An Improved Method for Interest Point Detection in Human Activity Video

Zhe Hu, Hong Zhang*, Yang Yang, Chujun Yang
Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China
zhangh@swu.edu.cn

Abstract. An improved method for human activity recognition in RGB-D video using improved color-depth local spatio-temporal features (CoDe4D) detection method is presented in this paper. Firstly, histogram equalization and correction function are employed to suppress noise of rgb frame and depth frame, respectively. Then a saliency map is constructed by the improved CoDe4D method. In this way, feature points can be obtained by the saliency map, which integrates both depth information and RGB information. Next, the depth cuboid similarity feature (DCSF) is utilized to describe feature vectors. Meanwhile, visual words are generated by bag of feature method. To further improved the estimation accuracy, the SVM with generalized histogram intersection kernel is applied to train and predict categories. It shows good performance on MSR Daily Activity 3D datasets.

1. Introduction
Human Activity recognition performance has been improved due to the rapid development of local spatio-temporal (LST) features, but it is limited to the sensing device. However, with the advent of the depth sensor, the depth information is now available. Previous methods did not satisfy the new situation. The new situation has become an emerging problem with increasing information from Human Activity videos. However, how to deal with the increasing information is a challenging problem with several solution from traditional Human Activity recognition approaches. The increasing information from depth sensor 1) provides 3-D structure information; 2) is generally not sensitive to illumination changes; 3) and is cheaper because of the emergence of Microsoft Kinect, Asus Xtion Pro Live, and PrimeSense color-depth sensors. In darkness, the color sensor is significantly influenced by the illumination changes. Human shapes, poses and textures are hardly detected in such situation. Yet the depth sensor can also work in darkness, and is much easier to build 3-D structure of the scene. Therefore, color information and depth information are combined to detect local spatio-temporal feature point. In darkness and some other situations, depth information is vital clues for detecting LST feature points. For example, when a person shows in video with background color clothes, it’s hard to tell apart human shapes, poses and textures from background. From this perspective, it is important to develop a human activity recognition method that can work for above situation as well as general situation. Excitingly, many human activity recognition algorithms are available to detect LST features with color information. Therefore, both color and depth information can be effectively used to accomplish human activity recognition.

Usually, local appearance and motion information around each interest point are used to described LST features. LST features are widely used to encode human actions in color videos [1], [2]-[7] because of its relatively invariant to image scaling, translation and rotation, their partially invariant to
illumination changes, and its robust to partial occlusion \[8\], \[9\]. Besides, extracting LST features don’t require preprocessing steps like human detection and tracking, which brings about potential failures.

To date, only few works on human activity recognition for incorporating color and depth information have been carried out \[1\], \[10\]-\[12\]. Zhongwei Cheng et al propose Comparative Coding Descriptor (CCD) to represent the depth information. By fusing color and depth features, their method achieves good performance \[10\]. Hao Zhang and Lynne E. Parker propose a new 4-dimensional (4D) local spatio-temporal feature that combines both color and depth information \[1\]. An-An Liu et al propose a multi-domain and multi-task learning (MDMTL) method, which is the first supervised framework to jointly realize domain-invariant feature learning and task modeling for multi-domain action recognition \[1\]. Hao Zhang and Lynne E. Parker \[12\] fuses both color and depth information acquired from RGB-D cameras, proposing the new CoDe4D LST feature \[12\].

Most of them focus on human in the center of the video, ignoring the other human not in the center of the video. The improved CoDe4D method can detect the other human. The overview of the paper is posted in Fig.1. The remainder of the paper is organized as follows. We briefly introduce related work in Section 2. Section 3 details the improved CoDe4D method. At last, report the experimental results on MSR Daily Activity 3-D dataset in Section 4. Section 5 concludes the paper.

2. Related Work

The related works can be roughly grouped into two categories: LST interest point detection and LST interest point description. The former focuses on fusing color and depth information and extracting LST interest point. The latter focuses on describe the LST interest point knowledge calculated from feature detection for human activity recognition.

