JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 9, No. 1, March 2015, pp. 29-38

Improvement of Accuracy for Human Action Recognition by Histogram of Changing Points and Average Speed Descriptors

Thi Ly Vu, Trung Dung Do, Cheng-Bin Jin, Shengzhe Li, Van Huan Nguyen, Hakil Kim*, and Chongho Lee
School of Information and Communication Engineering, Inha University, Incheon, Korea
vuthily@inha.edu, {dotrungdung, sbkim, szli, conghuan}@vision.inha.ac.kr, {hikim, chlee}@inha.ac.kr

Abstract

Human action recognition has become an important research topic in computer vision area recently due to many applications in the real world, such as video surveillance, video retrieval, video analysis, and human-computer interaction. The goal of this paper is to evaluate descriptors which have recently been used in action recognition, namely Histogram of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF). This paper also proposes new descriptors to represent the change of points within each part of a human body, caused by actions named as Histogram of Changing Points (HCP) and so-called Average Speed (AS) which measures the average speed of actions. The descriptors are combined to build a strong descriptor to represent human actions by modeling the information about appearance, local motion, and changes on each part of the body, as well as motion speed. The effectiveness of these new descriptors is evaluated in the experiments on KTH and Hollywood datasets.

Category: Smart and intelligent computing

Keywords: Human action recognition; Dense sampling; HOG; HOF; HCP; AS; SVM

I. INTRODUCTION

Human actions convey essential information in videos. With the rapid growth of computer vision, intelligent camera systems, and robotics in the society nowadays, human action recognition, which aims to recognize the human actions from a series of observations in the video, has become an important research topic.

Action recognition has many challenging problems. First, the number of human actions is various, so recognizing all kinds of human activity in a video is impossible and it depends on training datasets. Second, even for the current state-of-the-art algorithms, human action recognition requires the modelling of the actions in very high dimensional feature spaces; thus, it is very computationally expensive. Third, the intra-class variation of each human action is very large, due to the different camera viewpoints or different humans, leading to a low recognition rate [1-6].

The goal of this paper is to make the human action recognition system more practicable in real-time applications by improving the recognition rate and accelerating the process to be real-time. The most time-consuming step is the representation of actions in video sequences, because it has to deal with a large number of features. Hence, the overall performance can be enhanced by reducing the runtime in this step.

The appearance and motion information of action in the video can be encoded by the Histogram of Oriented Gradient (HOG) and the Histogram of Optical Flow (HOF)

Open Access http://dx.doi.org/10.5626/JCSE.2015.9.1.29 http://jcse.kiise.org

Received 9 December 2014; **Accepted** 9 March 2015

*Corresponding Author

pISSN: 1976-4677 eISSN: 2093-8020

[7]. In order to better represent the speed and movement characteristics of human actions, two novel descriptors are proposed, so-called the Histogram of Changing Points (HCP) and Average Speed (AS). The combination of these descriptors, e.g., HOG, HOF, HCP, and AS, is proven to be a stronger descriptor for describing the interested features by experimental results.

The main contributions of this work are as follows:

- Two new descriptors, namely HCP and AS, are proposed. The HCP descriptor provides the information on how much the interested points are changed, caused by the action with a low computational time. The AS descriptor presents the speed of actions.
- Evaluation of features for human recognition—the performance of the popular features like HOG and HOF—is evaluated in the aspect of human action recognition. Furthermore, the trajectory information, HOG, and HOF are combined with the newly proposed features of HCP and AS to build a stronger descriptor. By combining those descriptors, the shape, appearance, motion, and speed characteristics of human actions are modelled. The improvement of the combined descriptor is proven through several experiments on different public datasets below.

The paper is organized in 7 sections. Section II reviews the methods related to our work. Overview of our approach is presented in Section III. Section IV describes the descriptors used in this work, including the proposed descriptors. Section V details the use of these descriptors in combination with support vector machine (SVM) for the action classifier. Section VI presents the results of our experiments based on KTH and Hollywood2 dataset, and Section VII concludes the work and discusses the future study.

II. RELATED WORKS

Local space-time features have been successfully applied to action recognition recently. Many different space time feature detectors and descriptors have been proposed in recent years [6-10]. Laptev [7] introduced these features by outstretching the Harris detector for a video. Other approaches in this group are based on the Gabor filter [8], the Hessian matrix [11], and the dense sampling [1], and so on. The role of feature detectors is locating the stable feature points in the spatio-temporal space by maximizing specific saliency functions. As mentioned in [6], the spatio-temporal salient features are used to detect local motion, and they represent video sequences in space and time dimensions.

