FAST AND RELIABLE HUMAN ACTION RECOGNITION IN VIDEO SEQUENCES BY SEQUENTIAL ANALYSIS

Hui Fang ⋆, Jeyarajan Thiyagalingam †, Nik Bessis ⋆ and Eran Edirisinghe‡

⋆Department of Computer Science, Edge Hill University
† Department of Electrical Engineering and Electronics, University of Liverpool.
‡Department of Computer Science, University of Loughborogh.

ABSTRACT

Human action recognition from video sequences is a challenging topic in computer vision research. In recent years, many studies have explored the use of deep learning representations to consistently improve the analysis accuracy. Meanwhile, designing a fast and reliable framework is becoming increasingly important given the exponential growth of video data collected for many purposes (e.g. public security, entertainment, and early medical diagnosis etc.). In order to design a more efficient automatic human action annotation method, the sequential probability ratio test, one of the classical statistical sampling scheme, is adapted to solve a multi-classes hypothesis test problem in our work. With the proposed algorithm, the computational cost is reduced significantly without sacrificing the performance of the underlying system. The experimental results based on the UCF101 data set demonstrated the efficiency of the framework compared to the fixed sampling scheme.

Index Terms— Human action recognition, efficient video analysis, sequential analysis, sequential probability ratio test(SPRT), Convolutional Neural Networks(CNNs).

1. INTRODUCTION

Vision-based human action recognition has been paid much attention by many researchers [1, 2]. Understanding of the human activities has huge potential in many applications. For examples, it enables smart surveillance systems to enforce the public security by reducing the emergence response time; it brings convenience to video consumers by categorizing large amount of unannotated movies or clips to facilitate the high level semantic retrieval in a large scale video database; in addition, it provides the cornerstone of the modern HCI techniques once human behaviours are interpreted correctly by computers.

Recently, the recognition accuracy of video sequences collected from unconstrained environments such as the video clips shared on Youtube [3], has been improved siginificantly by exploring the Convolutional Neural Networks (CNNs) as well as the spatial-temporal fusion techniques [4]. However, due to the exponential increase of the user-generated video data, making meaningful semantic labels efficiently has become more and more important as the processing speed is a bottleneck to hinder the further development of a wide range of video applications. In particular, rising consumer interests of video streaming services require more efficient algorithms to support accurate video analysis. Therefore, the computational cost of processing a large number of frames in each video needs to be leveraged in order to make the vision based system applicable in many real world scenarios.

Sampling scheme is the most straightforward solution to make faster video annotation. The popularity of key frames extraction and video summarization techniques [5] indicates that there is huge amount of semantic redundancy in videos. However, the randomization of the frame selection causes performance drop although the sampled frames may provide sufficient information to identify the human actions. Inspired by the advantages of sequential sampling methods [6], it is believed that the adaptive scheme is suitable to achieve optimal balance between the speed and accuracy. The main contributions of our work are highlighted as: (1) we showed that the sequential analysis is able to maintain the recognition accuracy when sampling scheme is adopted to speed up the analysis; (2) we developed a probability density functions (PDF) of the sequential probabilistic ratio test (SPRT) by using the soft-max layer outputs from deep learning networks; and (3) we adapted the SPRT to solve a multi-classes recognition task named as MC-SPRT in our paper.

2. RELATED WORK

Human action recognition started by the work of identifying some basic actions, e.g. walking, running and jumping etc., in several standard data sets, such as KTH human motion dataset [7] and Weizmann action dataset [8]. In order to recognize the human actions under the well-controlled experimental environment, spatial-temporal features were normally extracted from the video sequences in order to match the pattern in each action class. One typical exampe is the Motion History Volumes (MHV) generated by using background sub-

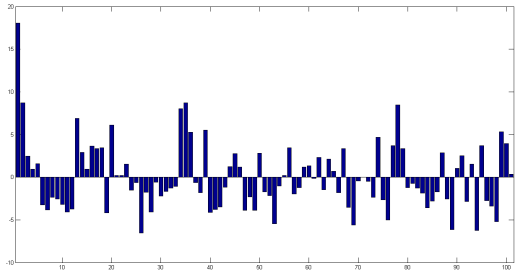
traction algorithm to provide a free viewpoint motion representation in [9].

Following the increasing demand of recognizing human action in realistic videos, much research focused on annotating videos collected under unconstrained conditions, e.g. Hollywood2 dataset [10], UCF Sport Action dataset [11], and UCF101 dataset [3] etc. In [12], dense motion trajectories and motion boundary histograms were proven to be better descriptor compared to KLT trajectories, SIFT trajectories and dense cuboids to represent the temporal features. CNNs, as the state-of-the-art image and motion feature representation method, have been widely used to solve the human action recognition task. For example, [13] integrated two ConvNets' representations, named as spatial ConvNet and temporal ConvNet, to complement the other and thus achieve better recognition accuracy; and in [14], long-term recurrent convoluational networks were trained to learn temporal dynamics of actions.

