Efficient Action Recognition based on Salient Object Detection

Hnin Mya Aye, Sai Maung Maung Zaw
University of Computer Studies, Mandalay
hninmyaaye26@gmail.com, saisaimmz@gmail.com

Abstract

Action recognition has become an important research topic in the computer vision area. This paper presents an efficient action recognition approach based on salient object detection. Recently, many features were directly extracted from video frames; as a result, unsatisfying results were produced due to intrinsic textural difference between foreground and background. Instead of whole frames, processing only on salient objects suppresses the interference of background pixels and also makes the algorithm to be more efficient. So, the main contribution of this paper is to focus on salient object detection to reflect textural difference. Firstly, salient foreground objects are detected in video frames and only interest features for such objects are detected. Secondly, we extract features using SURF feature detector and HOG feature descriptor. Finally, we use KNN classifier for achieving better action recognition accuracy. Experiments performed on UCF-Sports action dataset show that our proposed approach outperforms state-of-the-art action recognition methods.

1. Introduction

Action recognition is a fundamental task for many problems in the field of computer vision such as video surveillance, video retrieval, and human-computer interaction. Although a great deal of progress has been made in the recognition of human actions, it still remains a challenging task due to intra-class variations, inter-class similarities, background clutter, occlusions, high dimensionality and low quality of video data, and other fundamental difficulties [25].

The efficient video representation is mainly crucial part in action recognition. Actually, the features should be robustness to small variations in appearance, background, and viewpoint and action execution. In global representation, a preprocessing step is needed to mark the action region or segment the intended foreground object from the background. The common global representations are in the form of optical flow silhouettes or edges [4]. They are sensitive to partial occlusion and viewpoint variations. In local representation, the observation is described as a collection of independent patches. Compared with the global representation, local features are somewhat invariant to changes in viewpoint, person appearance and partial occlusions. Due to their advantage, local spatial-temporal features based on interest points are more and more popular in action recognition [15, 5].

Therefore, extracting informative and discriminative features from video frames has become an important issue in action recognition. Various successful methods based on local representations describing characteristics of local regions, and global representations describing video frame characteristics have been proposed to improve the accuracy.

The goal of this paper is to introduce more efficient action recognition approach by using a combination of global and local video representations. The main contribution is to estimate the intended foreground object by salient object detection, and only keep interest points on salient foreground objects in processing of action recognition. After that, local features are extracted by using SURF Detector and HOG feature descriptor that can yield local features invariant to geometric and photometric transformations.

The remaining sections of the paper are organized as follows. Section 2 commences related research work in the area of human action recognition by briefly reviewing the most relevant literature. Section 3 explains the detail of the main structure for action recognition. Section 4 discusses experimental results and finally, conclusions are drawn in Section 5.

2. Related Work

Various action recognition approaches have been proposed and these approaches showed the significant progress towards action recognition in realistic and challenging videos. Shape-based approaches build action representation models, shape contexts [20], motion history images (MHIs) [1] and space-time shapes [12] to recognize actions. Optical-flow approaches represent actions as histograms of optical flow by calculating the optical flow that encode the energy of the action. The representation based on local feature descriptions is more informative than the other approaches.

Laptev [13] introduced local features by outstretching the Harris detector for a video. Other approaches are based on the Gabor filter [15], the Hessian matrix [5], and the dense sampling [6], and so on. Laptev et al. [10] proposed the combined HOG/HOF feature descriptor. The former descriptor represents appearance and the second one represents a local motion by calculating optical flow [14]. Scovanner et al. [17] also proposed the spatiotemporal domain based SIFT descriptors which are invariant to changes of scale and rotation, and robust to noise. Willems et al. [5] proposed the extended SURF (ESURF) descriptor serving invariant of changing scales and orientations.

Various kinds of trackers have also been introduced in the action recognition tasks recently, such as the KLT tracker [16, 19], the SIFT tracker [11], and the dense sampling tracker [6]. Wang et al. [7] discussed the evaluation on these three trackers and proved that the dense sampling gives the best performance for action recognition task.

Oikonomopouls et al. [2] adapted the idea of saliency region selection in spatial images to spatiotemporal video space. Salient points are detected by measuring changes in the information content of the set of pixels in cylindrical spatiotemporal neighborhoods at different scales. Ashwan Abdulmunem et al. [3] introduced an approach considering saliency guided feature. With saliency as guidance, they extracted local and global features to encode video information.