The space descriptors of 1D or 2D are also proposed, which use gradient information, optical flow, and brightness information [4, 8] for the 3D spatio-temporal of image descriptors such as the 3D-SIFT [12], HOG [13], and LBP-TOP [2, 3, 14]. However, using the 1D or 2D space descriptors to model the video will lose the information in the temporal space; thus, the video is better represented by using the 3D time space descriptors.

Table 1. Comparison accuracy of human action recognition in previous researches

	KTH dataset (%)	Hollywood2 dataset (%)
Laptev et al. [4]	91.8	50.9
Wang et al. [1]	95.8	58.3

Many kinds of trackers have also been applied to the action recognition task recently, such as the KLT tracker [15, 16], the SIFT tracker [17], and the dense sampling tracker [1]. Among them, the dense sampling tracker shows better results on the sparse interest points for action classification [18]. Wang et al. [1, 19] worked on the evaluation for these three trackers and proved that the dense sampling gives the best performance for action recognition task. Table 1 lists the performance on the previous researches on the controlled dataset (KTH) and the uncontrolled dataset (Hollywoods2).

This paper therefore focuses on evaluating and improving descriptors for the recognition of human actions in uncontrolled scene videos. In the paper, the dense trajectory [2], HOG/HOF [4], HCP, and AS descriptors were combined to create a robust descriptor for representing the video sequences. This paper derives dense trajectory as the work of Wang et al. [1], where points are densely sampled and tracked by using the optical flow information. The HOG/HOF descriptor is detailed in [4] for characterizing the local motion and appearance; the HCP descriptor represents movement characteristics; and the AS descriptor describes the speed of human movement.

III. SYSTEM OVERVIEW

The diagram of the human action recognition system of this work is shown in Fig. 1. First, a human detector, such as background subtraction, Motion2D [20], and Calvin upper-body detection [21] which only detects upper-body, is applied to detect human regions. For KTH dataset, our method uses background subtraction method for the detection of human; this dataset has a simple background while Motion2D [20] is used in the Hollywood dataset because it has more complex scenes. Interested points are first detected by dense sampling in the video frame, and then tracked through L frames in the videos to provide the trajectory. Once the trajectory has an expected length, human regions will be described by descriptors such as HOG, HOF, HCP, and AS, based on these interested points.

In order to include the local changing information between human parts, the HCP descriptor is proposed; also, the average speed of human motion is modeled by

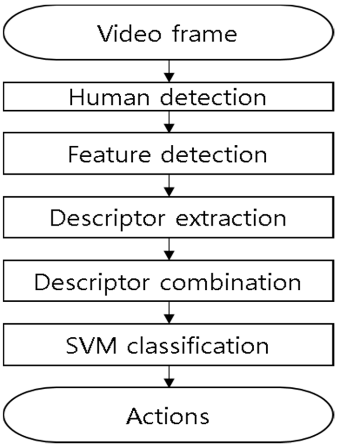


Fig. 1. System overview of human action recognition.

the proposed AS descriptor. The proposed descriptors are then combined with the well-known HOG and HOF to generate a stronger descriptor.

After extracting the descriptors, the classifier is trained by SVM [22] because the classification as SVM has been proved to be one of the best classifiers for the human action recognition task [1, 12, 15, 17, 19].

IV. DESCRIPTORS

A. Dense Sampling

Image representation plays pivotal roles in many computer vision tasks, such as image understanding, object recognition, and scene classification. Solid local feature extraction provides abundant information for the robustness to deformation and occlusions, which has been a fundamental topic since decades ago. Among them, two of the most common techniques are sophisticated interest points (Harris, Hessian matrix, Gabor filter, etc.) [7, 8, 11] and dense sampled points [1, 19]. Interested point detectors are focused on 'interesting' local regions that can help in building the correspondences between images of the same objects or scenes. To this end, high repeatability is required, which is guaranteed by the accuracy and reliability of the feature extraction procedure. Dense sampling descriptors on a regular grid can overcome several drawbacks of interest points, which include limited information, limited coverage regions, and subjective prior knowledge. Dense sampling method has recently demonstrated significant performance in image interpretation, classification, or other tasks.

