Extended Co-occurrence HOG with Dense Trajectories for Fine-grained Activity Recognition

Hirokatsu Kataoka[1 2], Kiyoshi Hashimoto[2], Kenji Iwata[3], Yutaka Satoh[3], Nassir Navab[4], Slobodan Ilic[4], Yoshimitsu Aoki[2]

[1]The University of Tokyo [2]Keio University [3]National Institute of Advanced Industrial Science and Technology (AIST) [4]Technische Universität München (TUM)

Abstract. In this paper we propose a novel feature descriptor Extended Co-occurrence HOG (ECoHOG) and integrate it with dense point trajectories demonstrating its usefulness in fine grained activity recognition. This feature is inspired by original Co-occurrence HOG (CoHOG) that is based on histograms of occurrences of pairs of image gradients in the image. Instead relying only on pure histograms we introduce a sum of gradient magnitudes of co-occurring pairs of image gradients in the image. This results in giving the importance to the object boundaries and straightening the difference between the moving foreground and static background. We also couple ECoHOG with dense point trajectories extracted using optical flow from video sequences and demonstrate that they are extremely well suited for fine grained activity recognition. Using our feature we outperform state of the art methods in this task and provide extensive quantitative evaluation.

1 Introduction

In the past years various techniques for visual analysis of humans have been studied in the field of computer vision [1]. Human tracking, body pose estimation, activity recognition and face recognition are just some examples of analysis of humans from videos that are relevant in many real-life environments. Recently, human activity recognition has become a very active research topic and several survey papers have been published, including those by Aggarwal *et al.* [2], Moeslund *et al.* [3], and Ryoo *et al.* [4]. The number of applications is vast and they include, but are not limited to video surveillance, sports video analysis, medical science, robotics, video indexing, and games. To put these applications into practice, many activity recognition methods have been proposed in recent years to improve accuracy. Activity recognition means determining the activity of the person from a sequence of images. In case of very similar activities with the subtle differences in motion we talk about fine-grained activities. The classical recognition pipeline starts with extracting some kind of spatio temporal features and feeding them into the classifiers trained to recognize such activities. However, in case of fine-grained activities minor differences between extracted features

frequently affect the classification of an activity. This makes visual distinction difficult using existing feature descriptors. For fine-grained activity recognition, Rohrbach et al. [5] confirmed that dense sampling of feature descriptors achieved better results than joint features based on posture information.

In this paper we propose Extended Co-occurrence HOG (ECoHOG) feature and integrate it with dense sampling and dense feature extraction approach in order to improve accuracy of fine-grained activity recognition. We rely on Co-occurrence Histograms of Oriented Gradients (CoHOG) [6] as a feature descriptor representing co-occurrence elements in an image patch. The co-occurrence feature clearly extracts an object's shape by focusing on co-occurrence of image gradients at the pairs of image pixels and in that way reduces false positives. We extend this feature by adding sum of the magnitude of the gradients as co-occurrence elements. This results in giving the importance to the object boundaries and straightening the difference between the moving foreground and static background. In addition we apply this descriptor on the dense trajectories and test it for fine grained activity recognition.

We tested influence of our ECoHOG feature coupled with dense trajectories on two fine-grained activity recognition datasets: MPII cooking activities dataset [5] and INRIA surgery dataset [7] and obtained increase of performance using only this features in contrast to the use of HOG [16], HOF (Histograms of Optical Flow) [11] and MBF (Motion Boundary Histograms) [17] used in Wang et al. [8][9].

2 Related Work

A large amount of activity recognition research has been undertaken in the past decade. The first noteworthy work is Space-Time Interest Points (STIPs) [10]. The STIP algorithm is an improvement of Harris corner detector for x-y and time t space. STIPs are three dimensional descriptors representing motion of corner points in time. The spatio-temporal sampling and feature description framework is widely used by the activity recognition community. Klaser proposed 3D-HOG *et al.* [12], while Marszalek *et al.* [13] described feature combination using the STIP framework. Recently, Everts *et al.* proposed color STIPs with four different color spaces added to standard STIP descriptor [14].

