applied sciences

MDPI

Article

Human Action Recognition Based on Foreground Trajectory and Motion Difference Descriptors

Suge Dong [1], Daidi Hu [1], Ruijun Li [1] and Mingtao Ge [1,2,*]

[1] Sias University, Zhengzhou 451150, Henan, China; dongsugett@126.com (S.D.); hudaidi123@126.com (D.H.); lirj100@163.com (R.L.)

[2] School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China

* Correspondence: stevenge@gs.zzu.edu.cn; Tel.: +86-150-9338-0110

Received: 1 April 2019; Accepted: 14 May 2019; Published: 24 May 2019



Abstract: Aimed at the problems of high redundancy of trajectory and susceptibility to background interference in traditional dense trajectory behavior recognition methods, a human action recognition method based on foreground trajectory and motion difference descriptors is proposed. First, the motion magnitude of each frame is estimated by optical flow, and the foreground region is determined according to each motion magnitude of the pixels; the trajectories are only extracted from behavior-related foreground regions. Second, in order to better describe the relative temporal information between different actions, a motion difference descriptor is introduced to describe the foreground trajectory, and the direction histogram of the motion difference is constructed by calculating the direction information of the motion difference per unit time of the trajectory point. Finally, a Fisher vector (FV) is used to encode histogram features to obtain video-level action features, and a support vector machine (SVM) is utilized to classify the action category. Experimental results show that this method can better extract the action-related trajectory, and it can improve the recognition accuracy by 7% compared to the traditional dense trajectory method.

Keywords: Action recognition; foreground trajectory; motion difference; SVM

1. Introduction

Human behavior recognition is an important research area in intelligent video analysis. It is widely used in human–computer interactions, video surveillance, motion analysis, virtual reality, and other fields. Usually, the realization of video human behavior mainly includes two steps: one is feature extraction based on video information, the other is video classification based on feature vectors; however, because of the problems of background interference, partial occlusion, and different viewpoints, it is still a challenging task to obtain efficient and accurate motion representation [1–3].

In recent years, the trajectory-based method has achieved great success in the field of behavior recognition [4–6]. Unlike the method of directly extracting local features, the trajectory-based method extracts space–time trajectories by matching feature points between adjacent frames and then representing human behavior [7–10]. Yun [11] used scale-invariant feature transform (SIFT) to match and track spatiotemporal context information between adjacent frames, and Matikainen [12,13] used the Kanade–Lucas–Tomasi (KLT) optical flow method to track feature points between adjacent frames and extract trajectories. Although these methods can capture the movement trends of behavior, there are still some problems. These are all sparse tracking methods, and sparse trajectories cannot adequately represent complex human behavior. In order to solve this problem, Wang [14] used dense optical flow to sample dense feature points on regular grids, forming dense trajectory (DT). Four descriptors, histogram of oriented gradients (HOGs), optical flow (HOF), motion boundary histograms (MBHs), and trajectory shapes (TSs), are used to represent human behavior [15,16]. However, Vig [17] found

that accuracy of human behavior recognition was maintained only by randomly selecting 30%–40% of descriptors in dense trajectory-based methods. This means that although dense trajectory descriptors can obtain a lot of information, most of them are redundant or even irrelevant.

In view of the above analysis, a human behavior recognition method is proposed based on foreground trajectory and motion differential descriptors. Firstly, the optical flow method is used to estimate the motion amplitude of each frame, and the foreground area is determined according to the motion amplitude of the pixels. Based on the extracted foreground area, the initial sampling point of the trajectory is placed, which reduces interference information caused by the background compared with the global dense trajectory. Secondly, in order to better describe the relative time information between different behaviors, a motion difference descriptor (MDD) is introduced to quantify the foreground trajectory. A direction histogram of motion difference is constructed by calculating direction information of the motion difference per unit time of the trajectory point. Finally, a behavior descriptor based on motion difference is constructed. A Fisher vector (FV) is used to encode the local features to the global features at the video level, and the features are inputted into a support vector machine (SVM) to determine the behavior category. Experimental results show that the proposed method can extract the trajectory more accurately than the traditional optical flow method and, thus, achieve higher recognition accuracy.