In order to deal with the scaling factor of humans in videos, the feature points are calculated on many image hierarchical resolution levels. These points are densely sampled on a grid spaced by $W \times W$ pixels. Experimental results show that $W = 5$ is the optimal grid size for all actions in our experimental datasets. In this work, three scales for video frames are considered, and the spatial scale is modified by a factor of $1/\sqrt{2}$. The number of tracked points is downsized by the algorithm of Shi and Tomasi [23]. Depending upon this work [23], the points on the grid are removed; and the sampled points, whose eigenvalues of the auto-correlation matrix are smaller than a threshold T, are deleted in the trajectory.

B. Dense Trajectory

The dense trajectory is generated based on dense local patches. In this study, the dense trajectory is adopted by the work of Wang et al. [1]. After filtering, sampled points are tracked though L video frames to give enough length for the trajectory. These frames are then represented by the HOG, HOF, HCP, and AS descriptors, which will be presented in detail in the next sections.

To compute the dense trajectories, the local patches are sampled as the dense sampling, as described in Section IV-A. Tracking is performed on patches by the median filter in a dense optical flow. Each point $P_t = (x_t, y_t)$ at frame t is matched to another patch (or point) P_{t+1} at frame t+1 by the following equation:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + H \times w \tag{1}$$

where H is the median filter, and $w = (u_t, v_t)$ is the optical flow. The dense optical flow is computed by using the Farneback algorithm [24]. The drifting problem is avoided by setting the maximum length of the trajectory. Experiments in this work show that using the maximum length of 15 gives the best results. This step is applied to the trajectories whose length exceeds a threshold T.

The shape of the trajectory (the shape of motion) is presented by encoding the local motion patterns. This work defines that the vector of the sequence is $T = (\Delta P_t, \Delta P_{t+1}, ..., \Delta P_{t+L+1})$ where $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. In order to make the trajectory shape descriptor invariant to scale changes, the concatenated vector is normalized by the overall magnitude of motion displacements:

$$TS = \frac{T}{\sum_{j=t}^{t+L-1} |\Delta P_j|} \tag{2}$$

where $L = 15$ is maximum length of trajectories. Thus, the trajectory descriptor has the length of 30 (for both x-direction and y-direction).

C. Histogram of Optical Flow/Histogram of Oriented Gradient

1) Histogram of Optical Flow

The optical flow is calculated by using Farneback algorithm [24]. The aligned window sequence is used to compute optical flow in the consecutive frames. As shown

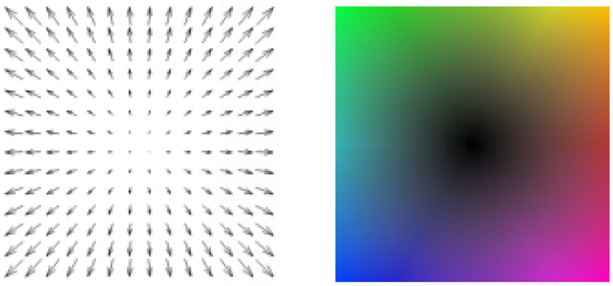


Fig. 2. An illustration of the optical flow coloring scheme. Reproduced with permission from [25].

in Fig. 2, the optical flow vectors calculated between two frames indicate the directions of motion. Therefore, the histogram of optical flow field contains huge information about motions in a video. Let $\theta_i(x, y)$ be the direction of optical flow vector at the pixel (x, y) in frame I; and θ_i is defined by:

$$\theta_i(x, y) = \arctan\frac{d_y}{d_x} \tag{3}$$

where d_x and d_y are the displacements in the x and y directions, respectively, $\theta_i \in [0^\circ, 360^\circ]$. These orientations in the human block are accumulated into a histogram, which is then normalized as in Eq. (9) through the length of trajectory L.

$$Desc_{HOF_j} = \sum_{i=j-L}^{j} f_i \tag{4}$$

The vector of HOF value is calculated by Eq. (4), where j is the current frame, L is the length of trajectory, and f is HOF value of i-th frame. The dimension of the HOF in one frame is 8.

2) Histogram of Oriented Gradient

The HOG, introduced in [26] for action recognition, is a descriptor used for the purpose of object detection in image processing and computer vision areas. This method counts the occurrences of gradient orientation in the localized portions of an image.

The HOG descriptor has some advantages in representation of the moving object. Because the HOG descriptor operates on localized blocks, it is invariant to the local geometric and photometric transformations. The HOG descriptor is particularly suitable for describing actions.