However, up to date the best approach for activity recognition is arguably "dense trajectories" proposed by Wang *et al.* [8][9], which is a trajectory-based feature description on dense sampling feature points. Using these trajectories histograms of oriented gradients (HOG) [16], histograms of optical flow (HOF) [11], and motion boundary histograms (MBH) [17] can be acquired. Rohrbach *et al.* claimed that dense trajectories outperformed other approaches in terms of accuracy on the MPII cooking activities dataset, which is a fine-grained dataset with 65 activity classes. Dense trajectories are also superior to posture-based feature descriptors on the dataset [15]. Dense sampling approaches for activity recognition have also been proposed in [18], [19], [20], [21], [22]) after the introduction of the first dense trajectories [8]. Raptis *et al.* implemented a middle-level

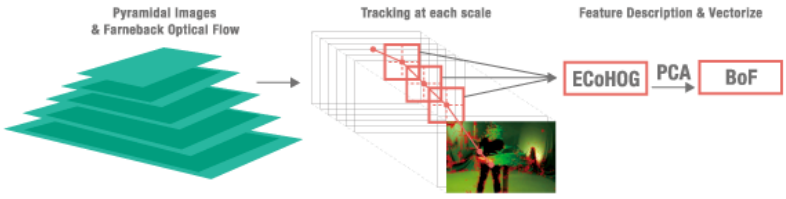

Fig. 1. Proposed framework: Pyramidal image capture and dense optical flow extraction are the same as the original dense trajectories. In feature extraction, we incorporate improved co-occurrence features (ECoHOG) in the dense trajectories. The co-occurrence vectors are reduced for effective vectorization as a bag-of-features (BoF) [23]. Finally, the co-occurrence vector is merged into other feature vectors (HOF, MBHx, MBHy, and Trajectory).

trajectory yielding simple posture representation with location clustering [18]. Li *et al.* translated a feature vector into another feature vector at a different angle using the "hankelet" approach [19]. To eliminate extra-flow, Jain *et al.* applied affine matrix [20] and Peng *et al.* proposed dense optical flow capture in the motion boundary space [21]. Wang *et al.* realized improved dense trajectories [22] by adding camera motion estimation, detection-based noise canceling, and a Fisher vector [24].

Several noise elimination approaches have been considered in this field to improve recognition performance, however, feature extraction is not enough. Thus, we introduced an improved feature descriptor into dense trajectories which we call Extended Co-occurrence HOG. This feature relies on the co-occurrence of the image gradients inside an image patch described with the normalized sum of the gradient intensities of co-occurring gradients. This helped distinguishing the moving foreground from the static background while capturing a subtle differences inside the image patches of the moving human body parts.

3 Proposed Framework

In this paper, we propose an improved co-occurrence feature ECoHOG and use it for fine-grained activity recognition. ECoHOG feature represents the gradient magnitude in a co-occurring image gradients located on the image patch and emphasize the boundary between the human and the background and between the edges within the human. Figure 1 shows the proposed framework applied in the context of fine-grained activity recognition. In essence, we have implemented the original dense trajectories [9] that find trajectories and extract features on the points along the trajectories. Using this framework and the concept of dense trajectories we integrated our improved co-occurrence feature (ECoHOG) into the HOF, MBH, and trajectory vectors. Finally we performed the dimensionality

reduction in order to convert the co-occurrence feature into a bag-of-features (BoF) vector [23].

The rest of the paper is organized as follows. In the next section we describe the extended co-occurrence feature descriptor and its vectorization using the BoF. In the next section we presents our experimental results using fine-grained activity datasets. Finally in the last section we conclude the paper.

4 Feature Description & Vectorization

4.1 Co-occurrence Histogram of Oriented Gradients (CoHOG)

HOG feature descriptor is calculated by computing the histogram of oriented gradients in the overlapping block inside the image patch. In practice gradient magnitude are accumulated in a corresponding orientation histogram and normalized within the block. Although the HOG can capture the rough shape of a human it often results in false positive detections in cluttered scenes when applied in tasks such as human detection. The Co-occurrence HOG (CoHOG) is designed to accumulate co-occurrences of pairs of image gradients inside the non-overlapping blocks of the image patch. Counting co-occurrences of the image gradients at different locations and in differently sized neighborhoods reduces false positives. For example, a pixel pair of the head and shoulders is described at the same time meaning that these two body parts, i.e. their edges, should always co-appear. As reported in [6] this proved to be more robust to the clutter and occlusions then standard HOG for human detection. In CoHOG eight gradient orientations are considered and co-occurrence of each orientation with each other orientation has been counted. This results in $8 \times 8 = 64$ dimensional histogram called co-occurrence matrix. In practice not only direct neighbors with offset one have been considered, but co-occurrence has also been regarded for larger offsets resulting in up to 30 co-occurrence histograms per image block. The co-occurrence histogram is computed as follows:

$$g(x, y) = \arctan \frac{f_y(x, y)}{f_x(x, y)} \tag{1}$$

$$f_x(x, y) = I(x + 1, y) - I(x - 1, y) \tag{2}$$

$$f_y(x, y) = I(x, y + 1) - I(x, y - 1) \tag{3}$$

$$C_{x,y}(i, j) = \sum_{p=1}^{n} \sum_{q=1}^{m} \begin{cases} 1, & \text{if } d(p, q) = i \text{ and } d(x + p, y + q) = j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $I(x, y)$ is the pixel value, $g(x, y)$ is the gradient orientation. $C(i, j)$ denotes the co-occurrence value of each element of the histogram, coordinates (p, q) depict the center of the feature extraction window, coordinates $(p+x, p+y)$ depict the position of the pixel pair in the feature extraction window, and $d(p, q)$ is one of eight the quantized gradient orientations.

