TEMPORAL SALIENCE BASED HUMAN ACTION RECOGNITION

Salah Al-Obaidi and Charith Abhayaratne

Department of Electronic and Electrical Engineering, The University of Sheffield
Sheffield, S1 3JD, United Kingdom
Email: s.alobaidi@sheffield.ac.uk, c.abhayaratne@sheffield.ac.uk

ABSTRACT

This paper proposes a new approach for human action recognition exploring the temporal salience. We exploit features over the temporal saliency maps for learning the action representation using a local dense descriptor. This approach automatically guides the descriptor towards the most interesting contents, i.e. the salience region, and obtains the action representation using solely the saliency information. Outperforming results on Weizmann, DHA and KTH datasets confirm the efficiency of the proposed approach as compared to the state-of-the-art methods, in terms of accuracy and robustness to the variations inside the action and similarities among actions. The proposed method outperforms by 2.7% with DHA, 1% with KTH and it is comparable in the case of Weizmann.

***Index Terms*—** Human Action Recognition (HAR), Temporal Salience, Salience-based HAR, Histogram of Oriented Gradients of Salience (HOG-S).

1. INTRODUCTION

Human action recognition (HAR) [1–10] is significantly used in a variety of applications, such as, video surveillance, human computer interaction, healthcare monitoring, smart homes. Vision-based HAR is still a challenge due to different limitations, such as light conditions, occlusion and cluttering background. These problems can be overcome by acquiring a set of features and training a classifier leading to promising results. However, uncorrelated and lost information may occurr during the feature extraction [1]. Thus, the performance of any recognition system is based on both the action representation model and the action classification method [2]. Many works have been proposed to represent actions using the local dense trajectories representation, such as, Histogram of Oriented Gradients (HOG) [3], due to its robustness [4]. The existing works on HOG-based HAR are categorised into two themes: 2D HOG [5–7] and 3D HOG [8–10] representations. In the first category, the dense features are extracted from a single image/frame to show the motion history. In the second category, a volumetric representation in space-time is exploited to represent the action. However, in both categories, redundant information, such as, the global / local motion, is generated. This affects the discriminating power of the descriptor, increases storage requirements of this information and makes the complexity high. Mainly, a little research has been done to address these problems.

This paper proposes a new approach for HAR exploring the temporal salience. The method does not include high complexity motion estimation algorithms. Instead it generates temporal salience maps, considering the spatial changes within successive frames followed by modelling the temporal salience maps using HOG leading to HOG of salience (HOG-S) features that are finally classified using the KNN classifier. The main contributions of this paper are:

1. Exploring the temporal saliency maps for HAR.
2. Proposing a salience based descriptor to encode each action using the HOG of salience (HOG-S).

The rest of the paper is organised as follows: In Section 2, the proposed method is presented, followed by the performance evaluation in Section 3 and conclusions Section 4.

2. THE PROPOSED METHOD

Fig. 1 illustrates the main steps of the proposed approach. Let $C = \{s_i^F, l_i\}_{i=1}^V$ be the actions dataset with V video samples and L classes, where s_i is the ith RGB video contains F frames and $l \in L$. The aim is to recognise the human actions using the HOG of the temporal salience. This approach includes two parts: temporal salience modelling and HOG of salience (HOG-S) feature extraction.

2.1. Temporal salience modelling

The temporal salience modelling approach is divided into the following four steps.

A) At the beginning, frame difference between each two consecutive frames, $(f_t, f_{t-1}) \in s_i$, where $t \in [1, \cdots, F]$, is obtained to define the changes in the pixel intensity. Then, this difference is compared with

