A Global Spatio-Temporal Representation for Action Recognition

Chao Deng, Xiaochun Cao
School of Computer Science and Technology
Tianjin University, Tianjin 300072, China
{dengchao,xcao}@tju.edu.cn

Hanyu Liu, Jian Chen
School of Computing
University of Southern Mississippi, USA
{hanyu.liu,jian.chen}@usm.edu

Abstract

In this paper we introduce an effective method to construct a global spatio-temporal representation for action recognition. This representation is inspired by the fact that human actions can be treated as 3D shapes induced by the silhouettes in the space-time volume. We estimate the silhouettes which contain detailed shape information of the action, and present an efficient sampling method to extract interest points along the silhouettes. The local interest point is represented by a spatio-temporal descriptor based on 2D DAISY. Our global space-time representation is the integration of these local descriptors in an order along the silhouette. In this manner, we not only utilize the static shape information, but also the spatial-temporal cue. We have obtained impressive results on publicly available action datasets.

1. Introduction

Action recognition has been a hot research topic for decades and many solutions have been proposed. The usage of features to describe the action and choice of category models for classification are two critical problems in action recognition.


Studies by Gorelick et al. [4] have shown that human action in video sequences can be seen as silhouettes of a moving torso and protruding limbs undergoing articulated motion. In this paper, we propose a method to construct a global representation that combines the shape and spatial-temporal features, and integrate such representations with SVM classification schemes for recognition. Background subtraction and Canny edge detector are applied to get the silhouette of the human body. Then we obtain the shape of body, and eliminate the intra-pattern variations such as appearance, cloth, and light. We sample the interest points with equal spatial intervals along the human silhouettes. These interest points are efficient to represent the silhouette of an action. Due to the computational efficiency of the 2D DAISY [10], the local interest point is represented by a spatio-temporal descriptor based on 2D DAISY. Descriptors of these interest points are then integrated to construct a global representation of an action. The contributions of this work are as follows:

• A new representation that globally combines the shape and spatial-temporal features of the actions, and is more discriminative.
• Formulation of 3D DAISY descriptor which is an extension of DAISY from 2D image to 3D video.

The rest of the paper is organized as follows. In Section 2, we describe the automatic silhouette computation and the method to extract the interest points, as well as the strategy to determine the class that an action belongs to. Section 3 introduces the local spatio-temporal descriptor. We show experiment results in Section 4. Section 5 concludes this paper.

2. Shapes of the actions

In our approach, we regard action as a sequence of space-time shapes. So, we first compute the silhouettes of the human body in the videos. Assuming the camera is fixed, we divide videos into two sets: body separable actions and inseparable actions as shown in Figure 1.
The former denotes the actions where the body moves from one position to others. Typical examples include running and walking. The latter are the situations where the people stay at one spot, for example, waving hands and bending. Considering the distinct properties of the actions, we adopt different methods to get the silhouettes.

2.1 The silhouette of separable actions

For the videos of the separable actions, the background can be rebuilt easily. In this work we simply use the median value of all pixels in the temporal direction. As the background is obtained, we compute the shape of the body by subtracting the background from each frame, shown as in the first row of Figure 1. The pixels highlighted in green denote foreground.

2.2 The silhouettes of the inseparable actions

For the inseparable actions, it always has a part of background occluded by the human body. Therefore, it’s unlikely to rebuild the background. To handle this problem, we adopt canny edge detection.

For inseparable actions, the background subtraction method fails since there are areas of background consistently occluded by the human body. To handle this situation, we adopt Canny edge detector to extract the edges of human body as well as clutters in the background. To remove the edges from background, we use frame differencing method to get the max bounding box which tightly contains the actor. We treat edges outside the box as background and remove them as shown in the second row of Figure 1.

The remained edges inside the box are the silhouette of the human body. We morphologically close the image because usually the edges are not consecutive. The mask of the human body is obtained by flood filling inside the silhouette.