The CoHOG can express the co-occurrence edge orientation acquired from two pixels and has higher accuracy than the HOG because of the co-occurrence edge representation. However, it faithfully counts all co-occurrence edges regardless of the edge magnitude. Human detection with a CoHOG results in false positives depending on the presence of an edge in a local image. Similar objects (e.g., trees and traffic signs) to a human have many elements whose histograms are similar. We believe that including the edge magnitude into the CoHoG is effective in creating a feature vector for human detection and therefore propose to extend CoHoG with magnitudes of the co-occurring gradients.

4.2 Extended Co-occurrence Histograms of Oriented Gradients (ECoHOG)

In this section, we explain the method for edge magnitude accumulation and histogram normalization, which we included in the ECoHOG. This improved feature descriptor is described below.

Accumulating Edge Magnitudes. Human shape can be described with the histograms of co-occurring gradient orientations. Here we add to it the magnitude of the image gradients which leads to improved and more robust description of the human shapes. In contrast to CoHOG, in our proposed framework we accumulate the sum of two pixel gradient magnitudes in the pairs of co-occurring pixel location inside the block of the image patch. The sum of edge magnitudes represents the accumulated gradient magnitude between two pixel edge magnitudes at different locations in the image block. In this way, for example, the difference between pedestrians and the background is more strengthened. The ECoHOG is defined as follows:

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} \|g_1(p,q)\| + \|g_2(p+x,q+y)\| \\ \text{if } d(p,q) = i \text{ and } d(p+x,q+y) = j \\ \\ 0 \;\; \text{otherwise} \end{cases} \tag{5}$$

where $\|g(p,q\|$ is the gradient magnitude, and $C(i,j)$, and all the other elements are defined as in Eqs. (1)–(3).

ECoHOG describes the magnitude for each pair of co-occurring pixel gradients and in that way creates a more robust co-occurrence histogram. It efficiently expresses the boundary between human and the background and also between the different textures of the clothing of the same human. Edge magnitude representation can define a boundary depending on the strength of the edges. This feature descriptor represents not only the combination of curves and straight lines, but also performs better than the CoHOG.

Histogram Normalization. The brightness of an image changes with respect to the light sources. The feature histogram should be normalized in order to

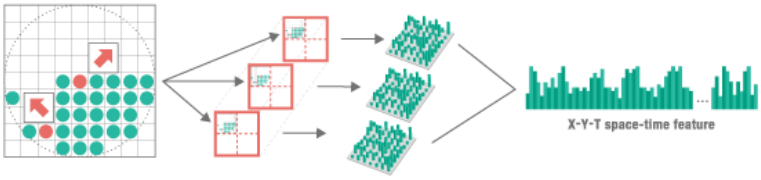

Fig. 2. Extended CoHOG in dense trajectories

be robust for human detection under various lighting conditions. The range of normalization is 64 dimensions, that is, the dimension of the co-occurrence histogram. The equation for normalization is given as:

$$C'_{x,y}(i,j) = \frac{C_{x,y}(i,j)}{\sum_{i'=1}^{8}\sum_{j'=1}^{8} C_{x,y}(i',j')}, \tag{6}$$

where C and C' denote histograms with and without normalization, respectively.

4.3 Extended CoHOG in Dense Trajectories

According to [9] the image is tessellated into a grid of 2×2 blocks and that small patch is tracked using optical flow in three consecutive frames. This results in dense trajectories of densely sampled image grid. [9] propose to compute multiple features like HOG, HOF and MBH in frames along the time trajectory and concatenate into one spatio-temporal histogram resulting in (X-Y-T) space–time block as shown in Fig. 2. Instead of computing HOG, HoF and MBH we compute our ECoHOG feature and concatenate the in time. Computation of co-occurrence description looks not in real-time, however, divided blocks can be calculated in parallel. Parallel processing allows us to calculate ECoHOG nearly efficient as HOG, HOF and MBH. Depending on the size of the neighborhoods different offsets are used to collect co-occurrence of the image gradients in ECoHOG. This results into a number of histograms per image block and continuous spatio-temporal feature has huge dimension.