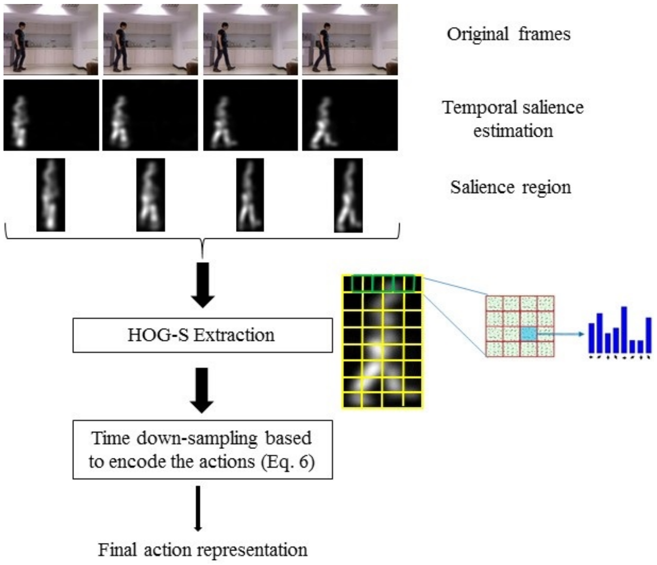
Original frames
Temporal salience estimation
Salience region
HOG-S Extraction
Time down-sampling based to encode the actions (Eq. 6)
Final action representation

Fig. 1. Proposed approach for temporal saliency estimation and feature extraction

a user-defined threshold, τ, to detect the moving pixels, as follows:

$$D_f(x,y) = \begin{cases} |f_t(x,y) - f_{t-1}(x,y)| \text{ if } |f_t(x,y) - \\ \qquad\qquad\qquad\qquad f_{t-1}(x,y)| \geq \tau \ , \\ 0 \qquad\qquad Otherwise \end{cases} \tag{1}$$

where $D_f(x,y)$ is the frame difference at location (x,y) after thresholding. However, it is difficult to determine the perfect value of τ. To make this representation more robust and generalised, we further vary τ into $\mathcal{H}$ values to obtain $\mathcal{H}$ difference maps. Each a difference map will be analysed independently.

B) The deference map based on the threshold τ_h, i.e. $D_f^{\tau_h}$, where $h \in [1, \cdots, \mathcal{H}]$, is then processed using an overlapped block-based two dimensions fast Fourier transform (2DFFT) based entropy in order to join the small regions to form the silhouette. Due to $D_f^{\tau_h}$ has magnitudes distributed over the foreground region, a $N \times N$ entropy based operator is adopted to link these isolated magnitudes and highlight the temporal pose. $D_f^{\tau_h}$ is partitioned into overlapped $N \times N$ blocks and then 2DFFT is applied on each block to analyse the frequencies in these blocks, separately.

C) After that, the normalised Power Spectral Density (NPSD) of the block is calculated. These probability densities are exploited to obtain the salience model of the object using the local Shanoon entropy. This entropy assigns a score for each pixel based on the contribution of the neighbouring magnitudes over the $N \times N$ block. Thus, the high magnitudes produce the high score. Since we have $\mathcal{H}$ difference maps are analysed, there are $\mathcal{H}$ entropy maps are obtained. These entropy maps are used to obtain the weighted entropy map, $\hat{\mathcal{E}}$, as follows:

$$\tilde{\mathcal{E}}(x,y) = \frac{\sum_{h=1}^{\mathcal{H}} \tau_h \mathcal{E}^{\tau_h}(x,y)}{\sum_{h=1}^{\mathcal{H}} \tau_h}, \tag{2}$$

where $\mathcal{E}^{\tau_h}(x,y)$ and $\tilde{\mathcal{E}}(x,y)$ are the entropy and the weighted entropy values at location (x,y), respectively.

D) The $\tilde{\mathcal{E}}(x,y)$ is then normalised to the [0 255] values and smoothed by applying a 2-D Gaussian kernel with $\sigma = 4$ in order to fill the small holes if found and obtain the final temporal saliency map.

The above procedure is an overlapped block-based processing. Since each value in the saliency map is calculated based on its neighbours, the salience value of each location is affected by the values of the neighbouring locations. Mainly, this locally entropy approach links fairly the small regions that are close to each other leading to construct the temporal salience of the silhouette of the foreground object. This method models the temporal salience without including complex motion estimation approaches. Rather, modelling the spatial changes within consecutive frames is used to compute the temporal salience map. This map is explored to extract new features by applying the HOG on the obtained temporal salience map.

