Feature Integration with Random Forests for Real-time Human Activity Recognition

Hirokatsu Kataoka†, Kiyoshi Hashimoto†, Yoshimitsu Aoki†
† Keio University

ABSTRACT

This paper presents an approach for real-time human activity recognition. Three different kinds of features (flow, shape, and a keypoint-based feature) are applied in activity recognition. We use random forests for feature integration and activity classification. A forest is created at each feature that performs as a weak classifier. The international classification of functioning, disability and health (ICF) proposed by WHO is applied in order to set the novel definition in activity recognition. Experiments on human activity recognition using the proposed framework show - 99.2% (Weizmann action dataset), 95.5% (KTH human actions dataset), and 54.6% (UCF50 dataset) recognition accuracy with a real-time processing speed. The feature integration and activity-class definition allow us to accomplish high-accuracy recognition match for the state-of-the-art in real-time.

Keywords: Activity Recognition, Feature Integration, Random Forests.

1. INTRODUCTION

In past years, various techniques for human sensing have been studied in the field of computer vision [1]. Human tracking, posture estimation, and face recognition are some examples, which have been applied in real-life environments. Recently, human activity recognition is the most popular research topic. We have been studying techniques of how to comprehend human activities and utilize them for our living spaces. In particular, the applications of interest are surveillance, sports video analysis, medical science, robotics, video indexing and games. To put these applications into practice, many recognition methods have been proposed in recent years to improve accuracy. Activity recognition means determining the activity of the person from a sequence of consecutive images. It is promising to computer vision applications if we can understand the human activity, for example, human activity understanding in daily scenes and activity based video indexing.

In this paper, we propose a real-time activity recognition applying feature integration with random forests. Our framework combines three different kinds of features namely flow, shape, and keypoint-based feature. The contributions of this paper are: (i) Three different features are integrated with random forests [2]. Three feature types are flow (HOOF [3]), shape (extended version of CoHOG [4]) and keypoint (SURF [5] + Bag-of-features (BoF) [6]). (ii) The activity-class consideration based on the international classification of functioning, disability and health (ICF) proposed by WHO.

Related works on activity recognition are discussed here: Several feature descriptors have been proposed to understand difficult activities; Laptev *et al.* [7] introduced Space-Time Interest Points (STIP) as an improvement of Harris corder detector. The STIP represents 3D (XYT) interested motion points to extract the same feature in an activity. This framework is widely used in activity recognition community. Klaser proposed 3D-HOG *et al.* [8] and Marszalek *et al.* [9] described feature combination by using STIP framework. Chaudhry *et al.* implemented a remarkable feature descriptor on a benchmark using a time series optical flow based method [3].

2. PROPOSED FRAMEWORK

Figure 1 shows the framework of the proposed method for activity recognition. The system comprises of three features, and random forests. The three features are flow, edge, and keypoint based feature. In feature extraction, we set space-time representation for all features. The whole features are consist of features from several frames. In addition, we define activity-class using the ICF. The experimental results show that our proposed method is of high-accuracy and real-time processable on Weizmann action dataset [10], KTH human actions dataset [11] and UCF50 dataset [12].