Dimensionality reduction and Bag-of-Features. In order to bring it to the reasonable size and make it computationally tractable we perform PCA in order to reduce dimension of our features. A low-dimensional vector is generally easier to divide into a collection of classes, i.e. to cluster into bags of features (BoF). In related work, the CoHOG required about 35,000 dimensions for pedestrian detection [6]. However, we use a low-dimensional vector of 4000 dimensions to compose a BoF vector for activity recognition. In the experiments, we define and analyze an effective parameter for dimensionality reduction.

The BoF effectively represents a visual vector in an image [23]. An image generally consists of a large number of small patches called visual words. The BoF calculates the distribution to form feature vectors. Following the original dense trajectories, we randomly select millions of vectors from a dataset. The K-means clustering algorithm categorizes them into 4000 cluster, i.e. into 4000 visual words. The value of K is the dimension of our ECoHOG used in our experiments. We use the sum of the squared difference to calculate the nearest BoF vector f for each input feature in each frame.

So in training for each fine grained activity dense trajectories are described using ECoHOG whose dimensions are reduced using PCA and they are all clustered into 4000 visual words. In ECoHOG representation, the feature models weighting gradient magnitude and it effectively evaluate edge features in co-occurrence elements. The ECoHOG features vectorize almost the same BoF vectors if activity is in the same class. The statistical learning allows us to classify a large number of activity classes, in other words, the feature can be better approach in fine-grained activity categorization.

5 Experiments

We carried out experiments to validate influence of our ECoHOG feature in fine-grained activity recognition. In this section, we discuss the datasets, parameter selection, and comparison of the proposed approach with state-of-the-art methods. The classifier setting is based on the original dense trajectories [9].

5.1 Datasets

We used two different datasets for fine-grained categorization. Visual distinction is difficult because the categories are often subtly different in the feature space. Moreover, they are difficult to distinguish using current activity recognition approaches. The INRIA surgery dataset [7] and MPII cooking activities dataset [5] are discussed below.

INRIA surgery dataset [7]. This dataset includes four activities performed by 10 different people with occlusions; e.g., people are occluded by a table or chair (see Figure 3). The activities include cutting, hammering, repositioning, and sitting. Each person performed the same activity twice, one for training and another for testing in this experiment.

MPII cooking activities dataset [5]. This dataset contains 65 activities (see Table 2) performed by 12 participants. These activities can be broadly categorized into a few basic activities, such as seven ways of "cutting", five ways of "taking", and so on. In total, the dataset comprises 8 hours (881,755 frames) in 44 videos. Performance is evaluated by leave-one-person-out cross-validation.

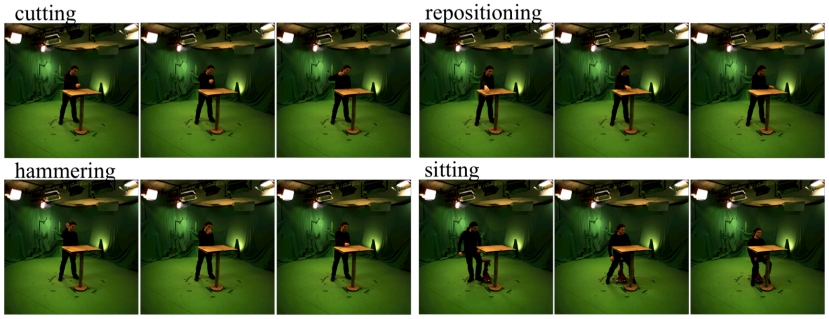

Fig. 3. INRIA surgery dataset [7], which includes four activities (cutting, hammering, repositioning, sitting).

5.2 Parameter selection

Figure 4 shows the relationship between the parameter settings and accuracy of the co-occurrence feature. In this situation, we must set the number of dimensions for PCA and offset size in the ECoHOG. Other settings for the dense trajectories and co-occurrence feature are based on [9] and [6], respectively. Here we focus on the seven "cutting" activities, which represent the most fine-grained category (Table 2 shows that activities 3 to 9 are the most confusing in terms of accurate recognition) in the MPII cooking activities dataset.

Figure 4(a) and (b) shows the number of dimensions and accuracy of the seven activities in the "cutting" category. The graph in Figure 4(a) shows that using a feature vector with 50 dimensions achieves the highest accuracy, and therefore, detailed results for 50 to 100 dimensions are depicted in Figure 4(b). From these results, we can judge the importance of balancing the "contribution ratio in PCA" and the "size of the feature space". As shown in Figure 4(b), 70 is the optimal value for creating the BoF vector in the ECoHOG feature.

