Evaluation of Vision-based Human Activity Recognition in Dense Trajectory Framework

Hirokatsu Kataoka[1], Yoshimitsu Aoki[2], Kenji Iwata[1], Yutaka Satoh[1]

[1]National Institute of Advanced Industrial Science and Technology (AIST)
[2]Keio University

Abstract. Activity recognition has been an active research topic in computer vision. Recently, the most successful approaches use dense trajectories that extract a large number of trajectories and features on the trajectories into a codeword. In this paper, we evaluate various features in the framework of dense trajectories on several types of datasets. We implement 13 features in total by including five different types of descriptor, namely motion-, shape-, texture- trajectory- and co-occurrence-based feature descriptors. The experimental results show a relationship between feature descriptors and performance rate at each dataset. Different scenes of traffic, surgery, daily living and sports are used to analyze the feature characteristics. Moreover, we test how much the performance rate of concatenated vectors depends on the type, top-ranked in experiment and all 13 feature descriptors on fine-grained datasets. Feature evaluation is beneficial not only in the activity recognition problem, but also in other domains in spatio-temporal recognition.

1 Introduction

Recently, activity recognition has become one of the most active topics in the field of computer vision. Since space-time interest points (STIP) [1] were proposed, many researchers have studied activity recognition. Several survey papers have been published in activity recognition such as Moeslund *et al.* [2] and Aggarwal *et al.* [3]. Moeslund *et al.* [2] introduced a large number of approaches, not only in activity recognition, but also in human detection and tracking in their paper, and Aggarwal *et al.* [3] listed several recognition styles such as single person's activity and interaction recognition.

In their study of activity representation, Wang *et al.* [4] evaluated several space-time features for activity recognition, e.g., STIP [5], cuboid [6], Hessian [7] and dense [4] features with more detailed experimental settings. This evaluation has led to the idea of dense trajectories (DT) [8], which outperform other space-time features. In follow-up work with improved dense trajectories (iDT) [9], they improved their idea by implementing estimating camera motion with speeded-up robust features (SURF) [10] and a homography matrix, human rectangles and Fisher vector [11]. The improvements induced outstanding performance rates such as UCF50 [12] (91.2%), and Hollywood2 [13] (64.3%). The current state-of-the-art approach on the side of accuracy is the combination of iDT and per

frame deep net features (6,7,8-layers) [14]. According to the THUMOS challenge, which consists of activity classification in a large-scale database [14], the iDT should be used to more completely understand all human activity, and not only deep net features.

Benenson *et al.* [15] cited and implemented over 40 approaches including various features and classifiers so as to detect a pedestrian in traffic scenes. The results of three familiar frameworks (random forests [16], deformable part model (DPM) [17] and deep learning [18]) are close if there are enough fine-tuned parameters. Thus, the comparison of various approaches will be a significant test to determine how much to change and how to apply the feature descriptors. In activity recognition, feature evaluation is important to gain knowledge of a more practical use of a space-time feature descriptor for activity recognition.

In this paper, we execute efficient evaluations with various dense trajectory-based feature descriptors on multiple types of datasets including traffic (NTSEL– self-collected), surgery (INRIA surgery [19]), daily living (MSR daily activity 3D [20]) and sports (UCF50 [12]) scenes. Moreover, the 13 features are assigned and divided into five feature properties: (i) trajectory (ii) shape (iii) motion (iv) texture and (v) co-occurrence. The performance rate of activity classification depends on the computational environment, i.e., activity codewords, trajectory patterns, classifier settings and cross-validation task. We furthermore evaluate various features in a fair experimental setting.

The rest of the paper is organized as follows. In the next section we describe the dense feature and 13 feature descriptors used in this paper. In section 3, we show the effectiveness of the 13 feature descriptors and their concatenated vectors in our experimental results by means of four datasets. Finally, in the last section we conclude the paper.