Fig. 2 shows the result of applying the above procedure on a sample frame of a walking sequence from DHA dataset. As we can see that our proposed saliency model further highlighting the most dynamic parts compared to other parts of the body. The moving parts represented with high salience magnitude, which reflects the optimal performance of the proposed saliency model. For the seek of robustness, we test several sequences of actions in the DHA dataset, and it shows an efficient performance by emphasising the most dynamic parts

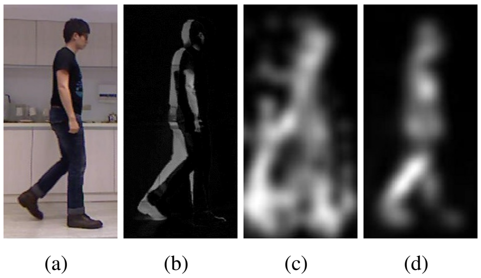

Fig. 2. Temporal saliency estimation of frame #15 from the walking sequence for the participant #1: (a) original frame, (b) temporal difference map, (c) temporal saliency map with a single threshold, $\tau = 4$ and (d) temporal saliency map based on the proposed weighted entropy.

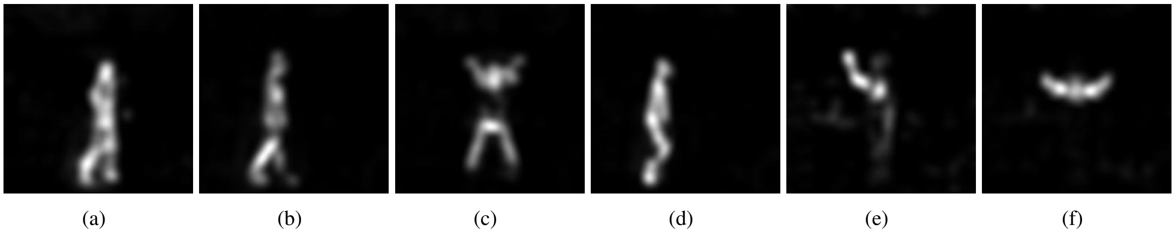

Fig. 3. Temporal saliency maps for action samples: (a) run, (b) walk, (c) jack, (d) jump, (e) One hand waving and (f) Two hands waving.

as shown in Fig. 3. For instance, Fig. 3e and Fig. 3f provide different saliency maps for One hand waving and Two hands waving respectively. This proposition is crucial to compute the HOG-S, which accurately identifies the variation between different actions.

2.2. HOG-S feature extraction

The HOG-S descriptor is obtained by calculating the horizontal and vertical gradients of the salience region, i.e. d_{x_s} and d_{y_s}, in each video frame. Then, the magnitude, G_S, and orientation, θ_S, of the gradients are computed as follows:

$$G_S = \sqrt{d_{x_s}^2 + d_{y_s}^2}, \tag{3}$$

$$\theta_S = arctan\left(\frac{d_{y_s}}{d_{x_s}}\right). \tag{4}$$

The gradient orientations are then quantized into 9-bins orientation histogram. Let G_S be an $K \times L$ matrix containing the gradient response. Thus, there are $B_K \times B_L$ blocks from witch HOG-S features are extracted. Each block in turn is subdivided into P patches and each patch give us 9-bins histogram. Thus, when we concatenate them all into one response vector,**v**, we obtain a $9 \times P \times B_K \times B_L$ dimensions vector. Then, **v** is normalised using (5) in order to make the description more light-invariant.

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{||\mathbf{v}||_2^2}, \tag{5}$$

where $\hat{\mathbf{v}}$ is the normalised HOG-S. However, $\hat{\mathbf{v}}$ can not be perfectly discriminating among features in according to the variation inside the action and similarities among the actions. To address this, we improve the HOG-S descriptor by proposing a time down-sampling to divide the frames of each sequence into two groups: even and odd time-based vectors. Thus, the set of normalised vectors $\hat{\mathbf{V}} = \{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \hat{\mathbf{v}}_3, \cdots, \hat{\mathbf{v}}_T\}$, is down-sampled by 2 with respect to T, i.e. $T \downarrow 2$. By this, $\hat{\mathbf{V}}$ is partitioned into even and odd time vectors. The final feature vector at time instant t, $\tilde{\mathbf{v}}_t$, is computed as

$$\tilde{\mathbf{v}}_t = \left| \sum_{k=0}^{t-1} \hat{\mathbf{v}}_{t-2k} - \sum_{k=0}^{t-1} \hat{\mathbf{v}}_{t-(2k+1)} \right|, \tag{6}$$

where $t - k > 0$.