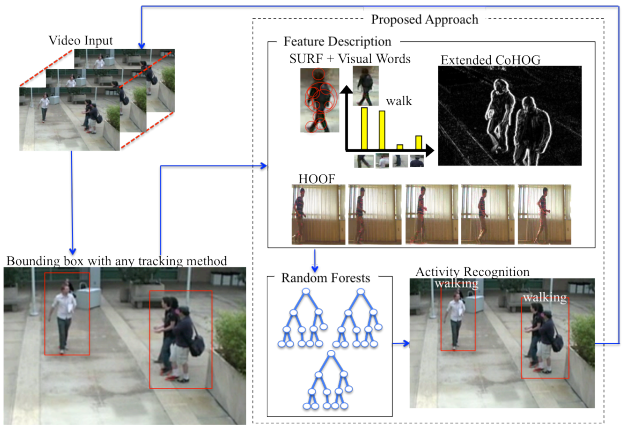

Figure 1. Framework

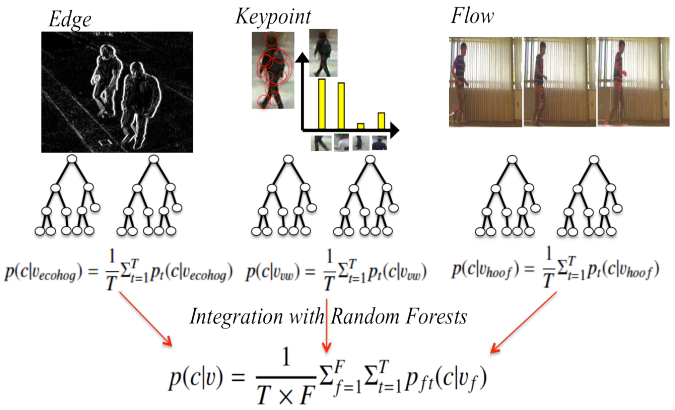

Figure 2. Feature Integration with random forests

The rest of the paper is organized as follows. Firstly, in Sections 3, we describe activity recognition framework. Secondly, Section 4 presents experimental results on human activity recognition using datasets. Finally, Section 5 concludes the paper.

3. ACTIVITY RECOGNITION

We believe the integration of features will improve the accuracy of activity recognition. In particular, using three different kinds of features should ensure good classification performance. In activity recognition, we apply three kinds of features which are flow, shape, and a keypoint-based feature. Random Forests are used to integrate the three different feature descriptors, namely, Histograms of Oriented Optical Flow (HOOF) [3], an improved version of Co-occurrence Histograms of Oriented Gradients (CoHOG) [4], and BoF [6] representation of SURF [5] for human activity recognition. We improved CoHOG as the Extended CoHOG (ECoHOG) by adding edge magnitudes from two different pixels, step extraction to give a spatial gap in the extraction window and a spatio-temporal representation. At the same time, BoF+SURF was improved by using a time-series representation.

3.1 Feature Descriptor for Activity Recognition

Histograms of Oriented Optical Flow (HOOF) [3]. The HOOF algorithm was proposed by Chaudhry *et al.* [3]. HOOF accumulates the optical flow as a histogram representing the human motion feature. This algorithm is independent of the scale of the moving person as well as the direction of motion. This spatio-temporal representation is effective in activity recognition. In HOOF extraction, the optical flow is computed for each frame of the human motion. Each flow vector is accumulated as quantized data comprising the feature histogram. HOOF stores four bins from the eight divided directions. This representation is independent of the human direction owing to the different motion. We use the original HOOF algorithm for flow-based feature representation. The HOOF is acquired from 25 images yielding a 100 dimension histogram.

Extended Co-occurrence Histograms of Oriented Gradients (ECoHOG). We believe that edge-magnitude information is beneficial in acquiring the detailed human shape to understand human activities. We improved CoHOG [4] as ECoHOG, by adding edge magnitudes representing the human shape and step extraction in order to capture a wide range of features. The ECoHOG can describe edge-pair information from two different pixels, that is the head and the shoulders. Watanabe *et al.* explained that a co-occurrence representation is better than Histograms of Oriented Gradients, which represent only a single edge's orientation.

Speeded-Up Robust Features (SURF) [5] + Bag-of-features (BoF) [6]. We can express the keypoint feature which represents important areas of the human body, as a histogram using BoF [6], while SURF extracts the specified features for each activity. SURF is a feature descriptor for creating codewords, formed from clipped images capturing human activities. We obtain features from 25 frames, each contributing 20 dimensions, giving a total of 500 dimensions in the feature vector.

3.2 Feature Integration with Random Forests

The process flow of random forests is shown in Figure 2. We integrated three values of feature descriptor with random forests [2]. The random forests are a connection of tree classifier for returning probability distribution. The probability value of classifier is output at each tree. In learning step, random forests randomly select feature dimensions to split tree nodes. Information gain is applied in node splitting. A leaf node has a probability distribution, and random forests perform classification with bagging algorithm.