Figure 4(c) shows the relationship between offset (feature extraction window) size and accuracy, with 5×5 being the optimal offset in this experiment. Most edge orientation pairs are extracted with a 1.0–3.0 pixel distance in commonly used edge orientation according to this figure. It is also important to consider "pixel similarity". Since near field pixels tend to have similar features, the feature vector should be designed to capture pixel similarity. Figure 5 shows the top 50 frequently used ECoHOG elements with an offset size of 11×11. According to the figure, neighboring pixels mostly support effective feature extraction. In the rest of the experiments we used 70 dimensional features and offset length of 5×5 pixel.

5.3 Comparison of Proposed and State-of-the-art Methods

In this section, we enumerate the experimental results on the INRIA surgery and MPII cooking activities datasets to compare the proposed approach with

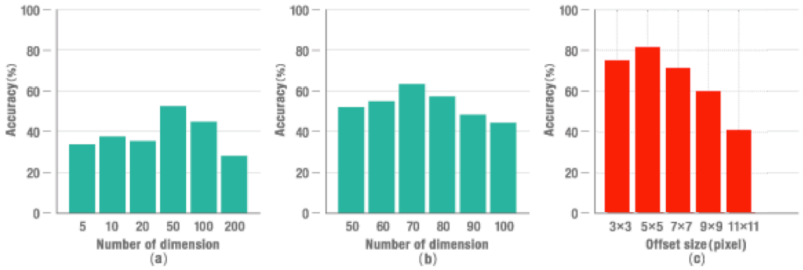

Fig. 4. Co-occurrence feature parameter settings for cutting activities (nos. 3–9 in Table 2): (a) number of dimensions for PCA; (b) detailed number of dimensions from 50 to 100 dimensions; (c) relationship between offset size and accuracy.

Table 1. Accuracy of the proposed and conventional frameworks on the INRIA surgery dataset.

Approach	Accuracy (%)
Tracking + HOG	40.16
Original Dense Trajectories (HOG, HoF, MBH, Trajectory)	93.58
CoHOG in Dense Trajectories (CoHOG)	81.05
ECoHOG in Dense Trajectories (ECoHOG)	96.36
Improved Dense Trajectories (ECoHOG, HOF, MBH, Trajectory)	97.31

state-of-the-art methods. In other words, we apply the original dense trajectories, CoHOG / ECoHOG in dense trajectories, and an integrated approach (ECoHOG, HOF, MBH, Trajectory). Simple HOG is consist of any tracking method and HOG description. We track a human and extract HOG feature in a tracked bounding box.

Experiment on INRIA surgery dataset. Table 1 shows the classification results on the INRIA surgery dataset. The original dense trajectories [9] achieved 93.58% accuracy thanks to dense sampling and the use of multi-type feature descriptions, whereas our proposed approach achieved 96.36%, applying only ECoHOG on the dense feature extraction on densely sampled trajectories. However, the integrated approach achieved a better result than the other two approaches (97.31% accuracy). ECoHOG improves CoHOG by including edge-magnitude accumulation, which expresses co-occurrence strength. ECoHOG comprehensively evaluates edge features and effectively generates more distinguishable BoF features in an image patch. ECoHOG represents the edge-boundary, which can

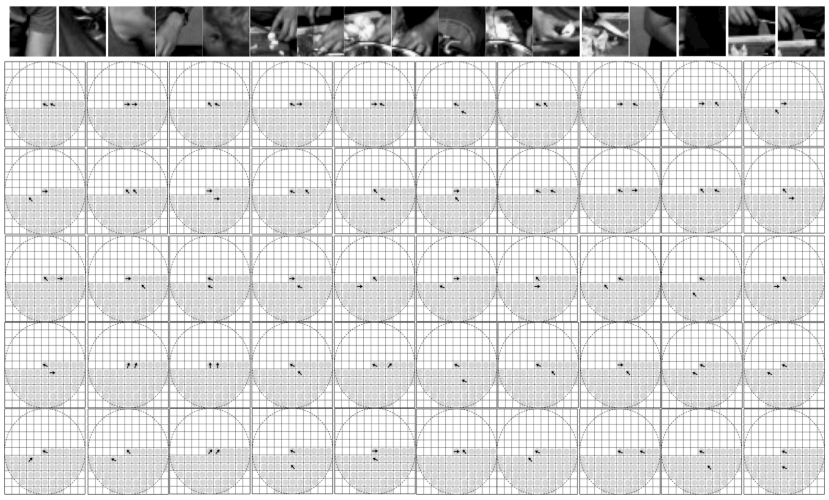

Fig. 5. (Top row) image patches on dense trajectories from MPII cooking dataset. (Remaining rows: left to right and top to bottom in decreasing order) top 50 frequently used offsets and orientations with 11 × 11 pixel offset.

effectively be added to the co-occurrence feature in fine-grained categorization. Overall, ECoHOG performed 15.31% better than CoHOG in the framework with dense trajectories, and 2.78% better than the original dense trajectories. The integrated approach performed slightly better (0.95%) than ECoHOG in the dense trajectories. In this context, other features, such as HOF, MBH, and Trajectory, supplement the ECoHOG feature in representing spatial and temporal image features. ECoHOG mainly represents the shape feature, while the other features capture motion features in the video sequences.