By using (5), the discriminating power of HOG-S descriptor is improved by increasing both the dissimilarity and the similarity between the actions and inside the action, respectively.

3. PERFORMANCE EVALUATION

To explore the robustness of the proposed framework, two publicly available datasets with different scenarios, i.e. indoor and outdoor environments, are used for evaluation. Weizmann dataset [11] contains $V = 93$ low-resolution (144×180, 50 fps) video sequences showing $L = 10$ actions achieved by 9 actors. Depth-included Human Action (DHA) video dataset [12] contains $L = 23$ action categories performed by participating 21 different individuals (12 males and 9 females). It is recorded using a static Kinect camera in three different scenes with 480×640 resolution. Finally, the third dataset, KTH [13], is a multiview scenario dataset including $L = 6$ actions with several variations. The actions are performed several times by 25 subjects in four different scenarios. It is recorded using a static camera with $25fps$ frame rate and the spatial resolution is 160×120 pixels.

In our experiments, the number of user-defined thresholds equals to 7 with values= 4, 8, 16, 32, 64, 128, 256. The size of the overlapped block is 3×3, which is chosen based on experiments. All the temporal saliency maps are firstly resized to the resolution 256×256 to apply the same parameters on both datasets. In our experiments, the size of the salience region is selected to be 168×72 resolution to crop all the salience contents. This bounding box produces a 23040 dimensions HOG-S feature vector. We found that the patch size 4×4 with $P = 16$ for each block achieves the best results.

Our method has resulted in overall accuracies of 99.65%, 99.39 and 99.06%, for the 3 datasets, Weizmann, DHA, and KTH, respectively. The corresponding confusion matrices are shown in Figures 5, 4 and 6, respectively. Table 1 shows how the proposed approach compares with the exiting methods for the 3 datasets. The proposed method provides the best performance for DHA and KTH datasets, while provides a close second best results for Weizmann dataset.

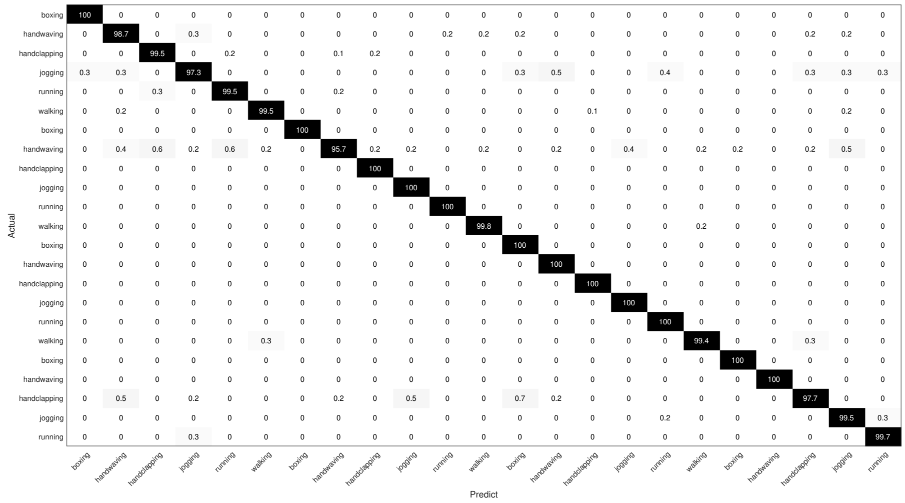

Fig. 4. Confusion matrix of human action recognition on DHA dataset (Overall accuracy: 99.39%)