2.1. Feature Detection

Researchers have developed various methods for achieving view-variant action recognition. Laptev \[3\] extended the Harris corner to 3D Harris, one of the spatio-temporal interest point (STIP) families. The pixel values of the neighborhood of these spatio-temporal feature points have significant changes in time and space. In this algorithm, the size of the neighborhood block can adapt to the time dimension and the space dimension. Dollar \[13\] pointed out that there is disadvantage in the above method, that is, the number of stable points of interest detected is too small, so Dollar alone uses the Gabor filter to filter in the time dimension and space dimension, so that the number of points of interest is detected. It will change as the local neighborhood size of time and space changes. Similarly, Rapantzikos \[14\] applies discrete wavelet transforms in three dimensions, selecting spatio-temporal salient points through the low-pass and high-pass filter responses of each dimension. At the same time, in order to integrate color and motion information, Rapantzikos \[15\] added color and motion information to calculate its salient points. Unlike the idea of detecting points of interest throughout the
human body, Wong [16] first detects points of interest in the subspace associated with the motion, which correspond to a portion of the motion, such as the arm swing, in these subspaces, some Sparse points of interest are detected. In a similar approach, Bregonzio [17] first estimates the focus of visual attention by calculating the difference in subsequent frames, and then uses Gabor filtering to detect significant points in these regions.

Lu Xia and J.K. Aggarwal present a filtering method to extract STIPs from depth videos (called DSTIP) that effectively suppress the noisy measurements. [18]. Hao Zhang and Lynne E. Parker fuses both color and depth information acquired from RGB-D cameras, proposing CoDe4D method [12].

He construction and geometrical dimensions of the composite slabs are shown in Fig.1. It is noted that the slab specimens consist of the thin-walled steel skeleton which also is formed with H-type main steel beams and steel channels, lightweight precast panels set upon the steel skeleton, shear keys connected to the main steel beams and post-pouring concrete layer.

2.2. Feature Detection
A local feature description is a description of a block in an image or video. The descriptor should be insensitive to background clutter, scale and direction changes. The spatial and temporal size of an image block typically depends on the size of the detected point of interest. Feature blocks can also be described by meshes based on local features, since a mesh includes locally observed domain pixels, treating them as a block, thus reducing the effects of local variations in time and space. The two-dimensional SURF feature [20] has been extended to three dimensions by Willems [21], and each cell of these eSURF features contains all Harr-wavelet features. Laptev [19] uses local HOG (gradient histogram) and HOF (light flow histogram). Klaser [22] extended the HOG feature to 3D, which formed 3D-HOG. Each bin of 3D-HOG is composed of regular polyhedrons, and 3D-HOG allows fast density sampling of cuboids at multiple scales. This work of extending the 2D feature point detection algorithm to 3D feature points is similar to extending the SIFT algorithm [23] to the 3D SIFT Scovanner [24]. In Wang [25], he compared various local description operators and found that the description operator that integrates gradient and optical flow information works best in most cases. There is also a description of the more popular, namely the word bag [26], [27], which is the word frequency histogram feature utilized.

Lu Xia and J.K. Aggarwal build a novel DCSF to describe the local 3D depth cuboid around the DSTIPs with an adaptable supporting size [18]. Hao Zhang and Lynne E. Parker’s multichannel orientation histogram descriptor applies a 4-D support region, which is adaptive to linear perspective view changes, on each interest point [12].

Though the research nowadays has achieved great performance for the action recognition, there are also existing noticeable disadvantages. In consequence, it is hard to extract discriminative and typical message by the model with only RGB information or depth information. Therefore, we propose an improved method to recognize complex spatio-temporal dynamic in activities based on current methods.

3. The Proposed Framework

3.1. Overview of Our Framework
The flowchart of our framework is illustrated in Fig.1. At first, an improved CoDe4D method is developed to extract local features from the video. Then, Depth Cuboid Similarity Feature (DCSF) [18] are proposed to describe spatio-temporal interest points of video by calculate the similarity among histograms in cuboids. Afterwards, the obtained representation of videos classified by bag of words to constructing visual words and generating feature vector. Ultimately, the feature vector is classified by SVM with generalized histogram intersection kernel to recognizing.