2. Foreground Trajectory Extraction

The first step in foreground estimation extraction is to select the initial sampling point. Compared with sampling the background region, the sampling points of the foreground salient region often contain more motion information; therefore, selection of key points is very important. For a pixel $P(x, y)$ in a frame, $H(x, y)$ is the Hessian matrix [18] of the pixel P, as shown in Formula (1).

$$H(x, y) = \begin{bmatrix} P_x^2 & P_x P_y \\ P_x P_y & P_y^2 \end{bmatrix}. \quad (1)$$

P_x and P_y are the first derivatives of the horizontal and vertical directions of the pixel P, respectively. If the determinant value $\left|H(x, y)\right|$ is greater than the threshold Ths, then P is considered an alternative key point. But in video, the key points of each frame will change with time. It is not enough to select the key points only by using the Heisen matrix. Therefore, this paper introduces an optical flow magnitude of each frame as an auxiliary measure of the key points.

Firstly, the dense optical flow field of each frame is calculated and expressed as $\omega = (u, v)$, where u and v represent the horizontal and vertical components of the optical flow field, respectively. For a point $P(x, y)$ in a frame, the definition of optical flow magnitude is shown in Equation (2).

$$M(x, y) = \sqrt{\left(u\big|_{x,y}\right)^2 + \left(v\big|_{x,y}\right)^2}, \quad (2)$$

where $u|_{x,y}$ and $v|_{x,y}$ represent the displacement values of points $P(x, y)$ in horizontal and vertical directions. Generally speaking, the motion amplitude of background points is less than the average motion amplitude of frames because the foreground points are usually less than the background points, and these background points are usually static. However, the contribution of background points to behavior recognition is often relatively small. In view of this, this paper introduces the idea of adaptive amplitude threshold, defined as $Th_m = \overline{M}$, where $\overline{M}$ is the average motion amplitude of the current frame. To sum up, for a certain pixel $P(x, y)$, if it satisfies $Th_m > \overline{M}$ and $\left|H(x, y)\right| > Ths$, then P is selected as the sampling point of the foreground trajectory

For a sampling point $P_t(x_t, y_t)$ in frame t, the corresponding position of the sampling point in frame $t + 1$ is defined as Equation (3).

$$P_{t+1} = (x_t, y_t) + (K \cdot \omega_t)\big|_{(x_t, y_t)}, \quad (3)$$

where K is a 3×3 median filter, and ω_t is a two-channel optical flow matrix of the first t frame. A motion trajectory (P_t, P_{t+1}, P_{t+2}, ... , P_{t+L}) is formed by connecting the sampling points in the continuous L frames in series. At the same time, in order to avoid the drift of sampling points, the sampling step and trajectory length are set to 5 and 15, respectively.

3. Second-Order Motion Descriptor

Many descriptors have been successfully applied to human behavior such as trajectory shape, HOG, HOF, and MBH. They describe trajectory shape characteristics, static behavior characteristics, motion velocity characteristics, and motion boundary characteristics from different aspects. Among them, both HOF and MBH are descriptors based on motion features, and they have better recognition effects than static descriptors [17]. Motion feature is an important part of human behavior recognition. Therefore, in order to better describe motion information contained in videos, the motion differential descriptor (MDD) was obtained by calculating temporal variation in optical flow along the trajectory, which was used to represent the relative motion information of the behavior.

For a trajectory $\{(x_{t+l}, y_{t+l})\big|l = 0, 1, 2, \ldots, L-1\}$ in frame t, L is the length of the trajectory, and MDD is defined as the time derivative of the optical flow displacement along the trajectory direction. Assuming that point $(\widetilde{x}_{t+l}, \widetilde{y}_{t+l})$ is any point in the neighborhood of point (x_{t+l}, y_{t+l}) on the trajectory, and point $(\widetilde{x}_{t+l+1}, \widetilde{y}_{t+l+1})$ is the corresponding point at the same position in the spatial neighborhood of point (x_{t+l+1}, y_{t+l+1}) on the trajectory in the next frame, the time derivative of the optical flow along the trajectory of point $(\widetilde{x}_{t+l}, \widetilde{y}_{t+l})$ is defined as:

$$\Delta u_{t+l}\Big|_{(\widetilde{x}_{t+l}, \widetilde{y}_{t+l})} = u_{t+l+1}\Big|_{(\widetilde{x}_{t+l+1}, \widetilde{y}_{t+l+1})} - u_{t+l}\Big|_{(\widetilde{x}_{t+l}, \widetilde{y}_{t+l})}; \tag{4}$$

$$\Delta v_{t+l}\Big|_{(\widetilde{x}_{t+l}, \widetilde{y}_{t+l})} = v_{t+l+1}\Big|_{(\widetilde{x}_{t+l+1}, \widetilde{y}_{t+l+1})} - v_{t+l}\Big|_{(\widetilde{x}_{t+l}, \widetilde{y}_{t+l})}. \tag{5}$$