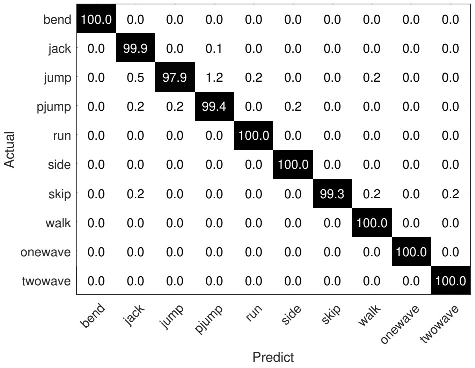

Fig. 5. Confusion matrix of human action recognition on Weizmann dataset (Overall accuracy: 99.65%)

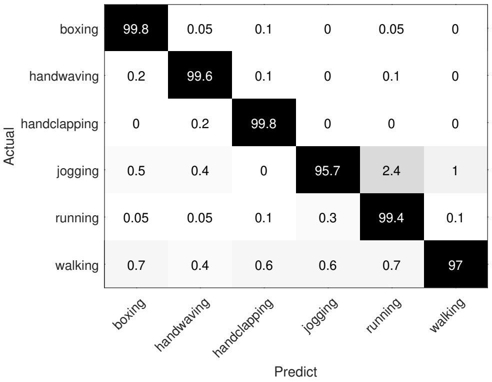

Fig. 6. Confusion matrix of human action recognition on KTH dataset (Overall accuracy: 99.06%)

Table 1. Recognition accuracy (%) of the proposed HOG-S and the state of the art for Weizmann, DHA and KTH datasets: a comparison

Method	Weizm-ann	Method	DHA	Method	KTH
[14]	97.98	[15]	86.5	[16]	94.8
[17]	99.1	[18]	95	[19]	94.2
[19]	98.9	[20]	95.45	[21]	96.80
[10]	100	[22]	96.69	[23]	98.16
Ours	**99.65**	**Ours**	**99.39**	**Ours**	**99.06**

4. CONCLUSIONS

This paper has addressed the problem of extracting an accurate discriminating representation for HAR based on temporal saliency maps. It relies on exploiting the temporal saliency to construct the HOG-S descriptor for HAR. The proposed framework achieved better performance in terms of accuracy over the exploited datasets compared to the state-of-the-art. This outperforming has been shown in Table 1 for KTH and DHA datasets. The confusion matrices verified that the new features are better discriminated between actions as well as reducing the variations inside the action itself.

5. REFERENCES

[1] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 252–264, 2015.

[2] K Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[3] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off," *International Journal of Multimedia Information Retrieval*, vol. 4, no. 1, pp. 33–44, 2015.

[4] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, 2017.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, vol. 1, pp. 886–893.

[6] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao, "Simple to complex transfer learning for action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 949–960, 2016.

[7] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Computer Vision*, vol. 10, no. 7, pp. 758–767, 2016.

[8] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proceedings of British Machine Vision Conference (BMVC)*, 2008, pp. 275–1.

[9] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010, pp. 635–648.

[10] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi, "3D-HOG embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[12] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *Proceedings of the International conference on Multimedia*. IEEE, 2012, pp. 1053–1056.

[13] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE, 2004, vol. 3, pp. 32–36.

[14] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 236–243, 2013.

[15] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the International conference on Multimedia*. IEEE, 2012, pp. 1057–1060.

[16] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1860–1870, 2013.

[17] K. Xu, X. Jiang, and T. Sun, "Two-stream dictionary learning architecture for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 567–576, 2017.

[18] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, 2015.

[19] M. Rodriguez, C. Orrite, C. Medrano, and D. Makris, "One-shot learning of human activity with an MAP adapted GMM and simplex-HMM," *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1769–1780, 2017.

[20] H. Liu, Q. He, and M. Liu, "Human action recognition using adaptive hierarchical depth motion maps and gabor filter," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1432–1436.

[21] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[22] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, "Action recognition using 3D histograms of texture and a multi-class boosting classifier," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4648–4660, 2017.

[23] M. Tong, Y. Chen, M. Zhao, and W. Tian, "A new framework of action recognition with discriminative parts, spatio-temporal and causal interaction descriptors," *Journal of Visual Communication and Image Representation*, vol. 56, pp. 116–130, 2018.