3. Main Structure of the Action Recognition System

The human action recognition system mainly consists of three steps. The first step is salient object detection, in which the salient foreground objects are detected and only interest points on the detected objects are used. Applying salient object detection makes reducing the number of feature descriptors, suppressing the background interference and also helps making the method more robust to background fluctuations, while at the same time reduces the running time. The second step is feature extraction to encode video information. Finally, KNN classifier is used to achieve action recognition. In this section, we will explain detailed descriptions of each step. Figure 1 shows the main structure of the action recognition system.

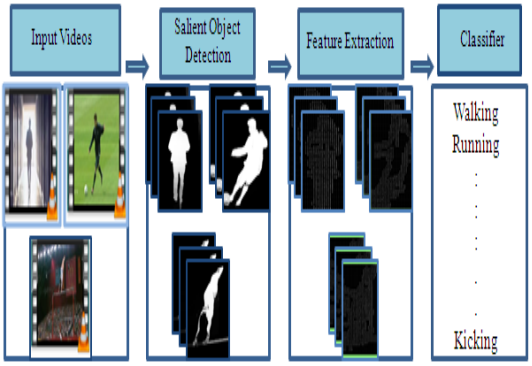


Figure 1. Main Structure of action recognition system (UCF-sports dataset)

3.1 Salient Object Detection

The first step is to detect salient foreground objects in video frames. For salient object detection, we use inner and inter label propagation based detection algorithm proposed by Hongyang Li et al., [9]. This algorithm estimates saliency in an image by propagating the labels extracted from the most certain background and object regions. To estimate the background appearance, the boundary cues are used because they are good indicators to distinguish salient objects from the background. The objectness cues are also used to emphasize on the salient object characteristics.

The affinity matrix construction is vital importance in the label propagation. It is constructed among superpixels by calculating the similarity of two image regions called superpixels (generated by SLIC algorithm [18]). The similarity is measured by a defined distance of the mean features in each region. The affinity entry wij of superpixel i (image region i) to a certain node j is defined as:

$$w_{ij} = \begin{cases} \exp\left(-\frac{D(f_i, f_j)}{\sigma^2}\right) & j = N(i) \\ 0 & i = j \text{ or otherwise} \end{cases} \quad (1)$$

where fi,fj denotes the mean feature vectors of pixels inside node i,j, σ is a turning parameter to control strength of the similarity, N(i) indicates the set of the direct neighboring nodes of superpixel i. A degree matrix D = diag {d1,...,dN} where di = Σjwij

is sum of the total entries of each node to other nodes. As an affinity matrix, the information of the background labels is propagated to estimate saliency measure of other superpixels. Given a dataset R = {r1, . . . rl, r l+1, . . . , rN} ∈ R D×N , where the former l regions serve as query labels and D denotes the feature dimension, a function V=[V(r1), … , V(rN)]T indicates the possibility of how similar each data point is to the labels. The similarity measure V(ri) satisfies

$$V_{t+1}(r_i) = \sum_{J=1}^{N} a_{ij} V_t(r_j) \qquad (2)$$

where aij is the affinity entry and t is the recursion step. For a given region, the similarity [23] is learned iteratively via propagation of the similarity measures of its neighbors such that a region's final similarity to the labels is effectively influenced by the features of its surroundings.

Figure 2 shows how the inner propagation algorithm works. In which, one superpixel region 6 is investigated and it can be seen that how its value V(r) changes during each iteration. It is assumed that there has 10 regions and the dash-outline regions (3, 6, 8) are not neighbors of region 6 and thus they are not considered in the propagation. The outline weight of each circle indicates the affinity weight. The red area inside each circle denotes the value of V(r). Since region 2, 4, 7 are assumed as background labels, they have red color fully filled within their circles in each iteration.