2 Feature evaluation strategy

Figure 1 shows the framework of the 13 feature descriptors in the dense trajectories framework. We applied 13 features– trajectory feature (traj.) [8], histograms of oriented gradients (HOG) [21], scale invariant feature transform (SIFT) [22], histograms of optical flow (HOF) [23], motion boundary histogram (MBHx & MBHy) [24], motion interchange patterns (MIP) [25], higher-order local auto correlation (HLAC) [26], local binary patterns (LBP) [27], improved LBP (iLBP) [28], local trinary patterns (LTP) [29], Co-occurrence HOG (CoHOG) [30], and Extended CoHOG (ECoHOG) [31] in this evaluation. We categorized the 13 features into five topics, namely: (i) trajectory– traj. (ii) shape– HOG and SIFT (iii) motion– HOF, MBH, and MIP (iv) texture– HLAC, LBP, iLBP and LTP, and (v) co-occurrence– CoHOG and ECoHOG. Moreover, the dense trajectories (DT) [8] + bag-of-words (BoW) model (not improved DT + Fisher vector [9]) was used in this evaluation because we evaluated the performance ability of the feature descriptors themselves. We used a support vector machine (SVM) as a multi-class classifier following [8].

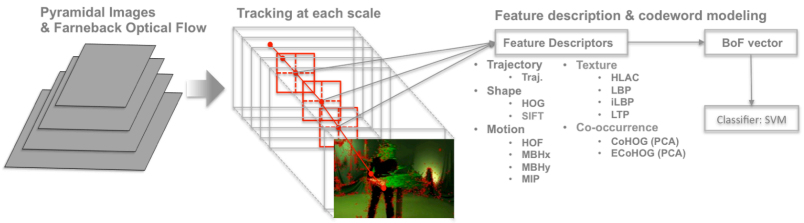


Fig. 1. The framework of the 13 feature descriptors in the dense feature framework.

2.1 Dense trajectories

We used Wang's dense trajectories [8] (DT) to create bag-of-words (BoW) vectors [32] for activity recognition. The idea of DT includes dense sampling and space-time feature extraction at the sampled points. Feature points at each grid cell are computed with Farneback optical flow. To take care of scale changes, the DT extracts dense flows in multiple image scales, where the image size increases by a scale factor $1/\sqrt{2}$. A large number of DT flows among the multiple scales are integrated into a feature vector based on BoW. This setup allows us to obtain a detailed motion at the specified patch. The length of the trajectory was set as 15 frames. Therefore, we recorded 0.5 seconds activities from a 30 fps video. Moreover, we set all BoW vectors to 4000 dimensions at each feature descriptor following [8].

2.2 Thirteen feature descriptors

Trajectory feature [8]. In activity analysis, the trajectory feature (traj.) [8] was extracted at each image patch. The size of the patch was 32×32 pixels, which is divided into 2×2 blocks. Here, the trajectory feature (T) is calculated as below:

$$T = \frac{(\delta P_t, ..., \delta P_{t+L-1})}{\Sigma_{j=t}^{t+L-1} ||P_j||} \tag{1}$$

$$\delta P_t = (\delta P_{t+1} - \delta P_t) = (x_{t+1} - x_t, y_{t+1} - y_t) \tag{2}$$

where L is the trajectory length. The feature represents a shape of connected optical flow.

Histograms of oriented gradients (HOG) [21]. HOG describes a feature vector with accumulating edge-magnitude into a quantized edge direction histogram. The process of feature extraction consists of edge calculation and normalization. Edge magnitude is accumulated into the quantized histogram by edge direction with $m(u,v) = \sqrt{f_u^2 + f_v^2}$ and $\theta(u,v) = arctan(f_v/f_u)$, where the magnitude and direction are $m(u,v)$ and $\theta(u,v)$, $f(x,y)$ is the differences between two pixels in the x and y directions. Feature extraction is executed

on overlapping blocks, and the feature vector is normalized every block with a norm.

Scale invariant feature transform (SIFT) [22]. This approach has characteristics of scale- and rotation-invariant features. SIFT contains keypoint detection and feature description; however, we mainly apply feature description to evaluate as a descriptor. To describe a feature vector, SIFT takes care of the image rotation by deciding a maximum direction. SIFT extracts 8 orientations divided into 4×4 blocks, giving 128 dimensions from an image patch.