The equation of random forests classification shown below:

$$p(c \mid v) = \frac{1}{T} \sum_{t=1}^{T} p_t(c \mid v) \tag{1}$$

where T is the number of trees in random forests ($t \in \{1, 2, ..., T\}$), v is the input feature vector, c is the class of activity. $p(c|v)$ is the predicted value of posteriori by t-th decision tree. Moreover, we integrate the three kinds of feature into one classifier with the predicted value of posteriori. The classifier is shown as below:

$$p(c \mid v) = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} p_{ft}(c \mid v_f) \tag{2}$$

where F is the number of feature ($f \in \{1, 2, 3\}$: HOOF, ECoHOG, and BoF+SURF), v_f is each feature vector. $p_{ft}(c|v)$ is the predicted value by t-th tree and f-th feature of decision tree. The sum of tree value allows us to integrate and select three features in classification step.

3.3 Activity Definition based on the International Classification of Functioning, Disability and Health

The International Classification of Functioning, Disability and Health (ICF), proposed by the World Health Organization (WHO) in 2001 [13]. Moreover, the ICF defined certain activities in daily life, from which we selected the activities for our framework.

4. EXPERIMENT

To show the effectiveness of our proposed method, we performed an experiment comparing our system with previous methods. Scenes were taken from the Weizmann action [10], KTH human action[11], and UCF50 datasets [12] to evaluate the effectiveness of our framework for complicated activities. Our algorithm runs in real-time at around 100 fps on a standard laptop PC (Intel Corei-7 2.7GHz, 4.0GB RAM). At this point, our recognition system operates on the clipped rectangle from the tracked human. Table 1 indicates the accuracy of proposed approach and related works.

Evaluation on the Weizmann dataset [10]. This dataset contains 90 videos separated into 10 activities (bend, jack, jump, pjump, run, side, skip, walk, wave, wave2) performed by nine persons [10]. Table 1 shows the accuracy of our proposed method compared with previous methods. Our approach obtained 99.2% when integrating the three improved flow, edge, and keypoint based feature descriptors. The proposed approach confuses to divide wave (waving hand with one hand) and wave2 (waving hand with two hands). We cannot divide these two activities using flow descriptor HOOF. Analyzing the feature descriptor, space-time ECoHOG was effective for classification of wave and wave2. Keypoint based feature is not specific solution because of the method cannot distinguish one hand or two hands, or left or right hand. Although it is not easy to differentiate the two activities, the proposed approach achieved the 98% and 96% accuracy on wave and wave2. In this experiment, proposed approach (99.2%) is superior to Wang *et al.* (97.2%) [14] in recognition accuracy. Moreover, our proposed approach runs at 109.2fps on the Weizmann action dataset that shows a remarkable achievement in processing time.

Table 1. Recognition accuracy on the Weizmann, KTH, and UCF50 dataset.

Framework	Accuracy (%) on Weizmann	Accuracy (%) on KTH	Accuracy (%) on UCF50
Proposed Method	99.2	95.5	54.6
Wang et al. [14]	97.2	94.2	-
Chaudhry et al. [3]	94.4	75.9	28.7
Laptev et al. [7]	88.8	91.8	47.9

Figure 3. Activity recognition with ICF

Evaluation on the KTH dataset [11]. The KTH human action dataset contains 600 videos separated into six activities (boxing, handclapping, handwaving, jogging, running, and walking) [11]. It includes camera motion and shadows in low resolution images (*160 * 120 pixels*). Table 2 gives the accuracy of our proposed method and previous methods.

Our proposed approach achieved 95.5% on the KTH human actions dataset. The proposed method is a close match for which Wang *et al.* with respect to accuracy, i.e. 95.5% as compared to 94.2%. Our framework, however, has a greater processing cost than the method in [14]. We verify the effectiveness of bagging not only for one-feature random forests but multi-feature random forests.

Evaluation on the UCF50 dataset [12]. The UCF50 dataset [12] comprises 1168 videos in 50 categories collected from YouTube. There are many categories in this dataset, for example, Baseball Pitch, Breaststroke, Playing Guitar, Jumping Jack, Punch, Tennis Swing and Walking with a dog. The dataset also includes several computer vision difficulties, such as camera motion, complicated backgrounds, occlusions and personal variations. Table 1 gives the results for our random forests technique and that of for the frameworks.