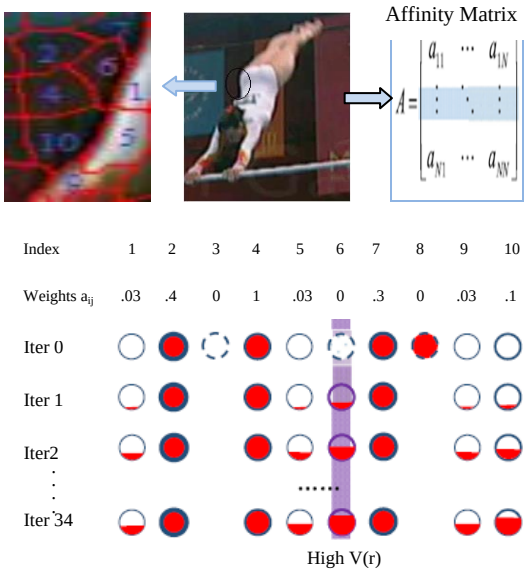


Figure 2. An example to illustrate how the inner propagation algorithm works

In some cases, the inner propagation via boundary labels alone has better results than a fusion of boundary and objectness labels due to the slight disturbance of objectness measures near the salient object. So, a compactness score is evaluated to determine the quality of the regional saliency map. Only the lower saliency maps score lower than a compactness criterion will be updated by the inter propagation via a co-transduction algorithm. Thus, to ensure high quality of the saliency maps and improve the computational efficiency, a co-transduction algorithm is devised to fuse both boundary and objectness labels based on an inter propagation scheme. The inter propagation algorithm can distinguish the foreground better from the background by enlarging the set of boundary labels from objectness cues. Figure 3 shows results of salient object detection algorithm.

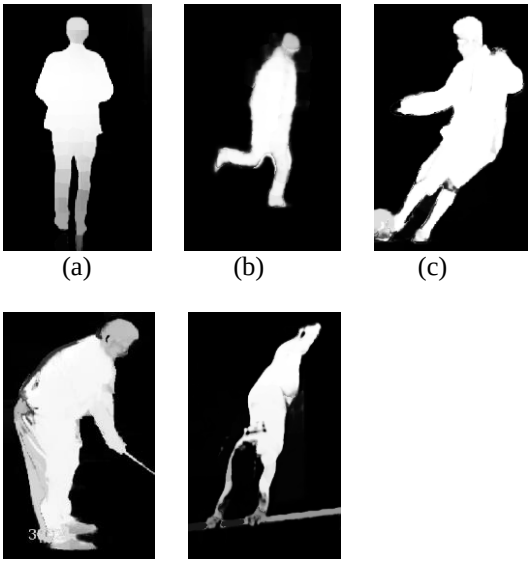


Figure 3. Results of salient object detection: (a) walking, (b) running, (c) kicking, (d) golf-swing, (e) swing-sideangle

3.2 Feature Extraction

In the second step of feature extraction, we need to perform two phases: feature detection and feature description. The role of feature detectors is locating the stable feature points in the spatio-temporal space by maximizing specific saliency functions [22]. The interested information relates to image regions which exhibit certain properties or some specific patterns. These patterns could be edges, corners, blobs, contours of objects, different kinds of junctions and many more things. The

collection of all these image patterns are labeled as image features or simply features.

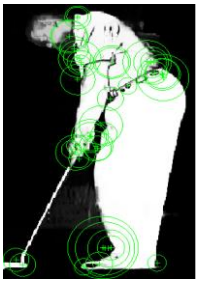
To detect features, we apply speeded up robust features (SURF) feature detector proposed by Herbert Bay (2006). It uses square-shaped filters as an approximation of Gaussian smoothing. It is very fast because of using an integral image in which the value of a pixel (x,y) is the sum of all values in the rectangle defined by the origin and (x,y) [26]. It also uses the Hessian matrix determinant as a measure of local change around the point and for selecting the scale. Given a point x = (x, y) in an image I, the Hessian matrix H (x, σ) in at scale σ is defined as follows :

$$H(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix} \quad (3)$$

where Lxx (x, σ) is the convolution of the Gaussian second order derivative with the image I in point x, and similarly for Lxy (x, σ) and Lyy (x, σ). Choosing scale spaces in SURF is implemented in image pyramids by applying box filters of different sizes. It is not need to iteratively apply the same filter to the output of a previously filter layer. Instead, such filter of any size can be applied at exactly the same speed directly on the original image. Therefore, the scale space is analyzed by up-scaling the filter size rather than iteratively reducing the image size [8].

In detecting features, we process only the interest feature points detected on salient objects, which carry robust information of an action. Consequently, the salient feature points are more precise and maximize the discriminative information of actions. Figure 4 shows the difference between the points of interest detected with and without salient object detection for UCF-sports action dataset (eg. Golf-swing).

(a)



(b)

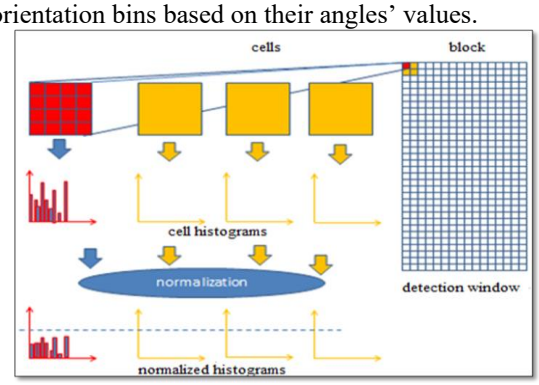
Figure 4. Points of Feature detection: (a) original frame without salient object detection (b) frame with salient object detection

In the feature description phase, we use Histogram of Oriented Gradients (HOG) proposed by Navneet Dalal and Bill Triggs (2005), which is a feature descriptor counting occurences of gradient orientation of pixels in overlapping windows of an image. It is invariant to geometric and photometric transformations.