Firstly, the optical flow derivatives of each point in the trajectory and its neighborhood are calculated. Then, the optical flow derivatives are accumulated into an 8-bit directional histogram according to the magnitude of the optical flow. We can get the motion difference descriptor f_l for a point in the trajectory. Based on this, the optical flow time derivative of the machine neighborhood of the trajectory point is calculated and quantified according to the trajectory length L. Finally, the motion difference descriptor of the trajectory is obtained as follows:

$$MDD = \frac{1}{L}\sum_{0}^{L-1} f_l. \tag{6}$$

4. Algorithmic Implementation Steps

The framework of the human behavior recognition algorithm based on foreground trajectory and second-order motion difference descriptors is shown in Figure 1.

The algorithm steps are as follows:

(1) Video streams were inputted, and dense optical flow fields were calculated. Sampling candidates with large amounts of information were initially selected.

(2) According to the magnitude of optical flow in each frame, the foreground region closely relating to behavior was determined, and the sample candidate points unrelated to the foreground were eliminated to obtain the starting point of the trajectory.

(3) The foreground trajectory starting point was tracked in eight scale spaces. The foreground trajectory was formed, and the scaling factor between each two scale spaces was $1/\sqrt{2}$. In this paper, the maximum length L of the trajectory was set to 15 to avoid the phenomenon of feature point drift.

(4) A space–time block with a size of $3 \times 3 \times 2$ was constructed along the foreground trajectory, and feature descriptors were extracted. HOG, HOF, MBH, and MDD proposed in this paper were used as human behavior descriptors, and a 492-dimensional feature vector was obtained, which

contained 96-dimensional HOG features ($2 \times 2 \times 3 \times 8$), 108-dimensional HOF features ($2 \times 2 \times 3 \times 9$), 192-dimensional MBH features ($2 \times 2 \times 3 \times 8 \times 2$), and 96-dimensional MDD features ($2 \times 2 \times 3 \times 8$).

(5) The 492-dimensional global eigenvector was optimized, and dimensionality reduction was carried out. Principal component analysis (PCA) was introduced. The ratio of feature dimensions before and after dimensionality reduction was set to 2:1, and the final feature length was $d = 246$.

(6) In order to further improve the ability of global features to represent video behavior, the Gauss mixture model (GMM) and FV were used to re-encode the features. In this paper, the number of Gauss kernels $K = 256$ was selected. Finally, an expectation maximization (EM) algorithm was used to solve the final feature dimension of $2 \times d \times K$ (i.e., 125,952 dimension).

(7) Encoded features were fed into the SVM for classification.

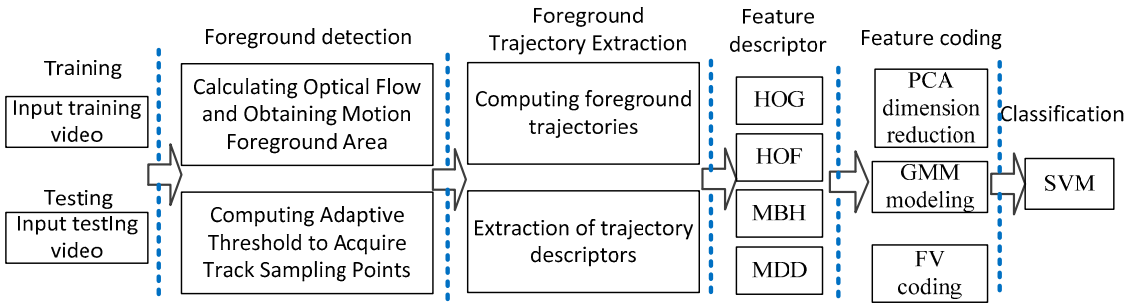


Figure 1. Flow chart of the human behavior recognition algorithm. HOG, histogram of oriented gradient; HOF, optical flow; MBH, motion boundary histogram; MDD, motion difference descriptor; PCA, principal component analysis; GMM, Gauss mixture model; FV, Fisher vector; and SVM, support vector machine.