Consideration. According to the experiments on three datasets, our proposed approach performs better than the other methods such as Laptev et al. [7] and Chaudhry et al. [3]. The proposed approach is based on feature integration with three types of descriptor. The three descriptors (HOOF, ECoHOG and BoF+SURF) represent different types of feature in an image, respectively. HOOF effectively expresses motion feature by using Lucas-Kanade optical flow and

vectorization. The motion feature is considering horizontal symmetry to handle the same activities depending on different directions (e.g. walking right-left and left-right should be dealt with the same feature). ECoHOG and BoF+SURF are similar approaches because of the features describe human shape in an image. ECoHOG represents whole human shape with co-occurrence feature, and BoF+SURF is consists of visual words from a human body. Moreover, random forests [2] are significant approach in order to combine three features. In random forests, each tree returns probability distribution from a feature vector. We learn and combine probability distribution in each feature descriptor as shown in 3.2. Random forests enhance classification ability with accumulation of probability distribution.

Evaluation on daily videos. We verified the effectiveness of the proposed framework in daily scenes. We have captured over 20 hours in laboratory and indoor-room. Figure 3 shows the examples of activity recognition and processing time. The activities are recognized based on ICF. The activity definition is ICF as shown in section 3. The proposed framework performs at around 50 fps with 70-80% accuracy in daily scenes. As a result of ICF definition, we can divide activities based on criterion whereas the simple feature descriptor would allow fast processing for daily scenes. For example, "walking" and "running" are confusing classes each other. In daily scenes, we can combine these two activities into "moving". The ICF defines moving as "d455: moving around". At the same time, we can improve recognition accuracy with activity definition. The activity definition should be defined depending on the situations and the activities.

5. CONCLUSION

We proposed a real-time recognition method for human activities. Three different kinds of feature are integrated in a random forests framework. The proposed approach achieved 99.2%, 95.5%, and 54.6% recognition accuracy on Weizmann, KTH and UCF50 dataset, respectively. In the future, we would like to include posture and object information for the ICF-based activity-class classification. The ICF contains a lot of activities that include a static posture with objects (e.g. d4301: carrying in the hands, d5402: putting on footwear).

REFERENCES

[1] T. B. Moeslund, A. Hilton, V. Kruger, "A survey of advances in vision-based human motion capture and analysis", Computer Vision and Image Understanding (CVIU), vol.104-2, pp.90-126, (2006).

[2] L. Breiman, "Random Forests", Machine Learning, vol.45-1, pp.5-32, (2001).

[3] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, "Histograms of oriented optical flow and binet cauchy kernels on nonlinear dynamical systems for the recognition of human actions", in CVPR, pp.1932-1939, (2009).

[4] T. Watanabe, S. Ito, K. Yokoi, "Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection", PSIVT2009, pp.37-47, (2009).

[5] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol.110, No.3, pp.346-359, (2008).

[6] G. Csurka, C. Dance, L. X. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints", ECCV International Workshop on Statistical Learning in Computer Vision, (2004).

[7] I. Laptev,"On Space-Time Interest Points", in IJCV, No. 64, pp.107-123, (2005).

[8] A. Klaser, M. Marszalek, C. Schmid, "A spatio-temporal descriptor based on 3D-gradients", in BMVC, (2008).

[9] M. Marszalek, I. Laptev, C. Schmid, "Actions in context", in CVPR2009, pp.2929-2936, (2009).

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, "Actions as space-time shapes", ICCV2005, (2005).

[11] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: A local SVM approach", ICPR2004, (2004).

[12] http://www.cs.ucf.edu/vision/public_html/data.html

[13] World Health Organization (WHO), "The International Classification of Functioning, Disability and Health (ICF)", World Health Assembly, (2001).

[14] Y. Wang, G. Mori, "Learning a discriminative hidden part model for human action recognition",NIPS2008, (2008).