It is computed through several steps. The gradients of an image are computed by filtering this image with horizontal kernel [–1, 0, 1] and vertical kernel [–1, 0, 1]−1. Then, magnitudes and angles are computed based on the computed gradients. Next, the image is separated into N×N overlapping windows.

Figure 5. HOG algorithm implementation scheme

For each window, angles are binned into B orientation bins based on their angles' values.



For each bin, sum of gradient magnitudes is calculated. After that, these sums, which are equal to the number of bins for each window, are normalized. At the end, N×N×B normalized numbers are obtained. These numbers are called the HOG feature descriptors for the image. Figure 5 illustrates the implementation scheme of the HOG algorithm.

3.3 Classification

For classification, K-Nearest Neighbor (KNN) classifier is used. It is a classification method based

on closet training examples in the feature space. The training examples are vectors in a multidimensional feature space, each with a class label. The testing vector is classified by assigning the label which is a majority vote of its K nearest neighbors. It is called lazy learning because the KNN will go over all training samples to find the nearest neighbor for testing sample.

4. Experimental Setup

In this section, we discuss the dataset, evaluation parameters and the experimental results using in our approach. To evaluate the performance, we conducted experiments on the UCF-Sports dataset. It contains 10 sport actions which are diving, golf-swing, kicking, lifting, riding horse, running, skateboarding, swinging-bench, swinging-side, and walking, with a total of 150 videos. This dataset is one of the most challenging datasets because of having complicated background and large intra-class variations.

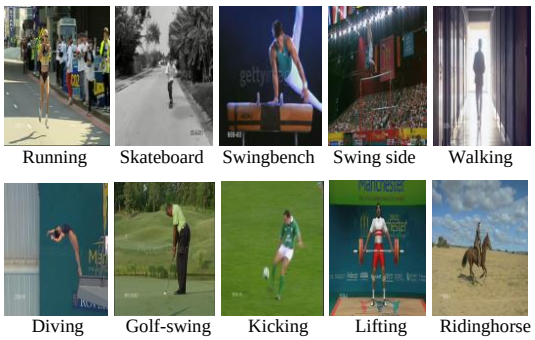


Figure 6. Sample frames from video sequences from UCF sports dataset

4.1 Experimental Results

Table 1 shows experimental results on the UCF-Sports datasets for cases with and without salient object detection. The salient object detection based recognition accuracy is obtained 88.2% while recognition accuracy without salient object detection is 83%. It shows that the proposed approach increases the accuracy by 5.2% than state-of-the-art approach. Our proposed action recognition approach can correctly classify most actions such as diving and swing side-angle. Most of the mistakes are intuitively reasonable because of various appearance variations; e.g., walking is confused with golf-swing, and kicking is confused with running. As the results, the proposed approach is superior to other action recognition approach without salient object detection.

Table 1. Recognition accuracy comparisons with and without salient object detection

Approach	Accuracy
Recognition approach with salient object detection	88.2%
Recognition approach without salient object detection	83%

4.2 Parameters

The number of octaves in SURF is specified as 2. The octave number 2 means having filter size 9-by-9, 15-by-15, 21-by-21, 27-by-27, and so on. The recommended values are between 1 and 4. Each octave spans a number of scales that are analyzed using varying size filters. Higher octaves use larger filters and subsample the image data. The number of scale levels per octave controlling the number of filters used per octave is 5. The recommended values are between 3 and 6. To detect more blobs at finer scale increments, this number can be increased.

The cell size in HOG is specified as [4 4]. To capture large-scale spatial information, the cell size can be increased. But, increasing the cell size may lose small-scale detail. The number of cells in a block is specified as [2 2]. A large block size value reduces the ability to suppress local illumination changes. Reducing the block size helps to capture the significance of local pixels. Smaller block size can help suppress illumination changes of HOG features. The number of orientation histogram bins is 8. To encode finer orientation details, increase the number of bins. Increasing this value increases the size of the feature vector, which requires more time to process.