Experiment on MPII cooking activities dataset. Table 2 shows the results on the MPII cooking activities dataset. The dataset contains 65 classes of cooking activities for measuring fine-grained activity recognition performance. The results for method 2, original dense trajectories, in Table 2 were taken from [5] as the baseline method. The original dense trajectories achieved 44.2% accuracy (without pose feature combination) on the dataset using HOG, HOF, MBH, and Trajectory features. ECoHOG in dense trajectories (method 4 in Table 2) achieved 46.6% better accuracy than the baseline. At the same time, CoHOG in dense trajectories (method 3 in Table 2) is superior to the baseline. According to these results, the co-occurrence feature effectively represents detailed space–time shapes in fine-grained activities. The combined method (method 5 in Table 2) achieved 49.1% better accuracy than the other approaches with all types of features (ECoHOG, HOF, MBH, Trajectory). According to these results, HOF/MBH/trajectory features complementarily extract image features from the

ECoHOG feature, which contains co-occurrence orientation and edge magnitude information on the edge extraction window. The co-occurrence feature mainly captures shape information; however, a detailed configuration is expressed from image patches on trajectories. On the other hand, the HOF/MBH/Trajectory features handle motion features between frames, which aid activity recognition in the fine-grained dataset. The combined approach achieves slightly better accuracy than ECoHOG in dense trajectories. In this experiment, performance of ECoHOG was 0.4% better than CoHOG and 2.4% better than the original dense trajectories. The combined model with ECoHOG/HOF/MBH/Trajectory features represents image features comprehensively and achieves 4.9% better accuracy than the baseline method.

The similarity of feature histogram is directly linked to significant BoF vector. We evaluate the similarities of CoHOG and ECoHOG feature as co-occurrence representation. Figure 6 shows the histograms of self-similarity on image patches (similar to Figure 5) from the MPII cooking activities dataset. In this case, 1,000 image patches were selected to calculate histogram similarity, giving the number of combinations ${}_nC_k$, $n = 1000$, $k = 2$ $(= 499500)$. We used the Bhattacharyya coefficient [25] to calculate histogram similarity:

$$S = \sum_{u=1}^{m} \sqrt{h_u^1 h_u^2} \tag{7}$$

where S is the similarity value $(0 \leq S \leq 1)$, h^1 and h^2 are feature vectors normalized as $\sum_{u=1}^{m} h_u^1 = \sum_{u=1}^{m} h_u^2 = 1.0$, and m denotes the number of histogram bins. The graphs show that ECoHOG has higher self-similarity scores; that is, ECoHOG tends to evaluate similar features and creates better BoF vectors.

6 Conclusion

In this paper, we proposed an improved co-occurrence feature in dense trajectories. The proposed approach, which achieves 96.36% accuracy on the INRIA surgery dataset and 46.6% accuracy on the MPII cooking activities dataset, is superior to state-of-the-art methods. The co-occurrence feature represents detailed shapes on temporally dense sampling points. Comparing ECoHOG with CoHOG, the magnitude accumulation yields the boundary division between objective motion and the background by magnitude weighting. We found that an integrated approach (ECoHOG, HOF, MBH, Trajectory) achieves better accuracy 97.1% and 49.1%, respectively. These values are 3.73% and 4.9% better than the proposed approach on the INRIA surgery dataset and MPII cooking activities dataset, respectively.

We also investigated the parameter settings in the video datasets, that is, offset length and number of dimensions in PCA of the co-occurrence feature. Optimal parameter values for creating the BoF vector in the co-occurrence feature are 5×5 (pixel) offset length and 70 PCA dimensions. "Pixel similarity", which is the co-occurrence pairs are extracted from neighbor area must be considered in the offset length to adjust the dimension size. Moreover, the PCA

