TEMPORAL SALIENCE BASED HUMAN ACTION RECOGNITION

Salah Al-Obaidi and Charith Abhayaratne

Department of Electronic and Electrical Engineering, The University of Sheffield, United Kingdom.

Email: s.alobaidi@sheffield.ac.uk, c.abhayaratne@sheffield.ac.uk, Web: http://svc.group.shef.ac.uk

For More Information

http://tiny.cc/demohar

@charithmetic

Introduction

- Human action recognition (HAR) 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 occur during the feature extraction.
- Usually, HAR explores the pixel domain to represent actions, which means large amounts of redundant data to be processed. In addition, these methods use complex motion estimation algorithms to model the actions, leading to high complexity.
- This work proposes exploring temporal saliency for human action modelling. Temporal salience models capture salience in a visual scene with respect to motion of objects, and automatically filter out the background.
- The main contributions of this work are:
  1. Exploiting the temporal saliency maps for HAR.
  2. Proposing a novel salience based descriptor to encode each action using the Histograms of gradients (HOG) of salience (HOG-S).

The Proposed Method

The proposed methodology consists of two stages:

A) Temporal salience based action modelling and

B) Histogram of oriented gradient of the temporal saliency maps for feature extraction.

A.1 - Define the temporal intensity change

Frame difference between each two consecutive frames, (f_t, f_{t-1}) is obtained to define the changes in the pixel intensity. Then, this difference is compared with an user-defined thresholds to detect the moving pixels.

A.2- Block based 2DFFT

The difference map based on the threshold is then processed using an overlapped block-based two dimensions fast Fourier transform (2DFFT). The difference map is partitioned into overlapped N x N blocks and then 2DFFT is applied on each block to analyse the frequencies in these blocks, separately.

A.3- Calculating the weighted local entropy

The normalised Power Spectral Density (NPSD) of each block is calculated. These probability densities are used to calculate the local Shanoon entropy. This entropy assigns scores for all pixel over the N x N blocks. Since we have several difference maps are analysed, there are corresponding entropy maps.

These entropy maps are used to obtain the weighted entropy map, $\tilde{\varepsilon}$, as in the next step.

A.4- Define the temporal intensity change

Frame difference between each two consecutive frames, (f_t, f_{t-1}) is obtained to define the changes in the pixel intensity. Then, this difference is compared with user-defined thresholds to detect the moving pixels.

No. of thresholds

Entropy map based on a specific threshold

$$\tilde{\varepsilon}(x,y) = \frac{\sum_{h=1}^{H} \tau_h \, \varepsilon^{\tau_h}(x,y)}{\sum_{h=1}^{H} \tau_h}.$$

