Human Action Recognition Algorithm Based on Improved Dense Trajectories

Yuling Sun[1, 2, a], Peng Gan[1, 2], Yu Xiao[1, 2]

[1]School of Electronics and Information Engineering, Tianjin Polytechnic University, Tianjin 300387
[2]Tianjin Key Laboratory of Optoelectronic Detection Technology and Systems, Tianjin, 300387

*Corresponding author Email: [a]895492215@qq.com.

Keyword: Action recognition, Improved dense trajectories, Camera motion elimination

Abstract: Objective Human action recognition is an interesting but challenging task for unconstrained videos with complex background, illumination variation and camera motion. In this paper, we present an improved dense trajectory-based algorithm to improve the accuracy of human action recognition. First, dense optical flow is utilized to track the scale invariant feature transform key-points at multiple spatial scales. The histogram of oriented gradient，histogram of optical flow，and motion boundary histogram are used to depict the trajectory efficiently. Second, eliminating the influence of camera motion based on trajectory direction consistency. The purpose of this operation is to improve the robustness of trajectories. Finally, the Fisher vector model is utilized to compute one Fisher vector for each descriptor separately, and then the linear support vector machine is employed for classification. Experimental results on KTH and YouTube datasets demonstrate that the proposed algorithm can effectively recognize human actions.

Introduction

Human action recognition is widely used in video surveillance, content-based video retrieval and human-computer interaction. It is a hotspot in the field of pattern recognition and computer vision. Due to the effects of occlusion, illumination variation, complex background and camera movement, the recognition behavior in real scene still faces great challenges.

In recent years, the method of using local features to describe video has become increasingly popular. Laptev [1] extends the Harris to the 3D video sequence, known as the Space-Time Interesting Point (STIP) detection algorithm. Dollar et al. [2] used two-dimensional Gaussian filter and one-dimensional Gabor filter in the spatial domain and time domain filtering, in order to obtain more dense spatial-temporal characteristics. Scovanner et al. [3] coded the spatiotemporal information using sub-histograms to obtain 3D-SIFT features. Willems et al. [4] extended the 2-D SURF descriptor to 3-D based on the 3-D Hessian matrix method of time-space interest detection. Laptev et al. [5] combined histogram of oriented gradients (HOG) and histograms of optical flow (HOF) as local features of behavior recognition, and achieved good results.

Trajectory is the change of spatial feature points in time series, and can capture motion information more effectively than Space-Time Interesting Point (STIP). Messing et al. [6] used the KLT tracker to trace Harris3D feature points to extract the trajectory and obtain good results. Sun et al. [7] obtained trajectories by matching the SIFT descriptors between adjacent frames, and rejecting distant matches. Raptis et al. [8] track the feature points of the region of interest and connect the HOG and HOF descriptors along the trajectory by computing the tracking descriptors.

The final descriptors are applied to behavioral modeling and video analysis. Wang et al. [9] performed a dense sampling of each frame of the video, using the dense optical flow field to track the feature points to obtain the trajectory, and extracting the surrounding HOG, HOF and MBH features with the trajectory as the center, obtaining better result than other newest methods.

Wang et al. [10] used the pedestrian detection technique to estimate the motion of the camera, and achieved good experimental results. Since pedestrian detection is a difficult problem and has a high computational complexity, this paper proposes a method to remove camera motion based on improved dense trajectory. Experimental results show that this method can effectively eliminate the influence of camera motion on action recognition.

Improved dense trajectory

An algorithm based on improved dense trajectories is proposed. The flow chart of the algorithm is shown in Fig.1.

