q. Peng. Boosting vlad with supervised dictionary learning and high-order statistics. European Conference on Computer Vision, 2014: 660–674.

[9] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. International Conference on Pattern Recognition, 2004: 32-36.

[10] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from video in the wild. IEEE Conference on Computer Vision and Pattern Recognition, 2009: 1996–2003.

[11] H. Kuehne, H. Jhuang, and T. Serre. HMDB: A large video database for human motion recognition, IEEE International Conference on Computer Vision, 2011: 2556-2563.