In this section, the improved CoDe4D is firstly introduced. Then, DCSF descriptor is applied to describe the interest points to constructing feature vector. And feature vector is generated by utilizing the bag of feature method. In addition, SVM is employed to train and predict these videos. Lastly, the
accuracy is discussed between the improved CoDe4D approach and several recently proposed human activity recognition methods based on the same dataset. To formalize the problem, a list of primary symbols used in the method are first given in Tab.1.

| Symbol | Description |
|--------|-------------|
| $x$    | X coordinates of pixels in each frame of videos |
| $y$    | Y coordinates of pixels in each frame of videos |
| $t$    | Each frame of the video |
| $Q(x, y, t)$ | The gray value computed from RGB values |
| $S(x, y, t)$ | Depth of pixels in each frame of videos |
| $Q_s(x, y, t)$ | Smoothed gray values |
| $S_s(x, y, t)$ | Smoothed depth values |
| $F(x, y | \sigma)$ | Gaussian filter |
| $\sigma$ | Control the spatial scale of the Gaussian filter |
| $Q_s(x, y, t)$ | Combined variations of gray values |
| $S_s(x, y, t)$ | Combined variations of depth values |
| $B(t | \tau, \omega)$ | A complex-valued Gabor filter |
| $\omega$ | Sinusoidal wavelength |
| $\tau$ | Control the temporal scale of Gabor filter |
| $Sup(x, y, t_0 | \tau)$ | Correction function to reduce noise caused by depth sensor |
| $n(x, y)$ | The number of flips during time $[t_0 - \tau, t_0 + \tau]$ at $(x, y)$ |
| $\Delta t_i(x, y)$ | The duration of the $i$-th flip |
| $Sup_0$ | The threshold of correction function |
| $Sup$ | The final value of correction function |
| $P(x, y, t)$ | A spatio-temporal saliency map |
| $\mu$ | A mixture weight to balance between gray and depth information |
| $Sim(A, B)$ | the similarity between histogram $z_A$ and histogram $z_B$ |
| $K(l, h)$ | The generalized histogram intersection kernel function |

### 3.2. Improved CoDe4D Feature Detection

Most of the color images now use the RGB color mode. When processing the image and video, the three components of RGB are processed separately. In fact, RGB does not reflect the morphological features of the image, but the color is configured from the optical principle. In order to reduce the amount of raw data of the RGB video, to reduce the amount of calculation for subsequent processing and to extract morphological features in images and videos, the RGB values of images and videos are converted to grayscale values. To extract LST interest points that represent motion features effectively, histogram equalization is used to enhance the image and video frame.

In general, there are two key stages in a practical data set based interest point detection: Gaussian filter and 2-D Gabor filter. The Gaussian filter links the image frequency domain processing with the time domain processing. As a low-pass filter, it can filter low-frequency energy (such as noise) for image smoothing. First of all, we use a 2D Gaussian filter to smooth video dimension

$$Q_s(x, y, t) = Q(x, y, t) * F(x, y | \sigma)$$  \hspace{1cm} (1)

$$S_s(x, y, t) = S(x, y, t) * F(x, y | \sigma)$$  \hspace{1cm} (2)

Where the * denotes convolution. The 2-D Gaussian filter $F(x, y | \sigma)$ is applied to smoothing the image of each frame of the video.

$$F(x, y | \sigma) = \frac{1}{2\pi \sigma^2} e^{-\frac{|x^2+y^2|}{2\sigma^2}}$$  \hspace{1cm} (3)

After Gaussian filtering, the next step is to use Gabor filtering to detect the edge of the human body.
\[ Q_{st}(x, y, t) = Q_s(x, y, t) \ast B(t \mid \tau, \omega) \]  \hspace{1cm} (4)

\[ S_{st}(x, y, t) = S_s(x, y, t) \ast B(t \mid \tau, \omega) \]  \hspace{1cm} (5)

The Gabor filter is very similar to the visual stimuli response of simple cells in the human visual system. It has good characteristics in extracting the local space and frequency domain information of the target. Gabor filter are sensitive to the edges of the image provide good direction selection and scale selection characteristics, and are insensitive to illumination changes, providing good adaptability to illumination changes.

The Gabor filter \( B(t \mid \tau, \omega) \) is as follows.

\[ B(t \mid \tau, \omega) = e^{-\frac{t^2}{\tau^2}} \cdot e^{i(2\pi\omega t)} \]  \hspace{1cm} (6)

we set \( \omega = 0.6/\tau \).