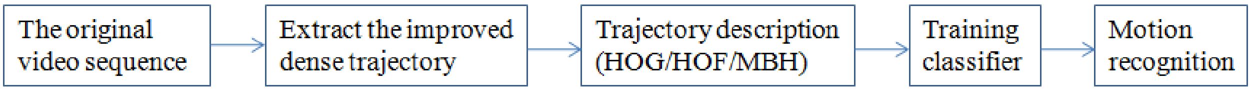

The original video sequence
Extract the improved dense trajectory
Trajectory description (HOG/HOF/MBH)
Training classifier
Motion recognition

Fig.1 Flow chart of the algorithm in this paper

Here, according to the method of literature [9], an image pyramid with 8 scales is set up for each frame, and the scale factor of the adjacent layer is$\sqrt{2}$. At the same time, the interval ω (ω=5) in each scale space is detected. Considering that there is no structural information in the flat area of the image, the criterion of Shi and Tomasi [11] is used to remove the point whose eigenvalue of autocorrelation matrix is less than a certain threshold. The autocorrelation matrix is calculated as in Equation (1) and the threshold is set as in Equation (2).

$$M = w(x,y)\begin{bmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{bmatrix} \tag{1}$$

$$T = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2) \tag{2}$$

Where, in the formula (1), I_x and I_y are the gradient of image I in the x and y directions respectively, and w is the window function. In (2), $\lambda^1{}_i$, $\lambda^2{}_i$ are eigenvalues of the autocorrelation matrix M at point i in the image I. By setting the threshold value, an effective feature point can be obtained.

For a given frame I_t, the dense optical flow field $\omega_t=(\mu_t, \nu_t)$, μ_t and ν_t are the horizontal and vertical components of the optical flow. Consider a point $P_{t+1}=(x_t, y_t)$ of the t frame, the position P_{t+1} in the t+1 frame is obtained by median filter method, as in (3).

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)\big|_{(x_t, y_t)} \tag{3}$$

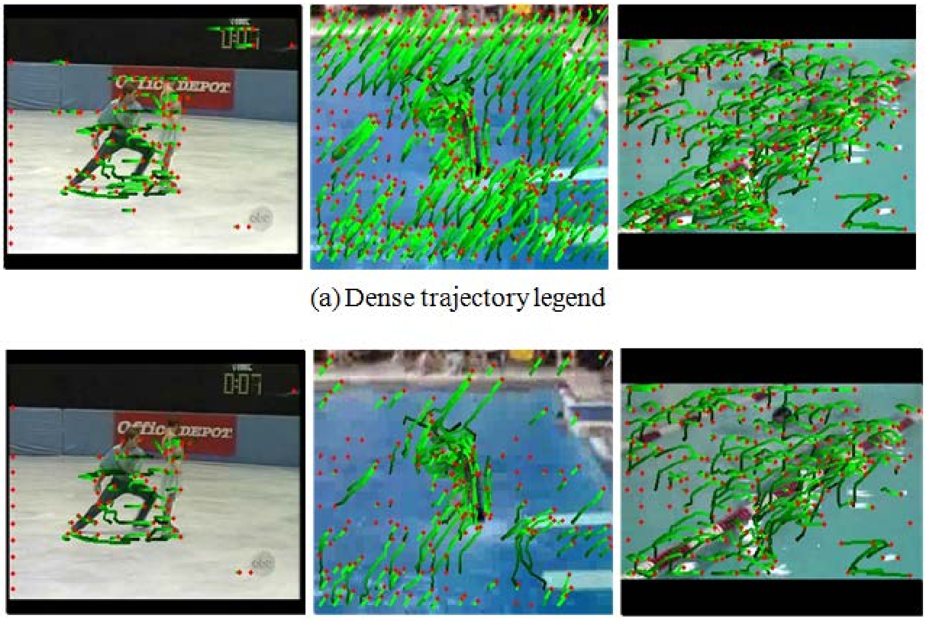
Where, M is the 3 × 3 median filtering kernel.

A trajectory is an ordered set of points in a video sequence (P_t,P_{t+1},P_{t+2},…). In order to avoid the problem of drifting in tracking, only L(L=15) frames are tracked in each spatial scale. If the displacement of trajectory is below the min-threshold or beyond the max-threshold, it is considered as an invalid trajectory. The results of dense trajectory are shown in Fig. 2 (a).