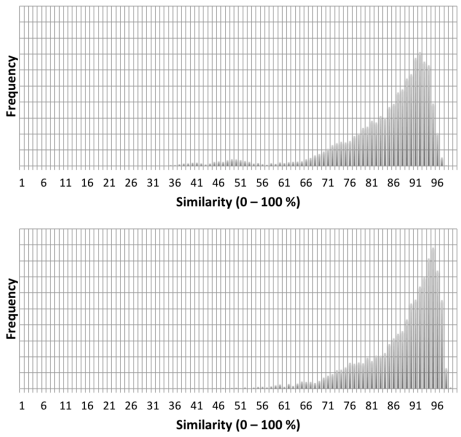

Fig. 6. Self-similarity of CoHOG (top) and ECoHOG (bottom) for 1,000 randomly chosen image patches from MPII cooking activities dataset: ${}_nC_k$ self-similarities in the graph ($n = 1000$, $k = 2$), the vertical axis denotes frequency and the horizontal axis gives the percentage similarity.

dimension should balance the "contribution ratio in PCA" and "size of the feature space" for the BoF vector. Given the above, we experimentally chose 70 dimensions from the original 640 ECoHOG dimensions.

References

1. T. B. Moeslund, A. Hilton, V. Kruger, L. Sigal, "Visual Analysis of Humans: Looking at People", Springer, 2011.
2. J. K. Aggarwal, Q. Cai, "Human Motion Analysis: A Review", Computer Vision and Image Understanding (CVIU), vol.73-3, pp.428-440, 1999.
3. 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.
4. M. S. Ryoo, J. K. Aggarwal, "Human activity analysis: A review", ACM Computing Surveys (CSUR), vol.43-3, 2011.
5. M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, "A database for fine grained activity detection of cooking activities", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
6. T. Watanabe, S. Ito, K. Yokoi, "Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection", Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp.37-47, 2009.
7. C.-H. Huang, E. Boyer, N. Navab, S. Ilic, "Human Shape and Pose Tracking Using Keyframes", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

8. H. Wang, A. Klaser, C. Schmid, C. L. Liu, "Action Recognition by Dense Trajectories", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3169-3176, 2011.
9. H. Wang, A. Klaser, C. Schmid, C. L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", International Journal of Computer Vision (IJCV), Vol.103, pp.60-79, 2013.
10. I. Laptev, "On Space-Time Interest Points", International Journal of Computer Vision (IJCV), No.64, pp.107-123, 2005.
11. I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1-8, 2008.
12. A. Klaser, M. Marszalek, C. Schmid, "A spatio-temporal descriptor based on 3D-gradients", British Machine Vision Conference (BMVC), 2008.
13. M. Marszalek, I. Laptev, C. Schmid, "Actions in context",IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2929-2936, 2009.
14. I. Everts, J. C. Gemert, T. Gevers, "Evaluation of Color STIPs for Human Activity Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2850-2857, 2013.
15. A. Zinnen, U. Blanke, and B. Schiele, "An analysis of sensor-oriented vs. model - based activity recognition", IEEE International Symposium on Wearable Computers (ISWC), 2009.
16. N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", IEEE Conference on Computer Vision and Pattern Recognitino (CVPR), pp.886-893, 2005.
17. N. Dalal, B. Triggs, C. Schmid, "Human Detection using Oriented Histograms of Flow and Appearance", European Conference on Computer Vision (ECCV), pp.428-441, 2006.
18. M. Raptis, I. Kokkinos, S. Soatto, "Discovering discriminative action parts from mid-level video representation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1242-1249, 2013.
19. B. Li, O. Camps, M. Sznaier, "Cross-view Activity Recognition using Hankelets", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1362-1369, 2012.
20. M. Jain, H. Jegou, P. Bouthemy, "Better exploiting motion for better action recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2555-2562, 2013.
21. X. Peng, Y. Qiao, Q. Peng, X. Qi, "Exploring Motion Boundary based Sampling and Spatial Temporal Context Descriptors for Action Recognition", British Machine Vision Conference (BMVC), 2013.
22. H. Wang, C. Schmid, "Action Recognition with Improved Trajectories", International Conference on Computer Vision (ICCV), pp.3551-3558, 2013.
23. G. Csurka, C. Bray, C. Dance, L. Fan, "Visual Categorization with Bags of Keypoints", European Conference on Computer Vision (ECCV) Workshop on Statistical Learning in Computer Vision, pp.59-74, 2004.
24. F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-scale image classification", European Conference on Computer Vision (ECCV), pp.143-156, 2010.
25. A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions", Bulletin of the Calcutta Mathematical Society, Vol.35, pp.99-109, 1943.