2.3 Extract interest points

Starting from its top point, we extract interest points with equal interval along the silhouette. Comparison of our sampling method and random pick in 3D SIFT [9] is shown in Figure 2. We focus on the points along the silhouette, because it accurately captures the information of an action. As shown in Figure 2(b), interest points sampled using our method is sufficient to represent a silhouette of an action. However, the random sampling will pick some insignificant points which are not informative for the action, as shown in Figure 2(c) and (d). After sampling, a 3D DAISY descriptor is constructed for each interest point. Then these local descriptors are integrated in clockwise from the top point to form a global representation of the action.

2.4 Action Classification

With the space-time representation of the actions, we use LIBSVM [1] to train representation models for each kind of action. The space-time representations are inputs as feature vectors. Suppose the representation set $V = \{ r_i \}$ of a video, where each representation $r_i$ will get a predicted label $p_{r_i}$ after classification. The score of representation $r_i$ for class $c$ is denoted by

$$ s^c(r_i) = \begin{cases} 1, & p_{r_i} = c; \\ 0, & p_{r_i} \neq c. \end{cases} $$

(1)

The final decision for a video is defined as:

$$ \hat{c} = \arg \max_{c \in \{1,2,...,C\}} \sum_i s^c(r_i), $$

(2)

where $C$ is the number of actions.

3 Create 3D DAISY descriptors

In this work, we adopt a recent descriptor, DAISY, and extend it to the spatio-temporal descriptor to represent the space-time video.

The original 2D DAISY is designed for dense wide baseline matching with less computational expenditure than the more widely used SIFT [5]. 2D DAISY is intrinsically invariant to rotation due to its particular design for the interest region.
3.1 Our 3D DAISY descriptors

As shown in Figure 3, we design our 3D DAISY descriptor as three concentric spheres with radius $R_1$ (red), $R_2$ (green), and $R_3$ (blue). The interest point is the center of the spheres. Let the center of spheres be $(x, y, t) = (0, 0, 0)$, four planes $\{(x, y, t)|t = 0\}$, $\{(x, y, t)|x = t\}$, $\{(x, y, t)|x = 0\}$, $\{(x, y, t)|x = -t\}$ intersect the 3 spheres, resulting in 12 great circles. The circles are denoted by $\Pi_{R, \theta}$, where $R$ is the radius of the circle, $\theta$ is the angle with plane $(x, y, t)|t = 0$. We evenly sample eight points on these circles. Without counting the redundant samples, a 3D DAISY descriptor has 79 samples. The sequence of the 12 circles is $\Pi_{R_1, 0}, \Pi_{R_2, 0}, \Pi_{R_1, \pi}, \Pi_{R_2, \pi}, \Pi_{R_1, \frac{\pi}{2}}, \Pi_{R_2, \frac{\pi}{2}}, \Pi_{R_3, \pi}, \Pi_{R_4, \pi}, \Pi_{R_3, \frac{\pi}{2}}, \Pi_{R_4, \frac{\pi}{2}}$.

In our approach, we compute the $x, y, t$ orientation gradients with specified Gaussian Kernel instead of 8 orientation maps in 2D-DAISY. In 3D DAISY, $h_{\Sigma}(u, v, t)$ is denoted by:

$$h_{\Sigma}(u, v, t) = [G^\Sigma_x(u, v, t), G^\Sigma_y(u, v, t), G^\Sigma_t(u, v, t)]^\top,$$

where $G^\Sigma_i$ denotes the $\Sigma$-convolved orientation map. Let $H_{R_i, \theta_j}$ be the vector made of values at circle $\Pi_{R_i, \theta_j}$:

$$H_{R_i, \theta_j} = (h_{\Sigma_x}(l_{\theta_j}(u_0, v_0, t_0, R_i)), h_{\Sigma_y}(l_{\theta_j}(u_0, v_0, t_0, R_i)), \ldots, h_{\Sigma_x}(l_{\theta_j}(u_0, v_0, t_0, R_i))),$$