In real scene video, which camera motion is intense, causes the massive trajectory to exist in the background. In order to eliminate the influence of camera motion on action recognition, a method of camera action estimation based on trajectory direction consistency is proposed. First, the picture is divided into n × n grid region, and then statistics the direction of each grid area of the trajectory. To accurately represent the main direction of the trajectory, the histogram is divided into 18 bins in the direction of 0 to 360 radians. In order to facilitate the statistics, we give each bin a direction label, and set the value of labels {0,1,…17}. The peak value of the histogram represents the principal direction of trajectories in the mesh region, and is labeled as the direction of the mesh. The direction of the n × n grid is counted and the trajectory of the direction larger than the threshold is eliminated to obtain the final improved trajectory. Threshold *Th* is set as in Equation (4).The results of the improved trajectory are shown in Fig.2 (b).

$$Th = \max \sum_{i=1}^{25} x_i \qquad (4)$$

Where, x_{i} is the direction label of the i-th block and x_{i} is taken as [0,17]. Compared with the dense trajectory, the improved trajectory is better on eliminating most of the camera motion, as long as keeping the trajectory of the human action.



(a) Dense trajectory legend

(b) Improved trajectory legend

Fig. 2 (a) shows a dense trajectory, and Fig. 2 (b) shows an improved trajectory obtained by uniform screening in the direction.

The trajectory shape descriptor. Set the trajectory length L, then the trajectory shape descriptor is described by the sequence S=($\Delta P_{\mathrm{t}},\ldots,\Delta P_{\mathrm{t+L-1}}$), where S is the improved dense trajectory. The corresponding trajectory displacement vector $\Delta P_{\mathrm{t}}=(P_{\mathrm{t+1}}-P_{\mathrm{t}})=(x_{\mathrm{t+1}}-x_{\mathrm{t}},\mathrm{y}_{\mathrm{t+1}}-\mathrm{y}_{\mathrm{t}})$. The vectors are unitized by displacement vectors and:

$$T = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L\text{-}1} \left\| \Delta P_j \right\|} \qquad (5)$$

In the above formula, T is called trajectory shape descriptor. With the literature [15], track length L is set to 15, and then 30-dimensional trajectory shape descriptor is obtained.

Structure and motion descriptor. In addition to the trajectory descriptor, this paper constructs HOG/HOF and MBH features to represent human action, as shown in Fig.3. In the L consecutive frames, the N×N neighborhoods centered on the feature points are tracked to form N×N×L time-space pipes. In order to embed the structural information, N×N×L space-time pipes are divided into $n_\sigma \times n_\sigma \times n_\tau$ space-time grids, and local descriptors (HOG, HOF, and MBH) are calculated in each space-time grid. The descriptors are concatenated to form the final descriptor. Here, N is set to 32, L is set to 15, n_σ is set to 2, and n_τ is set to 3. With the literature [9], for each type of descriptor, take square root each dimension after L1 normalization.

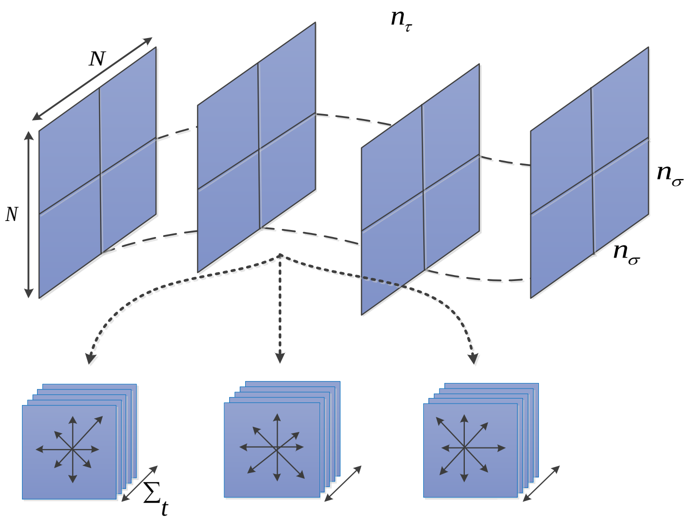


Fig.3 Local descriptor based on trajectory

Action classification

Fisher's vector is an extension of the Bag-of-visual-words (BOW) model. The Fisher kernel was first proposed by Jaakkol et al. [12]. Perronnin et al. [13] applied it to solve image classification problem, and had a good results. In order to reduce the computational complexity, we use Principal Component Analysis (PCA) to reduce the dimension of each descriptor to 1/2 of the original dimension. In the experiment, the number of Gaussian elements of the mixed Gaussian model is set to a fixed value K=256. Thus, each feature descriptors is represented as a Fisher vector of length 2KD. Where, D is the dimension of each descriptor after PCA dimension reduction. Finally, each Fisher vector is normalized by the square root and L2 normalization methods.