Table 2. Accuracy of the proposed and conventional frameworks on the MPII cooking activities dataset. Activity tags include 1.Background activity; 2.Change temperature; 3.Cut apart; 4.Cut dice; 5.Cut; 6.Cut off ends; 7.Cut out inside; 8.Cut slices; 9.Cut strips; 10.Dry; 11.Fill water from tap; 12.Grate; 13.Lid: put on; 14.Lid: remove; 15.Mix; 16.Move from X to Y; 17.Open egg; 18.Open tin; 19.Open/close cupboard; 20.Open/close drawer; 21.Open/close fridge; 22.Open/close oven; 23.Package X; 24.Peel; 25.Plug in/out; 26.Pour; 27.Pull out; 28.Puree; 29.Put in bowl; 30.Put in pan/pot; 31.Put on bread/dough; 32.Put on cutting board; 33.Put on plate; 34.Read; 35.Remove from package; 36.Rip open; 37.Scratch off; 38.Screw closed; 39.Screw open; 40.Shake; 41.Smell; 42.Spice; 43.Spread; 44.Squeeze; 45.Stamp; 46.Stir; 47.Sprinkle; 48.Take & put in cupboard; 49.Take & put in drawer; 50.Take & put in fridge; 51.Take & put in oven; 52.Take & put in spice holder; 53.Take ingredient apart; 54.Take out of cupboard; 55.Take out of drawer; 56.Take out of oven; 57.Take out of oven; 58.Take out of spice holder; 59.Taste; 60.Throw in garbage; 61.Unroll dough; 62.Wash hands; 63.Wash objects; 64.Whisk; 65.Wipe clean.: (1) original dense trajectories [5] (**44.2%**), (2) CoHOG in dense trajectories (**46.2%**), (3) ECoHOG in dense trajectories (**46.6%**), (4) combined model in dense trajectories with ECoHOG, HOF, MBH, and Trajectory (**49.1%**). The tracking + HOG model recorded **18.2%** on the MPII cooking activities dataset.

Activity Number	(1)	(2)	(3)	(4)	Activity Number	(1)	(2)	(3)	(4)
1	47.1	17.8	83.6	55.0	34	34.5	11.7	54.1	5.8
2	37.6	12.3	21.9	14.5	35	39.1	69.2	72.7	78.7
3	16.0	68.5	17.0	13.0	36	5.8	16.8	27.2	29.5
4	25.1	6.8	84.5	50.1	37	3.8	63.5	66.6	72.2
5	22.8	36.8	40.4	43.8	38	36.3	36.8	27.7	30.0
6	7.4	2.0	8.5	9.2	39	19.1	51.4	22.9	24.9
7	16.3	0.0	0.0	29.0	40	33.5	16.6	72.7	78.7
8	42.0	46.0	24.1	26.2	41	24.8	37.4	39.2	42.5
9	27.6	37.3	46.3	50.1	42	29.3	28.7	32.2	34.9
10	95.5	69.2	43.9	47.5	43	11.2	5.2	25.7	27.8
11	75.0	16.9	17.1	18.5	44	90.0	10.3	72.7	78.7
12	32.9	34.6	24.6	26.6	45	73.3	52.2	25.8	28.0
13	2.0	19.6	45.0	48.8	46	50.0	69.2	52.4	56.8
14	1.9	5.0	3.2	3.4	47	39.6	38.8	37.9	74.8
15	36.8	94.9	68.7	74.4	48	37.2	69.2	72.7	78.7
16	15.9	42.9	72.7	78.7	49	37.6	8.7	3.9	4.2
17	45.2	27.1	26.1	28.3	50	54.6	0.0	75.1	81.3
18	79.5	69.2	72.7	78.7	51	100	55.8	72.7	78.7
19	54.0	32.9	64.0	69.3	52	80.2	4.9	11.7	12.7
20	38.1	42.2	8.9	9.6	53	17.5	33.3	22.3	24.2
21	73.7	69.2	72.7	78.7	54	81.5	24.1	25.5	27.7
22	25.0	26.3	38.2	41.4	55	79.7	69.2	72.5	78.5
23	31.9	69.2	72.7	78.7	56	73.6	48.8	27.2	29.5
24	65.2	45.6	48.3	52.4	57	83.3	45.4	27.4	29.7
25	54.7	12.0	37.1	40.2	58	67.0	42.3	50.4	54.6
26	54.2	12.5	17.1	11.9	59	18.2	30.1	2.1	22.8
27	87.5	69.2	34.2	37.1	60	84.4	4.0	42.7	46.3
28	67.1	11.5	12.9	14.0	61	100	69.2	8.3	9.0
29	18.8	69.2	72.7	78.7	62	45.9	52.0	54.4	59.0
30	15.3	29.8	12.6	13.7	63	67.1	15.1	39.7	43.0
31	42.1	18.8	5.4	5.9	64	70.0	13.9	8.8	9.6
32	7.1	9.5	72.7	78.7	65	10.6	63.1	33.4	36.2
33	11.0	36.9	45.7	49.5	**Mean**	**44.2**	**46.2**	**46.6**	**49.1**