Histograms of optical flow (HOF) [23]. The captured optical flows are quantized into nine directions. Wang *et al.* implemented HOF with a 4-divided image patch in his paper [8], therefore, a 36-dimension feature is extracted in an image patch. The feature represents normalized optical flow on a human motion area.

Motion boundary histograms (MBH) [24]. The motion boundary calculates the difference between two temporal frames. Therefore, it is less susceptible to capturing background noise when the camera motion is stable. Usually MBH features include the x- and y-directions together. However, we separate MBH into each direction MBHx and MBHy to analyze the properties of the feature descriptor at each scene.

Motion interchange patterns (MIP) [25]. This feature basically extracts a feature vector with trinary encoded pattern changes from a noticed area. Three temporal frames are applied to construct a motion interchange pattern.

Higher-order local auto correlation (HLAC) [26]. HLAC describes a feature by counting 25 significant mask patterns. The patterns indicate the displacement of a human in an image patch. The 25 pattern count allows us to capture a high-level movement, and we capture the patterns from the edge and the binarized image.

Local binary patterns (LBP) [27]. The process of LBP is constructed using a binarization step and an encoding step. In the binarization step, we process each 3×3 pixel patch to compare two pixels at the center of patch. The values are binarized with magnitude correlation in the patch. We capture eight binarized values, then the values are translated into 0 – 255 as a feature (this is an encoding step).

Improved LBP (iLBP) [28]. The basic idea of binarization is close to the normal LBP. The iLBP compares the eight nearest pixels with the averaged value of the nine pixels in a 3×3 patch. The feature emphasizes an edge element compared with LBP.

Local trinary patterns (LTP) [29]. The improved feature descriptor with trinary patterns has the same description as LBP. The feature instinctively captures by preparing an additional neutral class from two binarized classes. Because of the third class, it has more powerful representation as a texture-based descriptor.

CoHOG [30]. Co-occurrence Histogram of Oriented Gradient (CoHOG) is able to describe more complex shapes by pairing the brightness gradient directions in HOG. The brightness gradient direction of the pair is calculated using

the co-occurrence matrix. The co-occurrence matrix is calculated by counting the number of brightness gradient directions of the pair that are a target pixel, and the specific positional relationship from the target pixel in the block.

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} 1 \\ (if\ d(p,q) = i\ and \\ \ d(p+x,q+y) = j) \\ \\ 0 \\ (otherwise) \end{cases} \tag{3}$$

where $C(i,j)$ is the co-occurrence histogram that accumulates pairs of the pixel of interest and an objective pixel. Coordinates (p,q) indicate the pixel of interest (center of window) and coordinates $(p+x,p+y)$ indicate the objective pixel. m and n are the width and height of the feature extraction window. $d(p,q)$ is a function that quantizes the edge direction as an integer from 0 to 7 at pixel (p,q).

ECoHOG [31]. Extended Co-occurrence Histogram of Oriented Gradient (ECoHOG) enables a more efficient feature description by deleting the feature dimensions in CoHOG and extracting only the valid features. ECoHOG makes improvements over CoHOG via the accumulation of edge strength, the step acquisition of edge pairs and time series feature representation. We describe each of these processes below. CoHOG generates a histogram by counting the number of pairs of brightness gradients. However, ECoHOG describes not only the shape but also the intensity of light and shade and the condition of the change by accumulating edge intensity.