The literature [14] present that the linear support vector machine (linearSVM) is more suitable for Fisher vector model and can be used to obtain better experimental results with smaller computational cost. In the experiment, the combination of HOG, HOF, MBH to form the final feature, uses the one-against-rest method to solve the linear SVM multi-classification problem.

Experiment and analysis

Experimental data set. In order to verify the effectiveness of the algorithm in this paper, we evaluate this algorithm on the two public datasets (KTH, YouTube), as shown in Fig.4. The KTH dataset contains six action classes, including jogging, running, boxing, clapping, walking and waving. These action videos are shooting in 4 different scenes by 25 volunteers. We follow the original setup: the second, third, fifth, sixth, seventh, eighth, ninth, tenth and twenty-second individuals were used as test samples, and the rest as training samples. The YouTube dataset contains 11 action classes: cycling, shooting, diving, horse riding, dancing, swing, tennis, trampolining, golfing, volleyball and dog walking. Due to complex background, a large number of camera motion, illumination variation, scale invariant and other factors, it is more challenging. The experiments use leave-one-out cross-validation to output the average accuracy of each class.

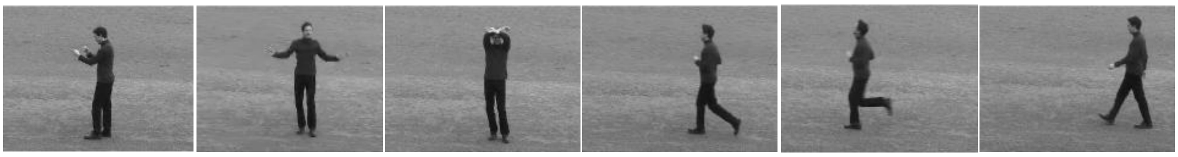


Fig. 4 Datasets used in the experiment

Experimental results and analysis

Eliminates the effects of camera motion. Table.1 is the experimental results of improved trajectories and dense trajectories. During the experiment, the improved trajectories and dense trajectories use the same parameters. HOG, HOF, and MBH represent experimental results using only the corresponding feature, and the results of combining the three types of features are shown in combination.

Table 1 Compare improved trajectory with dense trajectory

Action recognition method	Features	KTH Dataset (%)	YouTube Dataset (%)
Dense trajectory	HOG	87.0	72.6
	HOF	93.3	70.0
	MBH	95.0	80.6
	Combination	94.2	84.1
Improved trajectory	HOG	86.9	72.8
	HOF	93.7	71.5
	MBH	96.0	82.1
	Combination	95.3	86.7

For the HOG descriptor, the use of improved dense trajectory does not significantly improve the recognition result. The improved trajectory could eliminate a part of the background trajectory, so the number of features may be reduced and then the representation of HOG descriptor is not effective.

For the HOF and MBH descriptors, the average recognition rate is 1.1% and 2.6% higher in the two datasets using the improved trajectory than the dense trajectory. Because of camera motion, the dense trajectory algorithm can not be a good representation of motion information. The improved trajectory algorithm restrains a large number of background trajectories caused by camera movement through the consistent information, so the recognition result of improved trajectory algorithm is better than that of dense trajectory.