5. Experimental Results and Discussion

In order to verify the effectiveness of the human behavior recognition algorithm in this paper, experimental tests were conducted on two real scene datasets, YouTube [19] and UCF sports [20]. The YouTube dataset had 11 behavior categories as well as 1168 fragments that were divided into 25 groups. The UCF sports dataset had 10 sports categories from sports videos. Because of the interference of perspective, background, and light, there were large intraclass differences and some challenges. The software environment used in the experiment was OpenCV + Visual Studio. In both datasets, the ratios of the number of videos in the training set and test set were set to 2:1.

Firstly, this paper calculated the optical flow of the video to get the foreground region, and it extracted the trajectory based on the foreground region. Taking the UCF sports data set as an example, the results were compared with those in reference [12], as shown in Figure 2. From left to right are the original map, the traditional trajectory, and the foreground trajectory. As can be seen from Figure 2, the foreground trajectory obtained after foreground extraction suppressed the interference caused by background noise to a certain extent, greatly reducing the number of trajectories and improving the speed and quality of subsequent feature coding.

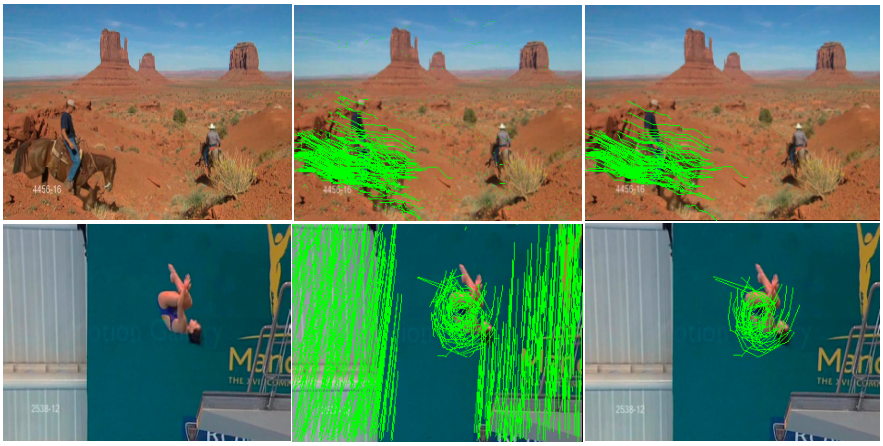
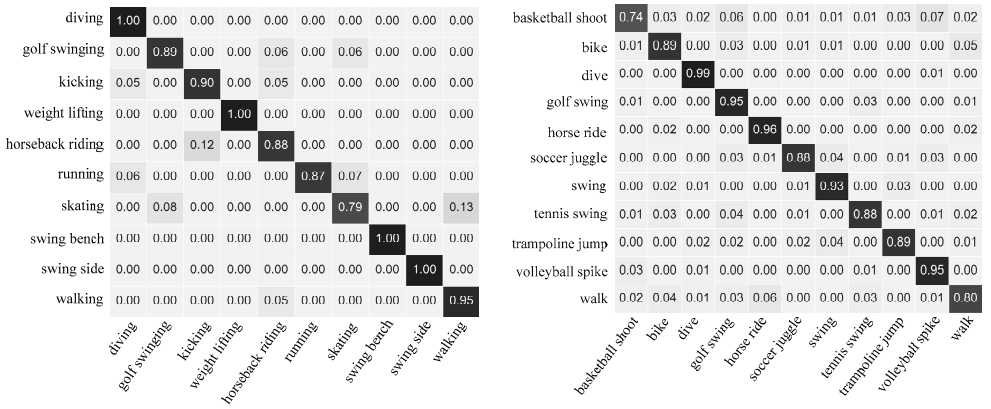


Figure 2. Comparison of traditional dense trajectory and foreground trajectory extraction in this paper.