At the first step, the x and y derivatives of the grey scale image are computed by applying the Sobel operator [26] as:

$$D_x = [-1 \;\; 0 \;\; 1] \tag{5}$$

$$D_y = [-1 \;\; 0 \;\; 1]^T \tag{6}$$

After calculating x, y derivatives (denoted by I_x and I_y, respectively), the magnitude and orientation of the gradient is computed as:

$$|G| = \sqrt{I_x^2+I_y^2} \tag{7}$$

and

$$\theta = arctan\frac{I_y}{I_x} \tag{8}$$

where $\theta \in [0^\circ, 360^\circ]$.

The second step is to compute 9-bin histograms at descriptor blocks, based on the orientation. The size of block is the same as that of the blocks of the HCP descriptor that will be detailed in the next section. For each pixel, the orientation is used to decide the bin for voting, with the weight of corresponding magnitude $|G|$. The histogram is accumulated through L frames to generate the HOG descriptor, and the L value is set up at 15 in our experiments.

The third step is to normalize the block's orientation histogram. Although there are three different methods for block normalization, the l_2-norm normalization is implemented as the following equation:

$$f = \frac{v}{\sqrt{\|v\|_2^2+e^2}} \tag{9}$$

where f is normalized vector of v which is the orientation histogram vector, and e is a constant of exponential number.

As described above, the HOG value is calculated as 9-bin histogram descriptor describing the local appearance, and it is considered in space-time volume by the following equation:

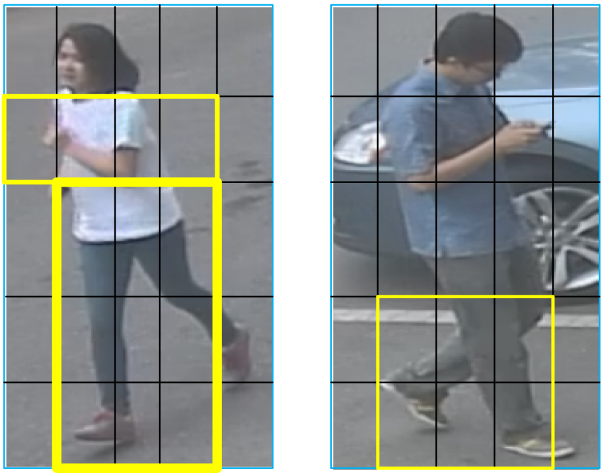
$$Desc_{HOG_j} = \sum_{i=j-L}^{j} f_i \tag{10}$$

where j is the current frame, L is length of trajectory, and f is HOG value at frame j. $Desc_{HOG_j}$ has 9 dimensions as the HOG dimension.

D. Histogram of Changing Points

In order to distinguish actions more accurately, this paper proposes a new descriptor named HCP to represent the local motion of each block inside the detection window. To compute the HCP descriptor, the detection window is divided into $n \times n$ blocks (the value of n is 5 in Fig. 3). As shown in Fig. 3, the change of interested points, in terms of position, was concentrated on the yellow regions where the thicker border indicates the higher change of interested points. The positions of interested points are much different between 'Running' action (Fig. 3(a)) and 'Walking' action (Fig. 3(b)) in the same camera view. Therefore, the HCP vector is an informative descriptor for distinguishing action labels for actions with large movements of body parts.

The procedure below is repeated in each frame. If the



(a) (b)

Fig. 3. Different HCP vector between running and walking action.

number of interest points within block k at the i-frame is IP_k^i, the change in number of interest points in each block φ_k^i is calculated frame by frame using the following equation:

$$\varphi_k^i = \left|IP_k^i - IP_k^{i-1}\right| \tag{11}$$

where $k=1,2,...,n^2$. Then, total change in interest points ω^i for a given frame i is calculated and accumulated through the operation, by using the equation below:

$$\omega^i = \sum_{k=1}^{n2} \left|IP_k^i - IP_k^{i-1}\right| \tag{12}$$

Finally, the changing percentage in the number of interest points (PC) for all of the blocks is calculated by the following equation:

$$PC_k^j = \sum_{i=j-L}^{j} \left(\frac{\varphi_k^i}{\omega^i}\right) \times 100\% \tag{13}$$

The value of PC_k^j is the changing percentage of the number of interest points in frame j at block k representing the magnitude of movement caused by the action in the block. It can be different in some actions carried out by the upper body and in other actions carried out by the lower body, such as ‘boxing’ and ‘walking’. Therefore, the HCP descriptor is an informative descriptor for distinguishing actions in videos, as described in the equation. The HCP is calculated by accumulating PC value from L frames, as shown in Eq. (14):

$$Desc_{HCP} = \cup_{i=1}^{n2} \sum_{j=0}^{L} PC_i^j \tag{14}$$

where $n \times n$ is the number of divided blobs, and PC_i^j is the change value of the block i at frame j. This vector represents the movement property in each part of an object caused by actions. Consequently, the HCP vector provides significant information to describe human actions in a video, which is proven as an effective descriptor for increasing the recognition rate in this problem.