Table 2 Comparison to the state-of-the-art results

Datasets	Behavior recognition method	Year of Publication	Average recognition accuracy /%
KTH	literature [9]	2011	94.2
	literature [8]	2013	94.8
	literature [15]	2014	93.2
	literature [16]	2014	94.4
	Improved trajectory(ours)	—	95.3
YouTube	literature [9]	2011	84.1
	literature [17]	2012	85.3
	literature [18]	2010	75.2
	Improved trajectory(ours)	—	86.7

Compared with the most advanced results. Table 2 is the average recognition accuracy of our improved human action recognition algorithm and the state-of-the-art experiment results in the recent years. As can be seen from Table 2, compared with other literature methods, our method can still significantly improve the behavior recognition results. It should be noted that the method based on extracted optical flow information to calculate the trajectory direction, will not bring about significant computation burden. At the same time, the reduction of memory space for trajectory can reduce the computation amount of subsequent descriptors. This is more pronounced for YouTube datasets where there is a lot of camera movement.

Conclusion

In order to effectively identify human behavior in natural environment video, this paper proposes an algorithm based on improved dense trajectory for human action recognition. The algorithm uses the direction uniformity of the background trajectory to suppress the influence of the motion of the camera, so as to obtain more robust trajectory characteristics. Experiments on KTH and YouTube datasets demonstrate the feasibility and effectiveness of the algorithm.

As the algorithm needs to deal with every frame of video, when the resolution of video is very high, it is difficult to calculate the video features in real time. Future work will focus on how to reduce the computational complexity of algorithms and how to improve the algorithm recognition results.

References

[1]. Laptev I On space-time interest points. International journal of computervision. Vol. 64(2005) No. 23, p. 107-123.

[2]. DollarP, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features. Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2005, p. 65-72.

[3]. Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. Proceedings of the 15th international conference on Multimedia. ACM, 2007, p. 357-360.

[4]. Willems G, Tuytelaars T, Gool L. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. European Conference on Computer Vision. Marseille, France, October 12-18, 2008, p. 650-663.

[5]. Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies. Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015, p.1-8.

[6]. Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints. Computer Vision, 2009 IEEE 12th International Conference on.Vol. 30(2009)No. 2, p. 104-111.

[7]. Sun J, Wu X, Yan S, et al. Hierarchical spatio-temporal context modeling for action recognition. Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009, p. 2004-2011.

[8]. Raptis M, Soatto S. Tracklet Descriptors for Action Modeling and Video Analysis. Computer Vision - ECCV 2010, European Conference on Computer Vision. Heraklion, Crete, Greece, September 5-11, 2010, p. 577-590.

[9]. Wang H, Klaser A, Schmid C, et al. Action recognition by dense trajectories. IEEE Conference on Computer Vision & Pattern Recognition.ColoradoSprings ,USA, 2011, p. 3169-3176.

[10]. Wang H, Schmid C. Action Recognition with Improved Trajectories. IEEE Conference on International Conference on Computer Vision. Sydney, 2013,p.3551-3558.

[11]. Shi J, Tomasi C. Good Features to Track. IEEE Conference on Computer Vision and Pattern Recognition.Seattle WA , USA, 1994,p. 593-600.

[12]. Jaakkola T S, Haussler D. Exploiting Generative Models in Discriminative Classifiers. Advances in Neural Information Processing Systems, Vol. 11(1998) No. 11, p. 487--493.

[13]. Perronnin F, Sánchez J, Mensink T. Improving the Fisher Kernel for Large-Scale Image Classification. Computer Vision - ECCV 2010. European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, p. 119-133.

[14]. Perronnin F, Dance C. Fisher Kernels on Visual Vocabularies for Image Categorization. IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007, p 1-8.

[15]. Mota V F, Perez E A, Maciel L M, et al. A tensor motion descriptor based on histograms of gradients and optical flow. Pattern Recognition Letters. Vol. 39 (2014) No. 4, p. 85-91.

[16]. Wang T, Wang S, Ding X. Detecting Human Action as the Spatio-Temporal Tube of Maximum Mutual Information. IEEE Transactions on Circuits & Systems for Video Technology. Vol. 24 (2014) No. 2, p. 277-290.

[17]. Bandouch J, Jenkins O C, Beetz M. A Self-Training Approach for Visual Tracking and Recognition of Complex Human Activity Patterns. International Journal of Computer Vision, Vol. 99 (2012) No. 2, p. 166-189.

[18]. N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions:Combining multiple features for human action recognition.European Conference on Computer Vision. Heraklion, Crete, Greece, September 5-11, 2010, p. 494-507.