Journal of Korea Multimedia Society Vol. 21, No. 4, April 2018(pp. 478-484)
https://doi.org/10.9717/kmms.2018.21.4.478

Improved DT Algorithm Based Human Action Features Detection

Zeyuan Hu[†], Suk-Hwan Lee[††], Eung-Joo Lee[†††]

ABSTRACT

The choice of the motion features influences the result of the human action recognition method directly. Many factors often influence the single feature differently, such as appearance of the human body, environment and video camera. So the accuracy of action recognition is restricted. On the bases of studying the representation and recognition of human actions, and giving fully consideration to the advantages and disadvantages of different features, the Dense Trajectories(DT) algorithm is a very classic algorithm in the field of behavior recognition feature extraction, but there are some defects in the use of optical flow images. In this paper, we will use the improved Dense Trajectories(iDT) algorithm to optimize and extract the optical flow features in the movement of human action, then we will combined with Support Vector Machine methods to identify human behavior, and use the image in the KTH database for training and testing..

Key words: Features Detection, Dense Trajectories, Improved Dense Trajectories, Action Recognition

1. INTRODUCTION

The main way for human understanding the world and receiving information is vision, it is not only pointed the perception of the external source, but also pointed a plurality of process such as information obtained, processed and understood, there is research demonstrates that there is 80% information obtained by human brain on the basis of vision, therefore, as the most important way for transmitting information among human, vision acts an important role in the life of human.

Human action recognition [1,2] in videos attracts increasing research interests in computer vision community due to its potential applications in video surveillance, human computer interaction, and video content analysis. However, action recognition remains as a difficult problem when focusing on realistic datasets collected from movies [3], web videos [4,5],Virtual Reality(VR) [6], and TV shows [7]. In recent years, with the development and widely popularization of film, internet and so on, vision has become a more important tool for obtaining information, and the need for automatically gathering and recognizing the information in the vision has become more and more.

With the rapid development and application of information technology and the popularity of the application, the use of computer vision technology in image processing and pattern recognition in the field, and the video image of human motion feature extraction and effective identification has become a hot topic of concern. Computer vision technology to identify the human body movement video or im-

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-1143, E-mail : ejlee@tu.ac.kr

Receipt date : Mar. 8, 2018, Revision date : Apr. 2, 2018
Approval date : Apr. 10, 2018

[†] Dept. of Information Communication Engineering, Tongmyong University (E-mail : dlhzy410@126.com)

[††] Dept. of Information Security Engineering, Tongmyong University (E-mail : skylee@tu.ac.kr)

[†††] Dept. of Information Communication Engineering, Tongmyong University

※ This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A3B03931003).

age is based on its video or image sequence analysis and processing; the detection of human motion targets for motor feature extraction and classification of identification, so as to achieve the purpose of understanding and describing their behavior.

It is mainly involved in a series of research fields such as image processing, multi-sensor technology, virtual reality, pattern recognition[8], computer vision and graphics, computer aided design, visualization technology and intelligent robot in the field of scientific research. For the human body motion image sequence analysis and processing of the human body visual analysis technology, under normal circumstances can be divided into the following process, moving target detection, moving target feature extraction and identification of complex background sport target identity. The detection, identification and tracking of the important parts of the human body (head, hand, etc.) are the basis for understanding human behavior. On the basis of solving these basic problems, more important and more difficult problems are action recognition and behavior understanding. Human body movement characteristics include: limb swing characteristics [9], gait characteristics [10], human body contour projection features [11], human symmetry features.

At present, motion characteristics in motion feature recognition include two components: structured component and dynamic component. Which is the structure of the static component of the static component, which is responsible for recording the movement of the body height, pace and other body shape information; While the dynamic component is the image of the movement in the process of showing the body's arm swing, limb tilt, legs and other sports characteristics, according to the above two types of components. Existing motion feature recognition algorithms are broadly divided into two categories: statistical-based and model-based approaches.

On the basis of taking into account the advantages and disadvantages of the different features and the scope of application, a new hybrid feature combining the static characteristics of the global shape description and the dynamic characteristics of the local optical flow description is proposed. Firstly, we use the background subtraction method to determine the approximate area of motion, get the silhouette of the human body, and use the silhouette contour vector to express the overall information of the human body appearance; And then extract the optical flow in the motion area and use the local optical flow information of the sub-region to represent the local characteristics of the human body movement, so as to improve the anti - noise ability of the optical flow; Finally, the global silhouette feature is combined with the local optical flow feature as a mixed feature. The experimental results show that the robustness and recognition performance of the mixed feature is better than that of the single feature.