5. Conclusion

In this paper, we have presented an efficient action recognition approach based on salient object detection. We firstly detect foreground salient objects in each video frame and process only interest feature points detected on salient objects, which carry robust information of an action. As a result, the salient feature points are more precise and maximize the discriminative information of actions. Experiments show that our proposed approach gives an effective improvement in the action recognition accuracy and outperforms than state-of-the-art action recognition methods.

References

[1]. Bobick, and J. Davis, "The Recognition of Human Movement Using Temporal Templates", *IEEE Trans Pattern Analysis and Machine Intelligence*, 2001, pp. 257 –267.

[2]. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal Salient Points for Visual Recognition of Human Actions", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 36, No. 3, 2005, pp. 710–719.

[3]. Ashwan Abdulmunem, Yu-Kun Lai, and Xianfang Sun, "Saliench guided local and global descriptors for effective action recognition", *Computational Visual Meida*, Vol.2, No.1, March 2016, pp. 97-106.

[4]. CHUANZHEN LI, BAILIANG SU, JINGLING WANG, HUI WANG, and QIN ZHANG, "Human Action Recognition Using Multi-Velocity STIPs and Motion Energy Orientation Histogram", *Journal of Information Science and Engineering*, 30, 2014, pp. 295-312.

[5]. G. Willems, T. Tuytelaars, and L. Van Gool, "An EfficientDdense and scale-invariant spatio-temporal interest point Detector", *Proceedings of 10th European Conference on Computer Vision (ECCV2008)*, Marseille, France, 2008, pp. 650-663.

[6]. 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, 2013, pp. 60-79.

[7]. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition*," Proceedings of British Machine Vision Conference (BMVC)*, London, UK, 2009, pp. 1-11.

[8]. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features", *European Conference on Computer Vision (ECCV2008).*

[9]. Hongyang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price, "Inner and Inter Label Propagation: Salient Object Detection in the Wild", *IEEE TRANSACTIONS ON IMAGE PROCESSING*, VOL. 24, NO. 10, OCTOBER 2015.

[10]. Laptev, M. Marszalek,C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

[11]. J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, 2009, pp. 2004-2011.

[12]. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, " Actions as Space-Time Shapes", *IEEETrans Pattern Analysis and Machine Intelligence*, 2007, pp. 2247–2253.

[13]. Laptev, "On Space-Interest Points", *International Journal of Computer Vision*, vol. 64, no. 2-3, 2005, pp. 107–123.

[14]. N.Dalal, B.Triggs, and C.Schmid, "Human Detection using Oriented Histograms of Flow and Appearance", *Proceedings of the 9th European Conference on Computer Vision*, Vol. 2, 2006, pp. 428–441.

[15]. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features", *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, (2005) , pp. 65–72.

[16]. P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: action recognition through the motion analysis of tracked features", *Proceedings of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshop)*, Kyoto, Japan, 2009, pp. 514-521.

[17]. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT Descriptor and Its Application to Action Recognition", *Proceedings of the 15th ACM International Conference on Multimedia*, 2007, pp. 357–360.

[18]. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels", EPFL, Lausanne, Switzerland, Tech. Rep. 149300, Jun. 2010.

[19]. R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked key points", *Proceedings of IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 104-111.

[20]. S.Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition using Shape Contexts," *IEEE Trans Pattern Analysis and Machine Intelligence*, 2001, pp. 509–522.

[21]. T. Liu et al., "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, Feb. 2011, pp. 353–367.

[22]. Thi Ly Vu, Trung Dung Do, Cheng-Bin Jin, Shengzhe Li, Van Huan Nguyen, Hakil Kim, and Chongho Lee, "Improvement of Accuracy for Human Action Recognition by Histogram of Changing Points

and Average Speed Descriptors", *Journal of Computing Science and Engineering*, Vol. 9, No. 1, March 2015, pp. 29-38.
[23]. X. Bai, X. Yang, L. J. Latecki, W. Liu, and Z. Tu, "Learning contextsensitive shape similarity by graph transduction," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 32, no. 5, May 2010, pp. 861–874.
[24]. X. Hou and L. Zhang, "Saliency detection: A spectral residual approach", *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.
[25]. Xiaojiang Peng, Yu Qiao, and Qiang Peng, "Motion Boundary Based Sampling and 3D Co-occurrence Descriptors for Action Recognition", *Image and Vision Computing*, Volume 32, Issue9, September 2014, pp. 616-628.
[26]. https://courses.cs.washington.edu/courses/cse57.6/13sp/projects/project1/artifacts/woodrc/index.htm.