Many works proved the recognition abilities of HOG, HOF, and MBH; therefore, these three descriptors were used as basic descriptors in this paper. At the same time, in order to verify the validity of the motion difference descriptor, the recognition ability of a single motion difference descriptor was tested and combined with three basic descriptors. Experimental results are shown in Table 1. Combination 1 represents the combination of three basic descriptors: HOG, HOF, and MBH. Combination 2 represents the combination of three basic descriptors and the motion difference descriptors mentioned in this paper, namely HOG, HOF, MBH, and MDD. "Combination 2" added to MDD was superior to the combination of traditional basic descriptors in recognition rates of both benchmark data and above, which illustrated the effectiveness of the method proposed in this paper. In addition, in order to show more intuitively the accuracy of classification and recognition results obtained by this algorithm, confusion matrices of recognition results on UCF sports and YouTube datasets were calculated respectively, as shown in Figure 3.

Table 1. Comparison of recognition results of different descriptors.

Descriptor	UCF Sports (%)	YouTube (%)
HOG	84.3	74.3
HOF	79.2	70.9
MBH	84.5	81.5
MDD	87.5	81.3
Combination 1	90.4	87.0
Combination 2	**92.8**	**89.6**



(**a**) Obtained obfuscation matrix on the UCF sports data set

(**b**) Obtained obfuscation matrix on the YouTube data set

Figure 3. Obfuscation matrixes obtained by the algorithm in this paper on different data sets.

From the obfuscation matrix in Figure 3 we saw that in the UCF sports data set, because there were some similarities between "skating" and "walking" actions, some "skating" actions were recognized as "walking" actions, so there was a false recognition rate of 0.13 between them. In the YouTube dataset, because of the instability of samples and the complexity of background, it was difficult to identify the dataset itself. A high false recognition rate was seen with "basketball shoot" because, compared to other actions, this action was more complex and was composed of multiple actions, so it was very difficult to identify.

Finally, in order to evaluate the overall performance of the algorithm, the recognition accuracy was used as the evaluation index, and the results were compared with those of references [2,3,10,14] on UCF sports and YouTubes datasets, respectively. The comparison results are shown in Table 2.

Table 2. Comparison of recognition accuracy with other similar algorithms.

Methods	UCF Sports (%)	YouTube (%)
Reference [2]	90.0	85.3
Reference [3]	92.0	88.9
Reference [10]	91.4	87.6
Reference [14]	89.7	86.1
Proposed	**92.8**	**89.6**

As can be seen from Table 2, the recognition accuracy of this algorithm was 92.8% on the UCF sports data set, which was 2.8%, 0.8%, 1.4%, and 2.9% higher than that of references [2,3,10,14], respectively. On the YouTube data set, the recognition accuracy of this algorithm was 89.6%, which was 4.3%, 0.7%, 2.0%, and 3.5% higher than that of references [2,3,10,14], respectively. Recognition accuracy was improved to a certain extent. This was because the proposed algorithm more accurately extracted trajectories related to behavior and ensured the accuracy of behavior representation. Meanwhile, the motion time was modeled effectively by MDD, which improved the recognition accuracy of the whole algorithm.

To evaluate the computational complexity of the proposed foreground trajectory, we computed the average time consumption per video for both trajectory extraction and descriptor representation on UCF sports. The experiment compared the proposed trajectory with reference [12] with Intel I7 (3.6GHz CPU), only a single CPU core was used, and both of the codes were not optimized. Time consumptions for the above two trajectories were compared in Table 3. Extraction of the foreground trajectory was somewhat more computationally expensive than [12]. This was because foreground trajectory used a motion-based sampling strategy, and it needed to traverse every moving point in the next frame, which was time consuming. But the number of the foreground trajectory was much smaller than [12], so the descriptor computation and encoding process for the former saved much more time. In addition, we did not count the time consumption of optical flow for PCA and GMM training because the processes between these two types of trajectories were similar.

Table 3. Comparison of the average time consumption per video for the trajectory.