E. Average Speed

This paper also proposes the AS descriptor that accumulates the average speed of all interested points in current frame to represent the speed of the human motion. The speed of a point can be decomposed into two directions, x and y, as the following equations show:

$$\Delta x = x_{P_i} - x_{P_{i-1}} \tag{15}$$

$$\Delta y = y_{P_i} - y_{P_{i-1}} \tag{16}$$

$$S_{P_i} = \sqrt{\Delta x^2 + \Delta y^2} \tag{17}$$

where (x_{P_i}, y_{P_i}) is the coordinate of the point P_i, and S_{P_i} is the speed of the point P_i. The speed of a point is calculated based on the distance of the interested point, between the current frame and the previous frame on x and y directions. Therefore, average speed is then calculated as a summation of the speed of interested points divided by the number of interested points:

$$AS = \frac{\sum_{i=1}^{T} S_{P_i}}{T} \times F \tag{18}$$

where AS is the average speed of interested points in the current frame, F is video frame rate value, and T is the number of interested points. The value of AS, which is the speed of human actions, represents average speed of interested points. This means that the descriptor is significant for distinguishing human actions with different speeds.

F. The Proposed Combination Descriptor

In this paper, the motions are described by combining the HOG, HOF, HCP, and AS descriptors. The HOG descriptor focuses on the static appearance information, whereas the HOF descriptor captures the local motion information. Calculating the HOG and HOF descriptors along the dense trajectories provides the tracking of appearance and local motions information. This combination guarantees that the features of motion are extracted with the dense trajectories to characterize shape and motion. Moreover, this method also proposes a descriptor for calculating the change of actions at different body parts, called the HCP descriptor. The combined descriptor also contains the trajectory information, namely the shape of trajectories derived from Eq. (2). Consequently, the combined descriptor contains information of HOG, HOF, HCP, and AS descriptors, as shown in Eq. (19):

$$Desc = TS \cup Desc_{HOG} \cup Desc_{HOF} \cup Desc_{HCP} \cup Desc_{AS} \tag{19}$$

where TS is the shape of the trajectory computed by Eq. (2).

At this step, the optical flow field is already calculated when generating the dense trajectory. By combining these descriptors, the new descriptor can represent human actions in more detail, in terms of appearance, local motion, motion change, and motion speed. The length of the combined descriptor is the total length of these five descriptors at $30 + 9 + 8 + n^2 + 1 = 48 + n^2$ dimensions.

Almost all previous approaches have calculated descriptors on the entire frame, which is leading to a time-consuming process. For reducing the descriptor computation time, this work only considers human body regions. In this paper, for the static scene dataset KTH, where there is only one actor, the background subtraction and Motion2D are used to detect the human region. For Hollywood2, the cameras are moving, so we use Motion2D and the Calvin upper-body detector [21]. This detector returns bounding boxes fitting the head and upper half of torso of a person, which segment the video frame into foreground/background corresponding to person/non-person regions. As shown in Fig. 3, the results of human detection have high probability of obtaining the objects of people, in terms of motion or a moving state, significantly reducing the processing time.

V. CLASSIFICATION

In machine learning, the SVMs, introduced in [22], are supervised learning models with associated learning algorithms that analyze data and recognize patterns which are used for classification and regression analysis. Given a set of training examples, each is marked as belonging to one of the two categories. The SVM model is a representation of the examples as points in space, which is mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

In classification, a multi SVMs model is utilized for training with the RBF χ^2 kernel (EMD kernel) [7]. As similar to [1], the different descriptors are combined in multi channels approach:

$$H(x_i, x_j) = \exp\left(-\sum_c \frac{D(x_i^c, x_j^c)}{M^c}\right) \tag{20}$$

where $D(x_i^c, x_j^c)$ is the χ^2 distance (or the EMD distance) between video x_i and video x_j, with respect to c-channel. M^c is the mean value of the χ^2 distance between the training samples for the c-channel.