In this paper, we will introduce the feature extraction method of iDT for firstly, and the second, we will use iDT feature to construct the classification method by SVM(Support Vector Machine), the last, we will use the KTH database to train and test the network.

2. DENSE TRAJECTORIES METHOD

The DT method divides the feature points separately on multiple scales of the image by way of meshing. Sampling on multiple spatial scales can ensure that the sampled feature points cover all spatial positions and scales, and that usually eight spatial scales are already sufficient, and if the image is large, it can be increased appropriately. Subsequent feature extraction is performed on each scale separately. The interval at which the feature points are sampled on W (the size of the grid) usually takes $W = 5$.

The next step is to track these feature points in the time series, but tracking the feature points in an area that lacks a change (such as a point in the middle of a white wall) cannot be achieved. So before the tracking, we remove some features of the previous point. The method here is to calculate the eigenvalues of the autocorrelation matrix for each pixel and set the threshold to remove the feature points below the threshold. The threshold is determined by:

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

Where $(\lambda_i^1, \lambda_i^2)$ is the eigenvalue of the pixel i in the image I. 0.001 for the experiment to determine a more appropriate value.

Let the coordinates of a feature point sampled in the step is $P_t = (x_t, y_t)$, then we can use the following formula to calculate the position of the feature point in the next frame image.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M \times w_t)|_{x_t y_t} \quad (2)$$

Where $w_t = (u_t, v_t)$ is the dense light flow field, calculated from I_t and I_{t+1}, u and v represent the horizontal and vertical components of the optical flow respectively. And M represents the median filter, the size of 3 * 3. Therefore, the formula is obtained by calculating the direction of the motion of the feature points in the middle of the optical flow in the neighborhood of the feature point.

The position of a feature point on a continuous L-frame image constitutes a trajectory, and subsequent feature extraction is performed along each track $(P_t, P_{t+1}, \ldots, P_{t+L})$. As the tracking of feature points there is drift phenomenon, so long tracking is unreliable, so every L-frame to re-dense sampling of a feature point, re-tracking. In the DT/iDT algorithm, select L = 15.

In addition, the trajectory itself may also constitute a trajectory shape feature descriptor. For a trajectory of length L, its shape can be described by $(\triangle P_t, \triangle P_{t+1}, \ldots, \triangle P_{t+L-1})$, where the displacement vector $\triangle P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$ is used. After the regularization, the trajectory feature descriptor can be obtained. Regularization is:

$$T = \frac{(\triangle P_t, \triangle P_{t+1}, \ldots, \triangle P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\triangle P_j\|} \quad (3)$$

So the final trajectory is characterized by a 15 * 2 = 30-dimensional vector.

In addition to the trajectory shape features, we also need more powerful features to describe the optical flow, DT/iDT used HOF (Histogram of Flow), HOG (Histogram of Oriented Gradient) and MBH (Motion Boundary Histograms) three characteristics.

Along the trajectory of length L of a feature point, the area of N×N around the feature point is taken on each frame, and a time-space body is formed. In this time-space body, in a grid division, the space is divided into n_σ in each direction, the time is evenly selected n_γ copies. So there $n_\sigma \times n_\sigma \times n_\gamma$ regions are used as a feature extraction in the time-space body. In DT/iDT, take N=32, $n_\sigma = 2$, $n_\gamma = 3$, then the details of the extraction of each feature are introduced.

HOG Features: The HOG feature calculates the histogram of the gray scale image gradient. The big number of the histogram is taken as 8. So the length of the HOG feature is 96 (2 * 2 * 3 * 8).

HOF features: HOF calculates the histogram of the optical flow (including the direction and amplitude information). The big number of the histogram is taken as 8 + 1. The first eight bits are the same as HOG, and the extra bit is used to count the pixels whose optical flow amplitude is less than a certain threshold. Therefore, the characteristic length of HOF is 108 (2 * 2 * 3 * 9).

MBH Features: MBH calculates the histogram of the gradient of the optical flow image and can also be understood as the HOG feature calculated on the optical flow image. Since the optical flow image includes the x direction and the y direction, MBHx and MBHy are calculated, respectively. MBH has a total feature length of 192 (2 * 96).

After the calculation, the normalization of the feature is also needed, and the HOG, HOF and

MBH are normalized in the DT algorithm.