$$m_1(x_1,y_1) = \sqrt{f_{x1}(x_1,y_1)^2 + f_{y1}(x_1,y_1)^2} \tag{4}$$

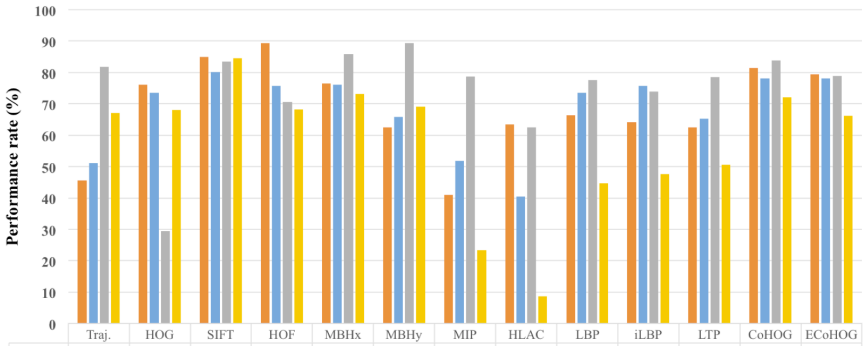
$$m_2(x_2,y_2) = \sqrt{f_{x2}(x_2,y_2)^2 + f_{y2}(x_2,y_2)^2} \tag{5}$$

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} m_1(x_1,y_1) + m_2(x_2,y_2) \\ (if\ d(p,q) = i\ and \\ \ d(p+x,q+y) = j) \\ \\ 0 \\ (otherwise) \end{cases} \tag{6}$$

$m_1(x,y)$ and $m_2(x,y)$ are the magnitudes of the pixel of interest (at the center of the window) and the magnitudes in the objective window, respectively.

3 Experiments

We carried out evaluations of feature descriptors in the framework of dense trajectories. Figure 2 shows the performance rates and Table 1 shows the top



	Traj.	HOG	SIFT	HOF	MBHx	MBHy	MIP	HLAC	LBP	iLBP	LTP	CoHOG	ECoHOG
NTSEL	45.7	76.1	85	89.4	76.5	62.5	41.1	63.5	66.3	64.1	62.5	81.4	79.5
INRIA	51.2	73.5	80.1	75.8	76.2	65.9	51.8	40.5	73.5	75.8	65.2	78.1	78.2
MSR	81.9	29.5	83.5	70.6	85.8	89.3	78.7	62.5	77.6	74	78.6	83.8	78.9
UCF50	67.2	68	84.5	68.2	73.1	69.2	23.4	8.7	44.7	47.7	50.5	72	66.2

Fig. 2. Overall rate of 13 features in the framework of dense trajectories [8] on the four datasets (NTSEL, INRIA surgery, MSR, and UCF50).

three features with dense trajectories on the four different types of datasets, namely in traffic (NTSEL traffic), surgery (INRIA surgery), daily living (MSR daily activity 3D) and sports (UCF50) scenes. Moreover, the classification with concatenated vectors is shown in Table 2. Here, Figure 3 shows examples of the datasets used in the experiments.

3.1 Results

NTSEL traffic dataset. We collected 100 videos with four pedestrian activities in traffic scenes (see Figure 3 top left). The activities include *walking*, *crossing*, *turning*, and *riding a bicycle*, where all of the activities indicate fine-grained pedestrian motion with three people. The dataset contains a cluttered background in small areas, making it difficult to capture optical flows. Presented activities are also fine grained as there are only small variations between *walking*, *crossing* and *turning* that is, they have a very few appearance and motion differences. We evaluated this dataset using 5-fold cross validation.

According to Table 1, HOF (89.4%), SIFT (85.0%) and CoHOG (81.4%) are the top three features in the NTSEL traffic dataset. The activities included in the NTSEL dataset are fine-grained walking activities. From these results, the optical flow vectorization (HOF) and detailed shape descriptors (SIFT, CoHOG) are effective for the classification of walking activities. HOF calculates 9 (dim.)×4 (blocks)×3 (frames) optical flow-based features, therefore it can significantly represent activities, e.g., *walking* and *crossing*, that show subtle differences when walking at a vertical or horizontal angle to a camera. The quantized optical flow vectors are able to classify fine-grained walking activities on the dataset. SIFT (4×4 block division) and CoHOG (co-occurrence feature with extraction

Fig. 3. Four different datasets: NTSEL traffic dataset (top left), INRIA surgery dataset (top right), MSR daily 3D activity dataset (middle), and a part of the UCF50 dataset (bottom)

window) are the detailed shape descriptors and extract temporal differences on a trajectory.

INRIA surgery dataset [19]. This dataset includes four activities performed by 10 different people with occlusions; e.g., people are occluded by a table or a chair (see Figure 3 top right). The activities include *cutting*, *hammering*, *repositioning*, and *sitting*. Each person performs the same activity twice, one for training and another for testing in this experiment.