The one-against-rest approach, which trains the classifier for each possible pair of classes, is used for multi classification and the class is assigned with the highest score. For each new test pattern, all binary classifiers are evaluated, and the pattern is assigned to the class that is chosen by the majority of classifiers.

VI. EXPERIMENTS

A. Dataset

1) KTH Dataset

The KTH is an action video dataset published in 2004 by Laptev et al. [27]. This dataset contains six kinds of actions as follows: walking, jogging, running, boxing, hand waving, and hand clapping. Each action type was performed several times by 25 people in four different scenarios as follows: outdoors *s1*, outdoors with scale variation *s2*, outdoors with different clothes *s3*, and indoors *s4*. Currently, this dataset contains 2,391 sequences. All sequences were taken over the homogeneous backgrounds with a static camera with 25 fps. The sequences were down-sampled to the spatial resolution of 160 × 120 pixels, and they have a length of four seconds on average. Therefore, there are 25 × 6 × 4 = 600 video files for each combination of 25 subjects, 6 actions, and 4 scenarios. We followed the original experimental setup of [1] that divides the samples into test sets (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and training sets (the remaining 16 subjects). This work trains and evaluates multi-class classifiers and reports the accuracy average over all classes.

2) Hollywood2 Dataset

The Hollywood2 action dataset contains 12 classes associated with 12 actions as follows: answer phone, driver car, eat, fight person, get out of car, hand shake, hug person, kiss, run, sit down, sit up, and stand up. All samples were collected by the means of automatic scrip-to-video alignment in combination with text-based scrip classification. This dataset includes the videos for training, testing, and automatic training subset; the test subset is provided with manually checked action labels. In total, there are 1,707 video sequences which are divided into a training set (823 sequences) and a testing set (884 sequences). Training and testing sequences come from different movies.

B. Experiments

1) Accuracy Evaluation

The accuracy of human action recognition is evaluated on two datasets, which are described in the previous section called KTH and Hollywood2 dataset. Table 2 displays the average accuracy over all classes of this work, compared with the state-of-the-art human action recognition methods recently used on the KTH dataset.

As shown in Table 2, the average accuracy achieves

Table 2. Mean average accuracy over all classes and processing time, comparing with state-of-the-art method on KTH dataset

Method	Average accuracy (%)
Laptev et al. [4]	91.80
Wang et al. [1]	95.88
Proposed method	96.20

Table 3. Evaluation of each descriptor on KTH and Hollywood dataset (%)

Dataset	HOG/HOF	HCP/AS	Combination
KTH	92.20	95.88	96.20
Hollywood2	46.30	58.90	60.01

HOF: Histogram of Oriented Gradient, HOF: Histogram of Optical Flow, HCP: Histogram of Changing Points, AS: Average Speed.

96.20% on the KTH dataset comparing with the previous methods. This table proves that our proposed descriptors (HCP, AS) and human detection algorithm for recognizing actions in a video is very accurate.

Table 3 evaluates the accuracy of each descriptor and combined descriptor. This table shows that almost all the chosen descriptors are impressive, regarding the precision of the recognition system. Among them, the HCP/AS descriptor is the most informative descriptor for human actions, with the accuracy as 95.88% and 58.90% in the KTH and Hollywood2 dataset, respectively. Consequently, the speed of actions and change characterized by each part of the body are informative descriptor for distinguishing usual actions. These two descriptors increase the accuracy of our human action recognition system, meanwhile reducing the processing time for each video frame. By combining HOG/HOF and HCP/AS descriptors with dense trajectory, our proposed approach shows that the accuracy of this system is 96.20% on KTH dataset and 60.01% on Hollywood dataset.

2) Processing Time

The results of Table 4 show a significant improvement achieved by our method, when compared to other recently reported results on the same dataset and obtained by using the same experimental settings. For evaluating our approach, several experiments on the KTH and Hollywood2 dataset were conducted.

The number of frames per second is also shown in Table 4, which proves that recognition actions inside human regions can improve the performances and processing time. Therefore, this method is meaningful for various descriptors, especially with the descriptors that generate a large number of features, such as the dense trajectory descriptor. This table also presents the effectiveness of our approach that detects human regions before calculating descriptors. The descriptors are only calculated on human regions and not on all pixels inside the frame, so the number of frames per second is dramatically increased to 10 frames per second.

Table 4. Comparison of number of frames per second with state-of-the-art method in computing descriptors process

Descriptor	Dense + HCP/AS	Dense + HOG/HOF	Combination
Wang et al. [19]	NA	2.4	NA
Present study	18	15	13

HOF: Histogram of Oriented Gradient, HOF: Histogram of Optical Flow, HCP: Histogram of Changing Points, AS: Average Speed.