For a video, there are a large number of trajectories, each of which corresponds to a set of traits (trajectory, HOG, HOF, MBH), so it is necessary to encode these feature sets to obtain a fixed-length coding feature for the final video classification.

3. IMPROVED DENSE TRAJECTORIES METHOD

The basic framework of the iDT algorithm is the same as that of the DT algorithm. The main improvement lies in the optimization of the optical flow image, the improvement of the characteristic regularization method and the improvement of the characteristic coding method.

In order to estimate the projection transformation accurately, the iDT algorithm uses two methods to obtain the matching point pair. Respectively, SURF features and optical flow characteristics. After obtaining the matching point, you can use the RANSAC algorithm to estimate the projection transformation matrix. The specific operation is: The gray scale images at time t and time t+1 are I_t and I_{t+1}, respectively, and the projection transformation matrix $H(I_{t+1} = H \times I_t)$ is calculated from two images. And then transform (warp) with the inverse of H. That is:

$$I_{t+1}^{warp} = H^{-1} \times I_{t+1} \tag{4}$$

I_{t+1}^{warp} represents the assumption that there is no image at time t+1 when the camera moves. I_t and I_{t+1}^{warp} can be used to calculate the optimized optical flow.

Eliminating the effects of camera movement from optical flow has two main advantages:

The motion descriptor (mainly HOF and MBH) can describe the action more accurately, and the accuracy of classification with a single descriptor is much higher than that in DT.

Since the trajectory is also performed using the optical flow, it is possible to eliminate the trajectory of the displacement vector in the optimized optical flow which is smaller than the threshold value by setting the threshold value.

In the iDT algorithm, the HOF, HOG, and MBH features are taken in a different way than the DT algorithm (L2 normalization) - the regularization of L1 is then squared to each dimension of the feature.

Feature coding phase The iDT algorithm no longer uses the Bag of Features method, but uses a better Fisher Vector code. The parameters of the Fisher Vector used in iDT are:

The length of the feature used for training: trajectory + HOF + HOG + MBH = 30 + 96 + 108 + 192 = 426.

Number of features used for training: 256,000 random samples from the training set.

PCA dimensionality reduction ratio: 2, ie dimension divided by 2, dimensionality after dimensionality is 213. First downs dimension, after encoding.

The number of Gaussian clusters in Fisher Vector K: K = 256.

Fig. 2 is in the MATLAB simulation gives the HOG features of video capture image, is a gray scale image gradient.

The MBH features is calculates the histogram of the gradient of the optical flow image and can also be understood as the HOG frature calculated on the optical flow image. The optical flow image includes the X direction and the Y direction ,so the

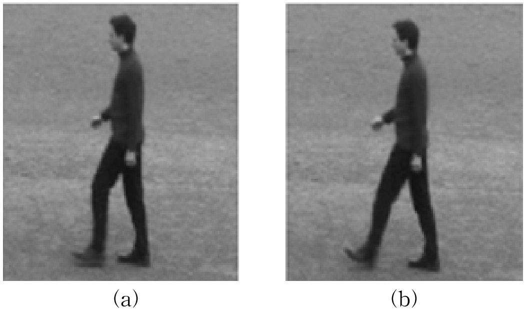

(a) (b)

Fig. 1. In experiment, (a) and (b) is a video capture the moment of T and T+1 point of two images.

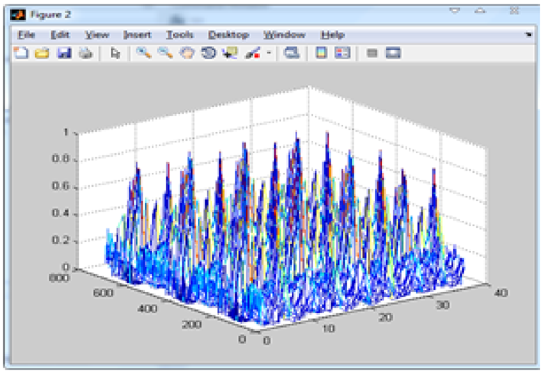


Fig. 2. HOG features.

images (a) and (b) are the histogram of the gradient in X direction, called MBHx. The images (c) (d) are the histogram of the gradient in Y direction, called MBHy.

So the number of bits after encoding is 2KD, that is 109056 dimensions. After encoding, SVM is used for classification. Use one-against-rest strategy to train multiple classifiers.