where $l_{\theta_j}(u_0, v_0, t_0, R_i)$ is the location with distance $R_i$ from $(u_0, v_0, t_0)$, in the direction $\{\theta_j, \phi_k\}$. $\theta_j = \{0, \frac{\pi}{3}, \frac{2\pi}{3}\}$ denote the rotation of the circles, $\phi_k = \{0, \frac{\pi}{2}, \frac{\pi}{3}, \frac{2\pi}{3}, \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{6}, \frac{7\pi}{6}\}$ indicate the angles of the 8 points in a circle. The 3D-DAISY is denoted by:

$$\mathcal{D} = H_{R_1, \theta_1}, H_{R_2, \theta_1}, H_{R_3, \theta_1}, H_{R_1, \theta_2}, H_{R_2, \theta_2}, H_{R_3, \theta_2}, H_{R_1, \theta_3}, H_{R_2, \theta_3}, H_{R_3, \theta_3}, h_{\Sigma_x}(u_0, v_0, t_0)]$$

Note that the redundant samples will be ignored.

In our approach we use the radii $R_1 = 3$, $R_2 = 2R_1$, $R_3 = 3R_1$ and $\Sigma_1 = 2$, $\Sigma_2 = 2\Sigma_1$ and $\Sigma_3 = 4\Sigma_1$. Thus our descriptor is made of $79 \times 3 = 237$ dimensions.

4 Experiment

We have tested our approach the action dataset provided by [4]. It contains 93 videos of different people performing 10 actions: running, walking, skipping, jumping-jacks, jumping forward on two legs, jumping in place on two legs, jumping sideways, waving with two hands and waving with one hand. This dataset is a popular public benchmark used in many action recognition papers.

We perform leave-one-out cross-validation, which uses all the representation for training except for those belonging to the testing video. There are two parameters involved in our global representation: sampling interval in the frame level $n_f$, and the sample points per frame $n_p$. Let $B$ be the set of $(n_f, n_p)$, $n_f \in \{1, 2, 3, 4, 5, 6\}$ and $n_p \in \{8, 12, 16, 20, 24\}$. The results are shown in Figure 4 (a). Without surprise, the more frames we have (smaller $n_f$), the better the performance. The average accuracy increases from 85.38% by sampling per 6 frames to 92.04% sampling every frame. However, sampling more points per frame does not always award. One reason is that the global representation is sensitive to incorrect boundaries due to the very low quality of videos. Only weak spatial relationship benefits. For the dataset in [4], sampling eight points per silhouette gives the best performance. Note that, sampling more points does not hurt this algorithm much as shown in the last row of Figure 4 (a). The mean accuracy 89.28% is better than 82.6% reported in bag of words based local 3D SIFT [9].

To demonstrate the discriminative power of the global representation, we also compare the results with the original bag of words 3D SIFT. Due to the fact that SIFT is more discriminative than DAISY, we use 3D SIFT as our local descriptor during this comparison. Accuracies of our method using 3D SIFT in all the combinations in $B$ are shown in Figure 4 (b). The results in total and average accuracies, time and space costs are compared in Table 1. As shown in this table, our method using global representation performs well with 97.85% max accuracy and 94.91% average accuracy. It outperforms the original bag of visual word approach based on local spatio-temporal descriptors with an increase of 12.31%. Figure 5 shows the confusion matrix of our method using 3D SIFT as local descriptor and the
5 Conclusion

In this paper we have proposed an efficient global representation that accurately captures the global spatio-temporal information of an action in a video. The results have demonstrated its discriminative power comparing to the local spatio-temporal representations.

Acknowledgement

This work was supported by NSFC (No. 60905019, 50735003), SRF for ROCS, SEM, Open Projects Program of NLPR, Tianjin University 985 research fund, and SKL of PMTI.

References