The top three features are SIFT (80.1%), ECoHOG (78.2%) and CoHOG (78.2%) from Table 1. Near to the NTSEL traffic dataset, the INRIA surgery dataset contains fine-grained activities in experimental surgery scenes. The camera angle is fixed and the arm swing activities are confusing, so the detailed shape features such as SIFT, ECoHOG and CoHOG are important for surgery activity classification. The difference between ECoHOG and CoHOG is edge-magnitude extraction and edge-pair counting. In this situation, ECoHOG fetches co-occurrence features with magnitude accumulation, which is slightly better than CoHOG.

MSR daily activity 3D dataset [20]. The dataset is basically used as a depth-based activity recognition with a Kinect sensor. Depth and 3D-posture are given in this dataset, and at the same time we can access RGB videos (see Figure 3, middle). In this experiment, *only* RGB information is assigned as input data to calculate feature vectors. There are 16 activities in the dataset: *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*. The experiment is executed with leave-one-person-out cross-validation. Moreover, the trajectory length (L) is 50 because the normal setting of dense trajectories (15 frame accumulation) is extremely short so as to capture feature elements from an activity in a daily living activity.

MBHx (89.3%), MBHy (85.8%) and CoHOG (83.8%) are listed as the top three feature descriptors. The original paper on dense trajectories [8] reported that the MBH effectively extracts motions for human activity. The motion boundary is a well-classified feature that is not dependent on horizontal and

Table 1. The top three performance rates with dense trajectories on the four datasets.

Dataset	Outline	Top three features
NTSEL	Fine-grained pedestrian activities.	HOF (89.4%)
		SIFT (85.0%)
		CoHOG (81.4%)
INRIA	Fine-grained surgery activities.	SIFT (80.1%)
		ECoHOG (78.2%)
		CoHOG (78.1%)
MSR	Daily living activities.	MBHy (89.3%)
		MBHx (85.8%)
		CoHOG (83.8%)
UCF50	Large-scale sports activities.	SIFT (84.5%)
		MBHx (73.1%)
		CoHOG (72.0%)

vertical direction. Large variations exist in understanding activities of daily living. For example, *sit still* & *lie down on sofa* are stable activities, and *cheer up* & *walk* are whole body motions. Although the dense trajectories accumulate over 50 frames to calculate a feature vector on the dataset, both MBH-based features perform better than the other features. The motion boundary feature enables subtle motions such as *sit still* and *lie down on sofa* to be distinguished, not only whole body activities. We believe that CoHOG contains a similar process to MBH in terms of feature description. CoHOG also extracts a detailed feature from a human area; however, CoHOG captures a co-occurrence shape with edge-pair counting.

UCF50 dataset [12]. The UCF50 dataset comprises 1168 videos in 50 categories collected from YouTube (see Figure 3 bottom for examples of the dataset). 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 elements that computer vision has difficulty with, such as camera motion, complicated backgrounds, occlusions and personal variations. Performance rate is calculated with leave-one-group-out cross-validation in 25 groups for each activity.

Table 1 shows that SIFT (84.5%), MBHx (73.1%) and CoHOG (72.0%) are better descriptors than the other 10 features on the UCF50 dataset. SIFT is an advanced feature descriptor itself in recognition tasks. Here, UCF50 is made large-scale enough by including an element of object recognition, and therefore SIFT outperforms other features on the dataset. The MBHx achieved the second highest score because this feature is likely to be robust to background noise. Only motion boundary is recorded as a feature. The co-occurrence feature stably accomplishes a good result on activity recognition by combining dense feature representations. The relationship between CoHOG and ECoHOG is competitive; however, CoHOG generally outputs a better score.