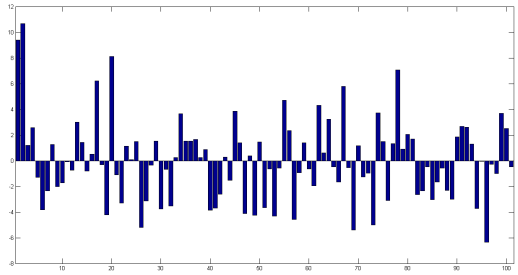
Efficient image/video analysis Improving the processing speed is a key requirement in order to apply the aforementioned video analysis architectures in real world applications. Many algorithms have been designed to target on fast image/video analysis based on efficient feature extraction [15, 16]. From the spatial feature process perspective, cascade classifier frameworks were adopted in a variety of methods to reduce the computational burden by filtering a large potion of candidate regions [17]. In [15], Zhang *et al.* used motion vector extracted from compressed video streams as an approximated feature of optical flow to reduce the processing speed of action recognition.

Predicting human activities from partial videos [18, 19] shares similar idea with our work. In [18], Ryoo presented a maximum-a-posterior(MAP) action prediction method to recognize activities from streaming videos by maximizing the activity prediction likelihood value so that only small number of frames were processed. Cao *et al.* [19] used sparse coding in a probability framework to predict human action. While the advantage of our work is that we propose an adaptive scheme to accumulate evidence over time and make a decision once a desired level is reached.

Sequential analysis has been extended and widely adopted into many real-time applications with limited resources, such as cyber-attack detection [20], sparse signal recovery [21] and a number of image processing and multimedia processing tasks [22, 23]. Lelescu *et al.* [22] used the statistical theory to provide a unified framework in order to tackle the challenging detection of gradual changes across video scenes, e.g. fades and dissolves, as well as the abrupt changes in a real-time video scene change detection application. In [23], Herout *et al.* proposed to use SPRT to adaptively sample a varying number of pixels for reliable edge classification.



(a) CNN Responses of a sample with correct classification



(b) CNN Responses of a sample with incorrect classification

Fig. 1. The feature representations from two sample frames in two video sequences. For a good sample, peak response is much larger than confusing response while for a bad sample, peak response is close to the confusing response. The x-axis represents the 101 classes and y-axis represents each response value.

3. PROPOSED METHOD

In this work, we aim at designing a fast and reliable video based human action recognition method by integrating deep learning feature representations into a SPRT framework. With this adaptive sampling scheme, a classification decision is made once the likelihood of an action class is accumulated to a high level by validating the probability values of a subset of frames. Let F be a subset of frames $\{F_1, ..., F_n\}$ in a video clip with cardinality $f = |F|$. Based on the statistical distribution of feature representations extracted from these frames, an action label is provided without processing more frames if the confidence value calculated from the log-likelihood ratio (LLR) is significant. The main components of our method, including CNNs' feature representations, probabilistic models and MC-SPRT, are explained in detail as follows.

3.1. CNNs' feature representations

CNNs is the state-of-the-art architecture to extract distinctive feature representations in most vision applications. Millions of nodes (neurons) which contain weights and biases enable the network to learn nonlinear mapping functions under more complicated recognition senarios, including the human action recognition task investigated in our work. As is demonstrated in [13], the fusion of two-stream convolutional neural net-

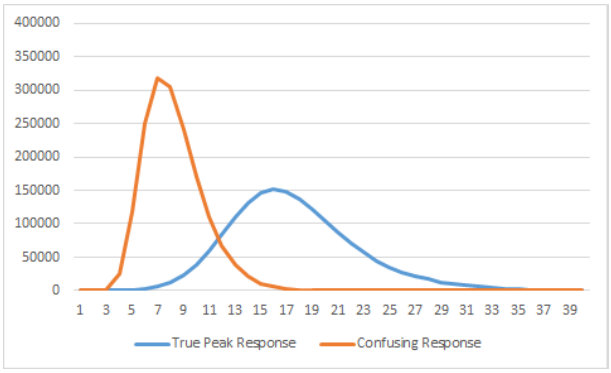


Fig. 2. The distribution of peak responses and confusing responses obtained from the training set. x axis represents the response values and y axis represents the frame number.

works, spatial network and temporal network, enhance the distinctive power as the feature representations obtained from the two networks are complementary with each other to improve the system performance. On one hand, the spatial configuration of salient objects in each frame indicates the potential human actions in a video sequence. On the other hand, each action has its own motion patterns to facilitate the recognition of the action.