VII. CONCLUSION

There are many detectors which can be used for detecting features in a video. Among them, the dense sampling is proven to bring high accuracy for action recognition in a video. This paper is different from the previous approaches toward action recognition, as we only processed human regions by the background subtraction, Motion2D, and Calvin detect upper-body algorithms, and not the entire video frame; so, the processing time is reduced appreciably. To achieve high accuracy for the human action recognition problem, the method to represent actions is important. This paper chose the combination of the dense trajectory, HOG/HOF, and HCP/AS descriptors to represent actions on a video frame, because it is a combination of the action speed, trajectory shape, appearance, and motion information of the video. Hence, they can provide meaningful information for recognizing human actions.

This work proposes the HCP/AS descriptor that is more operative if the human block is divided into many small blocks, with increasing *n*. However, when *n* is increased, this algorithm becomes more time-consuming. Therefore, we have to choose *n* that suits for both the processing time and accuracy. In our experiments, 5 is the best value of *n*.

For the feature improvement, this work can use algorithms to detect parts of the body, such as the poselet approach, then apply the HCP descriptor to represent each part of the body. This idea makes descriptor more distinguishable for human actions, especially for small actions.

AKNOWLEDGMENTS

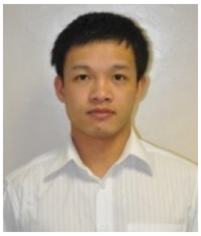
This research was funded by the Ministry of Science, ICT & Future Planning (MSIP), Korea in the ICT R&D Program of 2014.

REFERENCES

1. H. Wang, A. Klaser, C. Schmid, and C. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60-79, 2013.
2. L. Nanni, S. Brahnam, and A. Lumini, “Local ternary patterns from three orthogonal planes for human action classification,” *Expert Systems with Applications*, vol. 38, no. 5, pp. 5125-5128, 2011.
3. R. Mattivi and L. Shao, “Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor,” in *Proceedings of the 13th International Conference on Computer Analysis Images and Patterns (CAIP)*, Munster, Germany, 2009, pp. 740-747.
4. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, 2008, pp. 1-8.
5. M. Jain, H. Jegou, and P. Bouthemy, “Better exploiting motion for better action recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, 2013, pp. 2555-2562.
6. A. H. Shabani, D. A. Clausi, and J. S. Zelek, “Improved Spatio-temporal Salient feature detection for action recognition,” in *Proceedings of British Machine Vision Conference (BMVC)*, Dundee, UK, 2011, pp. 1-12.
7. I. Laptev, “On space-time interest points,” *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107-123, 2005.
8. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, Beijing, China, 2005, pp. 65-72.
9. H. Jhuang, T. Serre, L. Wolf, and T. Poggio, “A biologically inspired system for action recognition,” in *Proceedings of IEEE 11th International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007, pp. 1-8.
10. K. Y. K. Wong and R. Cipolla, “Extracting spatio-temporal interest points using global information,” in *Proceedings of IEEE 11th International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007, pp. 1-8.
11. G. Willems, T. Tuytelaars, and L. Van Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector,” in *Proceedings of 10th European Conference on Computer Vision (ECCV2008)*, Marseille, France, 2008, pp. 650-663.
12. P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional SIFT descriptor and its application to action recognition,” in *Proceedings of the 15th ACM International Conference on Multimedia*, Augsburg, Germany, 2007, pp. 357-360.
13. A. Klaser, M. Marszałek, and C. Schmid, “A spatio-temporal descriptor based on 3D-gradients,” in *Proceedings of British Machine Vision Conference (BMVC)*, Leeds, UK, 2008, p. 1-10.
14. T. Ahonen. A. Hadid, and M. Pictikainen, “Face recognition with local binary patterns,” in *Proceedings of the 8th European Conference on Computer Vision (ECCV2004)*, Prague, Czech Republic, 2004, pp. 469-481.
15. P. Matikainen, M. Hebert, and R. Sukthankar, “Trajectons: action recognition through the motion analysis of tracked features,” in *Proceedings of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshop)*, Kyoto, Japan, 2009, pp, 514-521.
16. R. Messing, C. Pal, and H. Kautz, “Activity recognition using the velocity histories of tracked key points,” in *Proceedings of IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 104-111.
17. J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and J. Li, “Hierarchical spatio-temporal context modeling for action recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, 2009, pp. 2004-2011.
18. L. Fei-Fei, and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005, pp. 524-531.
19. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *Proceedings of British Machine Vision Conference (BMVC)*, London, UK, 2009, pp. 1-11.
20. J. M.Odobez and P. Bouthemy, “Robust multiresolution estimation of parametric motion models,” *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348-365, 1995.
21. M. Eichner and V. Ferrari, “Calvin upper-body detector v1.04,” http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/.
22. N. Deng, Y. Tian, and C. Zhang, *Support Vector Machines-Optimization Based Theory, Algorithms, and Extensions*, Boca Raton, FL: CRC Press, 2013.
23. J. Shi and C. Tomasi, “Good features to track,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2005, pp. 593-600.
24. G. Farneback, “Two-frame motion estimation based on polynomial expansion,” in *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA)*, Halmstad, Sweden, 2003, pp. 363-370.
25. J. S. Perez, E. Meinhardt-Llopis, and G. Facciolo, “TV-L1 optical flow estimation,” *Image Processing On Line*, vol. 3, pp. 137-150, 2013.
26. Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, 2011, pp. 3361-3368.
27. Recognition of human actions, “http://www.nada.kth.se/cvap/actions/.”