3.2 Concatenated features for activity classification

The several-feature concatenation generally performs with a higher percentage on a vision-based classification. Here, we carried out the evaluations as to how to combine the 13 features based on the experiments in section 3.1. Table 2 describes how much the score changes with feature combinations. We prepared a variety of feature combination approaches: (i) original dense trajectories [8] (as baseline), (ii) five feature categories, in belief trajectory, shape, motion, texture and co-occurrence, and (iii) top-ranked shown in section 3.1 and all 13 feature combinations. We experimentally used two fine-grained datasets (NTSEL traffic and INRIA surgery dataset) to evaluate feature combination approaches.

Baseline. The original dense trajectories achieved an outstanding score as they are well-organized structures in activity classification. The four features (traj., HOG, HOF, MBH) are combined with kernel SVM for fine-grained recognition. The performance rates were 89.6% and 87.5% on the NTSEL traffic and INRIA surgery datasets, respectively.

Several types of feature. We categorized 13 features into five topics based on feature characteristics. The statements of several features are itemized here: trajectory– traj. feature, shape– HOG, SIFT, motion– HOF, MBHx, MBHy and MIP, texture– HLAC, LBP, iLBP and LTP, co-occurrence [31]– CoHOG, ECoHOG. The recognition scores were 45.7% (traj.), 85.2% (shape), 85.9% (motion), 60.7% (texture), 85.0% (co-occurrence) on the NTSEL, and 51.2% (traj.), 85.7% (shape), 82.5% (motion), 76.4% (texture) and 89.6% (co-occurrence) on the INRIA surgery dataset. The motion feature gave the best rate (85.9%), and the rates of the shape (85.2%) and co-occurrence (85.0%) features came next on the NTSEL traffic dataset. The motion feature included HOF and MBH, which are highly-accurate descriptors in the dataset. In the INRIA surgery dataset, the co-occurrence feature indicated a significant rate for classifying fine-grained activities with a precise description. The co-occurrence feature that contained CoHOG and ECoHOG showed stable vectorization in both datasets. The combination (co-occurrence feature & dense trajectories) of detailed description and dense representation is sophisticated in terms of fine-grained recognition.

Top-ranked and all 13 feature combinations. The top 3, 5, 7 and all 13 features should be combined to measure the ability of concatenated vectors. The scores of the concatenated vectors are in a narrow margin; however, the top seven (90.9%–HOF, SIFT, CoHOG, ECoHOG, MBHx, HOG and LBP) on the NTSEL traffic dataset and the top five (91.5%– SIFT, ECoHOG, CoHOG, MBHx and HOF) on the INRIA surgery dataset showed the best rates. The results demonstrate that all-feature concatenation does not always give the best accuracy on an activity recognition dataset. At this point, the 13-feature concatenation achieved 90.7% on both datasets. We believe that both datasets contain fine-grained activities, therefore the top-ranked features should be combined for high-accuracy fine-grained activity classification. Although the number of top-ranked features is an ad-hoc problem, the accuracy is easy to investigate in a classification experiment. In every case, we obtained a lower processing speed with a selected concatenated vector using dense trajectory-based feature extrac-

Table 2. The performance rate with concatenated features on the NTSEL traffic and INRIA surgery dataset.

Baseline	(%) on NTSEL	(%) on INRIA
Wang *et al.* [8]	89.6%	87.5%
Feature Type	(%) on NTSEL	(%) on INRIA
Trajectory	45.7%	51.2%
Shape	85.2%	85.7%
Motion	**85.9%**	82.5%
Texture	60.7%	76.4%
Co-occurrence [31]	85.0%	**89.6%**
Concatenated Vector	(%) on NTSEL	(%) on INRIA
Top 3 feature concatenation	89.6%	90.4%
Top 5 feature concatenation	90.2%	**91.5%**
Top 7 feature concatenation	**90.9%**	90.8%
All 13 features	90.7%	90.7%

tion and activity classification. We therefore improve both performance rate and processing speed by using a concatenated vector for fine-grained recognition.