We use the configuration in [4] to extract the feature representations: (1) the input to the spatial stream convolutional neural network is a 224X224X3 RGB colored image by resizing still frames; (2) the input to the temporal stream CNN is 2L channels of motion feature by stacking the horizontal and vertical scalar values from the optical flow vectors (L is the frame number and it is set to 10); and (3) the networks are trained by using VGG model [13] as initialization. Using the networks, the response vector output from the softmax layer is used as the feature to calculate the confidence level of the SPRT. As illustrated in Figure 1, the margin between the peak response, i.e the feature value of the correct classification, and the confusing response, i.e. the maximum response value by removing the peak response, is high in a good sampled frame while two response values are close to each other in a poor sampled frame.

3.2. Probabilistic models

In order to use the SPRT adaptive sampling method, probabilistic values need to be estimated in order to accumulate the confidence level. Considering that each video clip can be categorized into one of the K classes (101 action classes in the UCF101 data-set), for any frame index i and class k, let R_i^k be the response value extracted from the CNNs. We assume that R_i^k follows i.i.d. distribution and we refer to H_g as the hypothesis that the i_{th} frame is good for recognizing the k_{th} action and H_b the alternative. Therefore the hypotheses can be expressed as:

$$H_g : \vec{R_i^k} \sim P_g(\vec{R_i^k})$$
$$H_b : \vec{R_i^k} \sim P_b(\vec{R_i^k}) \quad (1)$$

Based on the law of large numbers (LLN), the central limit theorem is used to model the distribution of the true peak response values and confusing response values. The distributions are confirmed by Figure 2 which plots the histograms of the peak response values and the confusing response values from the data in the training set. In this figure, the number of frames for spatial and temporal features in the training set to generate the distributions are 1,776,542 and 1,690,709.

Algorithm 1 MC-SPRT: multi-classes SPRT to decide when to stop sampling frames for human action recognition

```
Input: Video clip V;
Output: Action class C_j;
for i = 1:length(V) do
    t = index of max(R_i)
    if t ∉ testArray then
        add t into testArray
    end if

    for k = 1:length(testArray) do
        comupte L^k
        if L^k < a then
            remove k from testArray
        else if L^k > b then
            C_j = k; break OUTER LOOP;
        end if
    end for
end for
```

3.3. MC-SPRT

SPRT is a statistical method to test the hypothesis in Equation 1 by computing the cumulative sum of the LLRs when samples are added sequentially. The LLR of multiple sampled frames is defined as:

$$\mathcal{L}^k(R_1^k, ..., R_j^k) \triangleq \sum_{i=1}^{j} log \frac{P_g(R_i^k)}{P_b(R_i^k)} \quad (2)$$

Here, the superscript k represents the action class and the subscript j represents the number of sampled frames to calculate the LLR. As a result, the fast annotation problem can be redefined as finding an optimal number of frames for reliably recognizing human actions in video clips by calculating the likelihood and accumulating confidence of the classification. Two thresholds, namely a and b, are defined based on type I error α and type II error β to accept or reject the hypothesis. In our experiments, the α (0.11) and the β (0.06) were obtained based on the training set. By using the standard SPRT confidence setting methods, the two thresholds a and b are -2.63 and 2.14 respectively.

Method	CNNs' Process %	Accuracy	STD.
sCNN Average	50%	72.8	-
tCNN Average	50%	78.8	-
sCNN & tCNN Average	100%	**86.2**	-
sCNN & tCNN Fixed 1%	1%	80.9	$\pm$ 0.003
sCNN & tCNN Fixed 2%	2%	84.1	$\pm$ 0.003
sCNN & tCNN Fixed 3%	3%	84.8	$\pm$ 0.002
Proposed method	$\approx$ 1.3%	**85.0**	$\pm$ 0.002

Table 1. Compared to fixed sampling schemes, the adatpive sampling scheme generates a very competitive result but processing fewer frames.

The binary SPRT is further extended to a multi-classes hypothesis test when new classes are added into the candidate list in the process because the correponding classes to the maximum reponses may vary when sampling frames sequentially. The pseudo-code of MC-SPRT method is provided in Algorithm 1. The testArray is for adding hypothesis test candidates into the list instead of testing all the 101 classes each time.

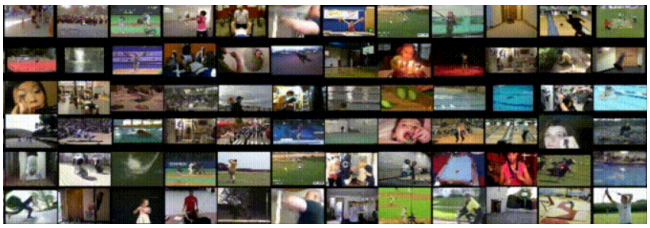


Fig. 3. The examples from the UCF101 action dataset [3].