In depth video, there is a noise that cannot be removed by smoothing filter. This kind of noise appears at the edge of the object, and the depth value bounces back and forth between the depth of the background and the depth of the object. Its signal flips much faster than human activity and occurs frequently in whole video. A correction function is as follows to reduce such noise.

\[ Sup(x, y, t_0 \mid \tau) = \sum \Delta \delta(x, y) \cdot \delta(x, y) \]  \hspace{1cm} (7)

To this end, the average time of each signal flip \( Sup(x, y, t_0 \mid \tau) \) which is used as a correction function is calculated. When the real motion occurs, the value of this correction function is higher, and a threshold is selected so that the correction function only affects the noise and doesn’t change the value of real motion.

\[ \overline{Sup} = \begin{cases} Sup, \text{if } Sup > Sup_0 \\ Sup_0, \text{else} \end{cases} \]  \hspace{1cm} (8)

Then grayscale and depth edge information are added to extract local maximum points as the feature points. Also, the RGB value need to be processed as the same of the depth value.

\[ P(x, y, t) = (1 - \mu) \cdot |Q_{st}(x, y, t)\ast\overline{Sup}|^2 + \mu \cdot |S_{st}(x, y, t)\ast\overline{Sup}|^2 \]  \hspace{1cm} (9)

3.3. DCSF Feature Description and SVM Recognition

Activate by Lu Xia’s and J.K. Aggarwal’s DCSF descriptor [18], a cuboid is extracted around the feature points. The cuboid size is adaptable to the depth. Dividing the cuboid pixels into plenty blocks, a histogram \( z_A \) which is counted from pixel values of block \( A \) is computed. And then Bhattacharyya distance is used to define the similarity between histogram \( z_A \) and histogram \( z_B \). The feature vector is composed of the similarity scores among the blocks of each cuboid.

\[ Sim(A, B) = \sum_{i=1}^{m} \sum_{l=1}^{n} \left( z_A^{(i)} \cdot z_B^{(i)} \right) \]  \hspace{1cm} (10)

The bag of words aims to turn feature vector into visual words, and reduces dimensionality. First, k-means algorithm with Euclidean distance is employed to clustering feature vector to build visual words. Then SVM with generalized histogram intersection kernel [28] is applied for classification.

\[ K(l, h) = \sum_{j=1}^{n} \min(|l_j|, |h_j|) \]  \hspace{1cm} (11)

Assign \( a = b = 1/4 \).

4. Experiment

The proposed method is evaluated on the MSR Daily Activity 3D, which has 16 kinds of daily human-object interaction actions, such as drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, sit down. Each action contains 20 videos, and is performed twice in sitting and in standing by 10 subjects. The size of RGB frames is 640×480, and the size of depth frame is 320×240.

As show in Fig.2, there are some red spots on the person in the background while parameters are set in proper values. But the red spots are too less to bring enough message, person with most red spots should be focused on. With regards to this, the relevant parameters are modified so that the red spots are concentrated on one subject.
Thus, before filtering, rgb frame and depth frame are resized into the same size. Some videos which subject is barely not moved was deleted. Then the parameters of the feature detector are set to $\sigma = 5$, $\tau = 3$, $\mu = 0.9$, and $Sup_0 = 4$. 350, 500, 800, 1000 and 1300 feature points are extracted. DCSF [13] descriptor is applied for generating feature vector, which is clustered into 100 clusters using K-means. Each center of clusters denotes a visual word of the video.

As shown in Fig.3, the red spots are 350 extracted interest points, which show trajectory of human activity.

As shown in Fig.3, the red spots are 350 extracted interest points, which show trajectory of human activity.

Under these settings, the extracted feature points are as fig.1. We get the accuracy of 94.1176% (Tab.2).

| Features                                      | Accuracy |
|-----------------------------------------------|----------|
| Local occupancy pattern (LoP) features [29]   | 42.5%    |
| Joint position features [29]                  | 68.0%    |
| Harries3D + Depth – HOG/HOF [19]              | 79.1%    |
| Depth cuboid detector and descriptor [2]      | 83.6%    |
| Depth cuboid similarity features [18]         | 85.8%    |
| CoDe4D + Adaptive MCOH [12]                  | 86.0%    |
| MDMTL [11]                                    | 93.8%    |
| Our improved CoDe4D+DCSF+SVM                  | 94.1176% |

5. Conclusion
This paper aims to solve the problem of the fusion of color information and depth information by using the improved CoDe4D method, and aims to focus on the subject not only in the center of the video but also in the background. In particular, the improved CoDe4D extracted several points from subject in the background. Besides, the improved method has good performance on the public action datasets MSR Daily Activity 3D by applying SVM with generalized histogram intersection kernel. Furthermore, categories with similar movements are difficult to distinguish. The question will be explored in the future.