4 Conclusion

In this paper, we evaluated 13 features in the framework of dense trajectories for more effective activity recognition. We carried out all experiments at fair settings in terms of activity codewords, captured trajectories, classifier settings and cross-validation. The four scenes are included in the experiments using traffic, surgery, daily living and sports datasets. The results describe the best performance rate at each dataset– HOF (89.4%) on the NTSEL traffic dataset, SIFT (80.1%) on the INRIA surgery dataset, MBHy (89.3%) on the MSR daily activity 3D dataset and SIFT (84.5%) on the UCF50 dataset. Detailed analysis indicated that the co-occurrence feature containing CoHOG and ECoHOG would be a stable descriptor in activity recognition. The co-occurrence feature becomes significant representation by using dense trajectories. In feature concatenation, we found that the combination of highly-selected features tends to give a better rate than the integration of all 13 features. Therefore, we experimentally chose sophisticated features with top-ranked listed features. Particularly in a fine-grained activity dataset, an extra feature descriptor should NOT be included for a fine-grained classification by means of only effective features. Moreover, selected feature concatenation allows us to improve both processing speed and accuracy.

For more detailed analysis, we visualize how to execute classification with various features. The subspace representation is one of the most important tasks in pattern recognition. Moreover, we try to correct for various durations of human activity.

Appendix:

This work was partially supported by JSPS KAKENHI Grant Number 24300078.

References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascaded of simple features, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2001)
2. Moeslund, T.B., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding (CVIU) (2006)
3. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review, ACM Computing Survey (2011)
4. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition to cite this version, British Machine Vision Conference (BMVC) (2009)
5. Laptev, I.: On space-time interest points, International Journal of Computer Vision (IJCV) (2005)
6. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features, Visual Surveillance and Performance Evaluation of Tracking and Surveillance (PETS) (2005) 65–72
7. Willems, G., Tuytelaars, T., V., G.L.: An efficient dense and scale-invariant spatio-temporal interest point detector, European Conference on Computer Vision (ECCV) (2008)
8. Wang, H., Klaser, A., Schmid, C.: Dense trajectories and motion boundary descriptors for action recognition, International Journal of Computer Vision (IJCV) (2013)
9. Wang, H., Schmid, C.: Action recognition with improved trajectories, IEEE International Conference on Computer Vision (ICCV) (2013)
10. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features, European Conference on Computer Vision (ECCV) (2006)
11. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification, European Conference on Computer Vision (ECCV) (2010)
12. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos, Machine Vision and Applications (MVA) (2012) 1–11
13. Marszalek, M., Laptev, I., Schmid, C.: Actions in context, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
14. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/ (2014)
15. Benenson, R., Omran, M., Hosang, J., Shiele, B.: Ten years of pedestrian detection, what have we learned?, European Conference on Computer Vision Workshop (ECCVW) (2014)
16. Breiman, L.: Random forests (2001)
17. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models (2010)
18. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)

19. Huang, C.H., Boyer, E., Navab, N., Ilic, S.: Human shape and pose tracking using keyframes, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
20. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
21. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005)
22. Lowe, D.G.: Distinctive image features from scale-invariant keypoints (2004)
23. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
24. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance, European Conference on Computer Vision (ECCV) (2006)
25. Kliper-Gross, O., Gurovich, Y., Hassner, T., Wolf, L.: Motion interchange patterns for action recognition in unconstrained videos, European Conference on Computer Vision (ECCV) (2012)
26. Kobayashi, T., Otsu, N.: Image feature extraction using gradient local autocorrelations, European Conference on Computer Vision (ECCV) (2008)
27. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution grayscale and rotation invariant texture classification with local binary patterns (2002)
28. Mohamed, A.A., Yampolskily, R.V.: An improved lbp algorithm for avatar face recognition, International Symposium on Information, Communication and Automation Technologies (ICAT) (2011)
29. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition, International Conference on Computer Vision (ICCV) (2009)
30. Watanabe, T., Ito, S., Yokoi, K.: Co-occurrence histograms of oriented gradients for pedestrian detection, PSIVT (2009)
31. Kataoka, H., Hashimoto, K., Iwata, K., Satoh, Y., Navab, N., Ilic, S., Aoki, Y.: Extended co-occurrence hog with dense trajectories for fine-grained activity recognition, Asian Conference on Computer Vision (ACCV) (2014)
32. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints, European Conference on Computer Vision Workshop (ECCVW) (2004)