	Trajectory Numbers	Trajectory Extraction (s)	Descriptor Computation (s)	Descriptor Encoding (s)	Total (s)
Reference [12]	26739	6.04	50.18	9.17	87.69
Proposed	2344	18.15	5.39	1.86	28.40

6. Conclusions

In view of the redundancy of the traditional dense trajectory method and the insufficiency of background interference, a foreground trajectory extraction method is proposed. Firstly, the motion amplitudes between video frames are estimated by optical flow method. The foreground region of motion is determined according to the motion amplitude of pixels, and the sample points of trajectory

are placed in the foreground region, which reduces the behavior-independent trajectory to a certain extent. The dense trajectory is extracted from the main action area, and the trajectory descriptor is constructed along the trajectory. Secondly, in order to better describe the relative time information between different behaviors, a motion difference descriptor is introduced to describe the foreground trajectory, and a motion difference direction histogram is constructed to represent the relative time information of motion. Finally, a Fisher vector is used to encode the feature descriptor and classify it with the SVM. Experimental results show that the proposed method can improve the effectiveness and comprehensiveness of the expression of behavior features in traditional dense trajectories in a certain range, and the results of behavior recognition are also better than those of traditional dense trajectory methods and some improved methods. To sum up, we have to admit that some problems still exist during this foreground-based sampling. This strategy may lead to tapping in local minima point drifting, which we will focus on in future work.

Author Contributions: S.D. and D.H. conceived and designed the experiments; R.L. performed the experiments and analyzed the data; and S.D. and M.G. contributed to the content and writing of this manuscript.

Acknowledgments: This work was supported by the Key Scientific Research Projects of Higher Education Institutions in Henan Province, China (16A510032).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
2. Pao, H.K.; Fadlil, J.; Lin, H.Y. Trajectory analysis for user verification and recognition. *Knowl. Based Syst.* **2012**, *34*, 81–90. [CrossRef]
3. Cancela, B.; Ortega, M.; Penedo, M.G. On the use of a minimal path approach for target trajectory analysis. *Pattern Recognit.* **2013**, *46*, 2015–2027. [CrossRef]
4. Wang, H.; Yun, Y.; Wu, J. Human action recognition with trajectory based covariance descriptor in unconstrained videos. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015.
5. Weng, Z.; Guan, Y. Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor. *EURASIP J. Image Video Process.* **2018**, *8*. [CrossRef]
6. Yang, Y.; Pan, H.; Deng, X. Human action recognition with salient trajectories and multiple kernel learning. *Signal Process.* **2013**, *93*, 2932–2941.
7. Phyo, C.N.; Zin, T.T.; Tin, P. Complex human—Object interactions analyzer using a DCNN and SVM hybrid approach. *Appl. Sci.* **2019**, *9*, 1869. [CrossRef]
8. Lu, X.; Yao, H.; Zhao, S. Action recognition with multi-scale trajectory-pooled 3D convolutional descriptors. *Multimed. Tools Appl.* **2019**, *78*, 507–523. [CrossRef]
9. Seo, J.J.; Kim, H.I.; Neve, W.D. Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection. *Image Vis. Comput.* **2017**, *58*, 76–85. [CrossRef]
10. Zhang, Y.; Tong, X.; Yang, T. Multi-model estimation based moving object detection for aerial video. *Sensors* **2015**, *15*, 8214–8231. [CrossRef] [PubMed]
11. Yun, Y.; Wang, H. Motion keypoint trajectory and covariance descriptor for human action recognition. *Vis. Comput.* **2017**, *34*, 1–13.
12. Matikainen, P.; Hebert, M.; Sukthankar, R. Trajectons: Action recognition through the motion analysis of tracked features. In Proceedings of the IEEE 12th International Conference on Computer Vision Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 514–521.
13. Yu, J.; Jeon, M.; Pedrycz, W. Weighted feature trajectories and concatenated bag-of-features for action recognition. *Neurocomputing* **2014**, *131*, 200–207. [CrossRef]
14. Zheng, H.M.; Li, Z.; Fu, Y. Efficient human action recognition by luminance field trajectory and geometry information. *Transplant. Proc.* **2009**, *42*, 987–989.
15. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.

16. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
17. Vig, E.; Dorr, M.; Cox, D. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In Proceedings of the European Conference on Computer Vision, Firenze, Italy, 7–13 October 2012; pp. 84–97.
18. Hiriart-Urruty, J.B.; Strodiot, J.J.; Nguyen, V.H. Generalized Hessian matrix and second-order optimality conditions for problems with C 1,1 data. *Appl. Math. Optim.* **1984**, *11*, 43–56. [CrossRef]
19. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 24–26 June 2008; pp. 1–8.
20. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 20–25.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).