Acknowledgments
This work was supported by "Southwest University Research Fund" No. 2018291059.
References

[1] H. Zhang and L. E. Parker, “4-dimensional local spatio-temporal features for human activity recognition,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Sep. 2011, pp. 2044–2049.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in Proc. IEEE Int. Workshop Vis. Surf. Perform. Eval. Tracking Surf. Oct. 2005, pp. 65–72.

[3] I. Laptev, “On space-time interest points,” Int. J. Comput. Vis., vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.

[4] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, “Simplex-based 3D spatio-temporal feature description for action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2014, pp. 2059–2066.

[5] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. González, “Selective spatio-temporal interest points,” Comput. Vis. Image Understand., vol. 116, no. 3, pp. 396–410, Mar. 2012.

[6] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” Int. J. Comput. Vis., vol. 103, no. 1, pp. 60–79, May 2013.

[7] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3551–3558.

[8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[9] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in Proc. Brit. Mach. Vis. Conf., Sep. 2009, pp. 127–137.

[10] Chang Z, Qin L, Ye Y, et al. Human Daily Action Analysis with Multi-view and Color-Depth Data[C]// Proceedings of the 12th international conference on Computer Vision - Volume 2. Springer-Verlag, 2012.

[11] Liu A A, Xu N, Nie W Z, et al. Multi-Domain & Multi-Task Learning for Human Action Recognition[J]. IEEE Transactions on Image Processing, 2018:1-1.

[12] Zhang H, Parker L E. CoDe4D: Color-Depth Local Spatio-Temporal Features for Human Activity Recognition from RGB-D Videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 26(3):1-1.

[13] Dollár, P., V. Rabaud, et al. (2005). Behavior recognition via sparse spatio-temporal features, IEEE.

[14] Rapantzikos, K., Y. Avrithis, et al. (2007). Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: potential in human action recognition, ACM.

[15] Rapantzikos, K., Y. Avrithis, et al. (2009). Dense saliency-based spatiotemporal feature points for action recognition, Ieee.

[16] Wong, S. F. and R. Cipolla (2007). Extracting spatiotemporal interest points using global information, Ieee.

[17] Bregonzio, M., S. Gong, et al. (2009). Recognising action as clouds of space-time interest points, IEEE.

[18] Xia L, Aggarwal J K. Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera[C]// 2013.

[19] Laptev, I., M. Marszalek, et al. (2008). Learning realistic human actions from movies, IEEE.

[20] Bay, H., T. Tuytelaars, et al. (2006). "Surf: Speeded up robust features." Computer Vision–ECCV 2006: 404-417.

[21] Willems, G., T. Tuytelaars, et al. (2008). "An efficient dense and scale-invariant spatio-temporal interest point detector." Computer Vision–ECCV 2008: 650-663.

[22] Klassen, A. and M. Marszalek (2008). "A spatio-temporal descriptor based on 3D-gradients."

[23] Mikolajczyk, K. and C. Schmid (2004). "Scale & affine invariant interest point detectors." International journal of computer vision60(1): 63-86.

[24] Scovanner, P., S. Ali, et al. (2007). A 3-dimensional sift descriptor and its application to action
recognition, ACM.

[25] Wang, H., M. M. Ullah, et al. (2009). "Evaluation of local spatio-temporal features for action recognition."

[26] Niebles, J. C., H. Wang, et al. (2008). "Unsupervised learning of human action categories using spatial-temporal words." International journal of computer vision79(3): 299-318.

[27] Schuldt, C., I. Laptev, et al. (2004). Recognizing human actions: A local SVM approach, IEEE.

[28] Boughorbel S , Tarel J P , Boujemaa N . Generalized histogram intersection kernel for image recognition[C]// IEEE International Conference on Image Processing. IEEE, 2006.

[29] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1290–1297.