4. EXPERIMENT and RESULTS

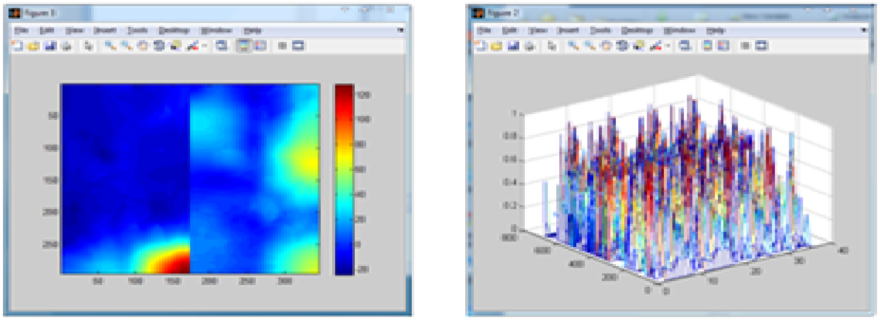
In order to verify the effectiveness of this algorithm, a large number of comparative experiments were made on the published KTH database.

This experiment is run in MATLAB2010b implementation. KTH database has six kinds of actions, respectively, boxing, hand-clapping, hand-waving, jogging, running, walking, each action by 25 different people in four scenes to complete a total of 599 video; Background is relatively static, in addition to the lens closer / pull away, the camera's movement is relatively slight, as shown in Fig. 5.

The experiment result showed in Table 1, and we do some comparison with some other method, such as Ahmad method include Optical Flow +Shape Flow, Sawant method include Silhouette + Local Optical Flow, and Original DT method. The result shows that the iDT algorithm has an excellent recognized rate 91.20%, as the best behavior recognition algorithm before the depth of learning, has excellent results and good robustness.

5. CONCLUSION

Human action recognition is an important research direction in the field of computer vision, In

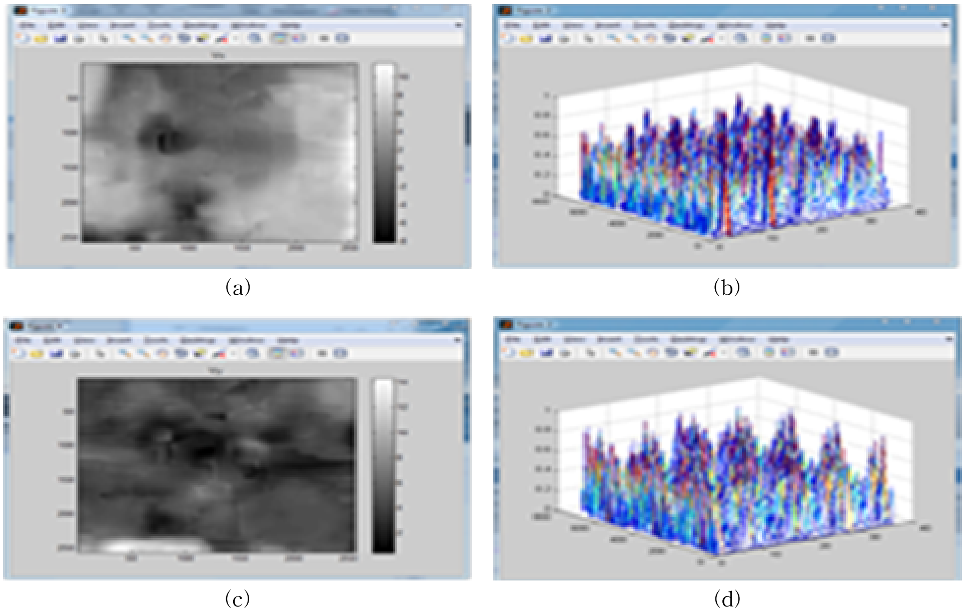


(a) (b)

Fig. 3. The image (a) is the histogram of the optical flow. The image (b) is the characteristics of HOF features. (Including the direction and amplitude information).

Table 1. The different features combine the corresponding recognition rates

Method	Feature	Rate
Ahmad [12]	Optical Flow+Shape Flow	88.29%
Sawant [13]	Silhouette+Local Optical Flow	90.80%
Original Method	DT(HOG, HOF, MBH)	84.50%
Improved Method	iDT(HOG, HOF, MBH)	91.20%



(a) (b) (c) (d)

Fig. 4. MBH features.

the field of pattern recognition and artificial intelligence also has important theoretical and practical significance. Gesture recognition technology is widely used in video surveillance, motion analysis, virtual reality, auxiliary medical, human-machine intelligent interaction, etc, the research has the very high application prospects and market value. However, due to the fact that the hard to avoid complex environmental factors, including the complex background, the action of the internal factors and external factors, shade, and illumination changes, etc.