Thi Ly Vu

Thi Ly Vu received B.S. and M.S. degrees from Le Quy Don Technical University, Vietnam, and Inha University, Korea in 2011 and 2014, respectively. She is currently working as a lecturer at Le Quy Don Technical University, Vietnam in Department of Information Technology. Her research interests include pattern recognition, image processing, and intelligent processing system.



Trung Dung Do

Trung Dung Do is a graduate student in the doctor's program of School of Information and Communication Engineering, Inha University, Korea. He received his bachelor's degree in Computer Technology from National Research Irkutsk State Technical University, Russia, in 2010, and master’s degree from Graduate School of Information and Communication Engineering, Inha University, Korea, in 2014. His research interests include computer vision, pattern recognition, image processing, and machine learning.

Cheng-Bin Jin

Cheng-Bin Jin received his B.S. and M.S. degrees from Yanbian University, China, in the Department of Computer Science and Technology, in 2009 and 2014, respectively. He is currently pursuing a Ph.D. in the Computer Vision Lab of the Graduate School of Information & Communication Engineering at Inha University, Korea. His research interests include action recognition in surveillance, human computer interaction, machine learning, and pattern recognition.

Shengzhe Li

Shengzhe Li is a student in mater and Ph.D. combined course at School of Information and Communication Engineering, Inha University, Korea. He received his bachelor’s degree in Software Engineering from Beihang University, Beijing, China in 2008. He is a co-founder and researcher at Global R&D Center, VisionIn, Inc. His research interests include biometrics, computer vision and medical image processing.

Van Huan Nguyen

Van Huan Nguyen received his B.S. from Hanoi University of Science and Technology, Vietnam in the Department of Applied Mathematics and Informatics, in 2005; and the M.S. and Ph.D. degrees in the Computer Vision Lab. of the Graduate School of Information & Communication Engineering at Inha University, Korea, in 2008 and 2012, respectively. He is currently working as a post-doc researcher in the Super Intelligence Technology Center, Inha University. His research interests include biometrics, pattern recognition, remote sensing, video processing, and video surveillance system.

Hakil Kim

Hakil Kim is a professor at School of Information and Communication Engineering, Inha University, Korea. He received the B.S. degree in Control and Instrumentation Engineering from Seoul National University, Korea, in 1983, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University, West Lafayette, IN, USA, in 1985 and 1990, respectively. His research area includes computer vision and pattern recognition. He has been actively participating in IOS/IEC JTC1-SC37 and ITU-T/SG17 WP2/Q.8 Telebiometrics as a project editor and a rapporteur.

Chongho Lee

Chongho Lee received his M.S. degree in Electrical Engineering from Seoul National University, Korea and Ph.D. degree from Iowa State University, USA in department of Computer Engineering. His research areas are Artificial Neural Networks, and Intelligent System. He is currently a Professor in School of Information & Communication Engineering at Inha University, Incheon, Korea. He joined in Dynamic Partial Reconfigurable FIR filter design, LNCS 3839, Mar. 2006, Design of CSVM Processor for Intelligence Expression, Journal of Electrical and Electronic Material, Feb. 2007 and DNA-inspired CVD Diagnostic Hardware Architecture, Journal of KIEE, Feb. 2008.