4. EXPERIMENTAL RESULTS

4.1. Dataset Description

The UCF101 dataset is used to test the efficiency of our proposed method given that it is one of the most popular and challenging datasets to evaluate the performance of human action recognition from video in the wild [3]. The dataset includes total number of 101 action classess with 13320 video clips. The video clips are downloaded from Youtube while the condition of these clips unconstrained to the lab configuration with various of illuminations. The list of the human actions can be found in [3] and some example frames of the dataset are illustrated in the Figure 3.

4.2. Results

The comparison results provided in the Table 1 demonstrated the efficiency of the proposed method. The result taht we obtained are consistent with the results reported in previous work [13], the recognition accuracy by fusing all the features extracted from the spatial network and temporal network reached 86.2% while the spatial CNN and temporal CNN obtained a lower performance of 72.8% and 78.8% respectively. Under the fixed sampling schemes, 1%, 2% and 3% were set to sample frames to extract CNNs' feature representations and these schemes ahieved 80.9%, 84.1% and 84.8% accuracy rate separately. By using the proposed adaptive sampling scheme, a very small numer of samples were used of approximately 1.3%. However, our system achieved a very competitive accuracy rate, 85.0%, compared to the method of processing all the frames in each sequence. The sampling schemes were run 50 times to analyze the variations of the methods. The standard deviation value showed that the performance of the proposed method is reliable.

5. CONCLUSION

In this paper, we designed a MC-SPRT method to improve the accuracy rate of human action recognition while a sampling scheme is used to accelerate the high sematic video analysis. The sequential probabilistic ratio test is used to evaluate the cumulative confidence whether a video clip can be classified to an action class by the multi-classes hypothesis tests. Once the probability is significant enough, the sampling process stops to avoid high computational cost of the video semantic analysis. The results based on the evaluation of UCF101 dataset prove the efficiency of the adaptive sampling scheme. Although many new achietectures of CNNs have been designed to improve the recognition rate consistently in these couple of years, we believe that the MC-SPRT can make a significant contribution towards fast and reliable action recognition by integrating the new methods into the statistical framework. In future work, we will improve the achietecture of current CNNs as well as test the framework in more public datasets.

6. ACKNOWLEDGEMENT

Authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal and Quadro M5000 GPU cards used for this research.

7. REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 2010, pp. 976–990, 2010.

[2] Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernndez-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633 – 659, 2013.

[3] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," in *CRCV-TR-12-01*, 2012.

[4] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 568–576. Curran Associates, Inc., 2014.

[5] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-aware surveillance video summarization," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5469–5478, Nov 2016.

[6] K. Cohen, Q. Zhao, and A. Swami, "Optimal index policies for anomaly localization in resource-constrained cyber systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 8, pp. 4224–4238, Aug 2014.

[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the IEEE International Conference on Pattern Recognition*, 2004, pp. 31–36.

[8] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 1395–1402.

[9] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249 – 257, 2006.

[10] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[11] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proceedings of CVPR*, 2009, pp. 1996–2003.

[12] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of CVPR*, June 2015.

[15] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang, "Real-time action recognition with enhanced motion vector cnns," in *Proceedings of CVPR*, June 2016, pp. 2718–2726.

[16] Z. Li, J. Jiang, G. Xiao, and H. Fang, *An Effective and Fast Scene Change Detection Algorithm for MPEG Compressed Videos*, pp. 206–214, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[17] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proceedings of BMVC 2015*, 2015.

[18] M. Ryoo, "Human activity prediction: early recognition of ongoing activities from streaming videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1036–1043.

[19] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *Proceedings of CVPR*, June 2013, pp. 2658–2665.

[20] Justin Grana, David Wolpert, Joshua Neil, Dongping Xie, Tanmoy Bhattacharya, and Russell Bent, "A likelihood ratio anomaly detector for identifying within-perimeter computer network attacks," *Journal of Network and Computer Applications*, vol. 66, pp. 166 – 179, 2016.

[21] M. L. Malloy and R. D. Nowak, "Sequential testing for sparse recovery," *IEEE Transactions on Information Theory*, vol. 60, no. 12, pp. 7862–7873, Dec 2014.

[22] Dan Lelescu and Dan Schonfeld, "Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream," *IEEE Transactions on Multimedia*, vol. 5, no. 1, pp. 106–117, 2003.

[23] Adam Herout, Istvan Szentandrasi, Michal Zacharias, Marketa Dubska, and Rudolf Kajan, "Five shades of grey for fast and reliable camera pose estimation," in *Proceedings of CVPR*, 2013, pp. 1384–1390.