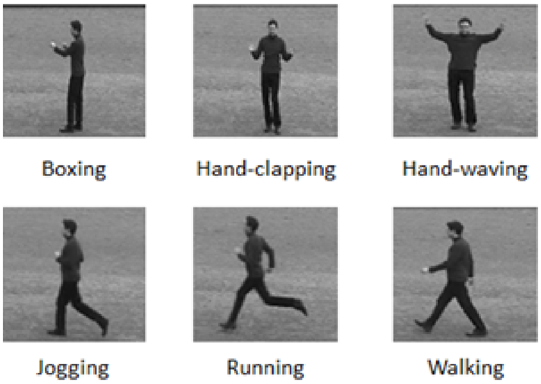


Fig. 5. Human action images of KTH database.

This paper during the training step we find that there are some factors (such as motion blur, the camera motion estimation is inaccurate when the figure is high), it can affect the results of activity recognition, but in vast majority of people and situation, this method can quickly and correctly train, and get a expect result.

REFERENCE

[1] J.K. Aggarwal, and M.S. Ryoo, "Human Activity Analysis: A Review," *Journal of Association for Computing Machinery(ACM) Computing Surveys*, Vol. 43, No. 3, pp. 102-145, 2011.

[2] H. Wang, and C. Schmid, "Action Recognition with Improved Trajectories," *Proceeding of IEEE International Conference on Computer Vision*, pp. 3551-3558, 2013.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio,

and T. Serre, “HMDB: A Large Video Database for Human Motion Recognition,” *Proceeding of International Conference on Computer Vision*, pp. 2556-2563, 2011.

[5] K. Soomro, A.R. Zamir, and M. Shah, *UFC101: A Dataset of 101 Human Actions Classes From Videos in The Wild*, Master's Thesis of Cornell University, 2012.

[6] K.G. Cho, "A Human Action Recognition Scheme in Temporal Spatial Data for Intelligent Web Browser," *Journal of Korea Multimedia Society*, Vol. 8, No. 6, pp. 844-855, 2005.

[7] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, “Structured Learning of Human Interactions in TV Shows,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2441-2453, 2012.

[8] L.B. Truong, S.H. Kim, and G.M. Jeong, “Pattern Recognition with feature feedback: Feature Mask By Using PCA," *Proceeding of Conference on Korea Multimedia Society*, pp. 14-16, 2010.

[9] Z. Guangyu, X. Changsheng, H. Qingming, and G. Wen, “Action Recognition in Broadcast Tennis Video,” *Proceeding of the 18th International IEEE Conference on Pattern Recognition*, pp. 251-254, 2006.

[10] A. Gritai, Y. Sheikh, and M. Shah, “On the Use of Anthropometry in the Invariant Analysis of Human Actions," *Proceeding of the 17th International IEEE Conference on Pattern Recognition*, pp. 923-926, 2004.

[11] W. Liang, and D. Suter, “Recognizing Human Activities from Silhouette: Motion Subspace and Factorial Discriminative Graphical Model,” *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[12] M. Ahmad, and S.W. Lee, "Human Action Recognition Using Shape and CLG-motion Flow from Multi-view Image Sequences,” *Journal of Pattern Recognition*, Vol. 41, No. 7, pp. 2237-2252, 2008.

[13] N. Sawant, and K.K. Biswa, “Human Action Recognition Based on Spatio-temporal Feature," *Proceeding of International Conference on Pattern Recognition and Machine Intelligence*, pp. 357-362, 2009.

Zeyuan Hu

He received his B. S. at Qingdao Institute of Technology in China (2011-2015). Currently, he is studying in the Department of Information and Communications Engineering in Tongmyong University, Korea for his Masters degree. His main research areas are image processing and pattern recognition.

Suk-Hwan Lee

He received his B. S., M. S. and Ph. D. in Electrical Engineering from Kyungpook National University. Korea in 1999, 2001, and 2004 respectively. He is cur rently an associate professor in Department of Information Security at Tongmyong University. His research interests include multimedia security, digital image processing, and computer graphics.

Eung-Joo Lee

He received his B. S., M. S. and Ph. D. in Electronic Engineering from Kyungpook National University. Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997, he has been with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a Professor. From 2000 to July 2002, he was a president of Digital Net Bank Inc. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.