A user-defined threshold

A.5- Normalising and smoothing the weighted entropy map

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

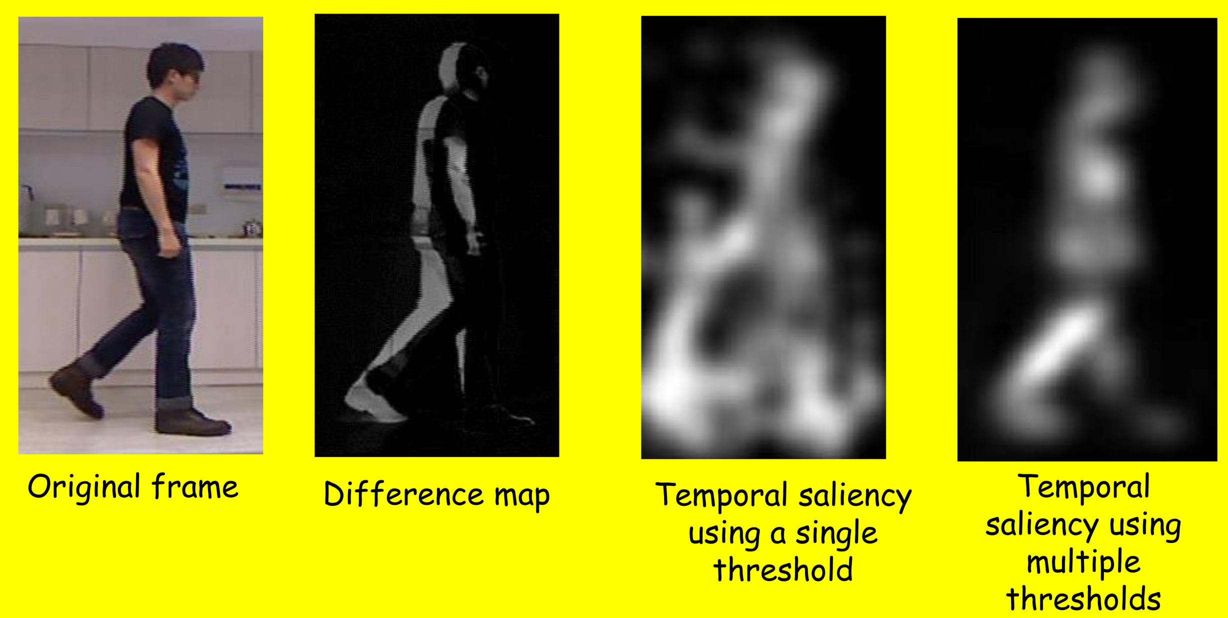

Fig. 1: Temporal saliency estimation

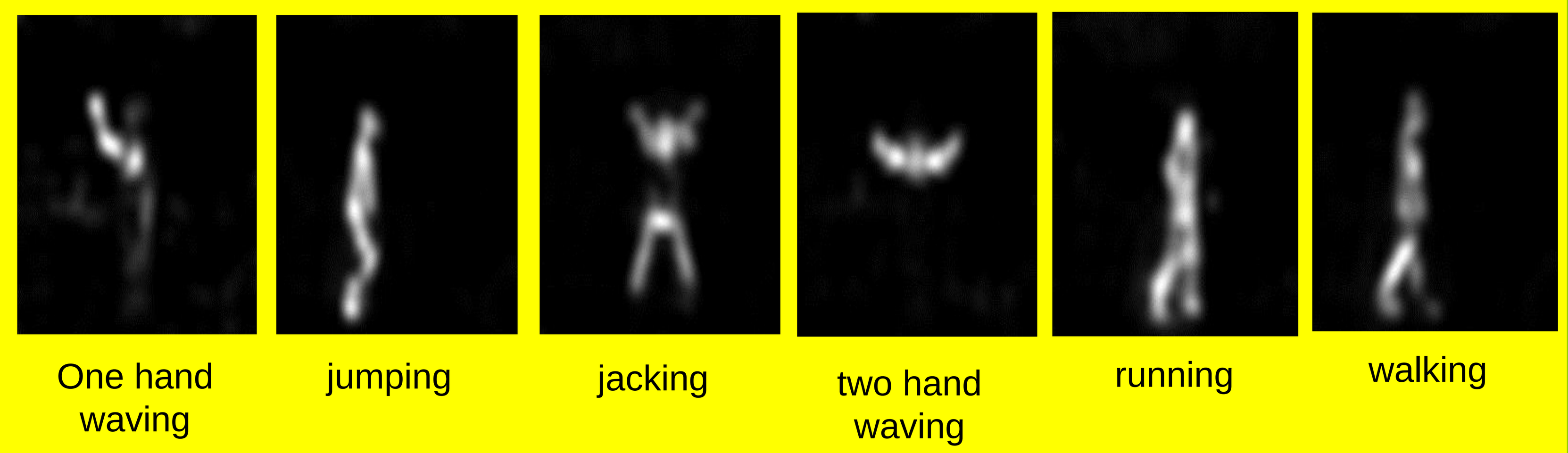

Fig. 2: Temporal saliency maps for action samples (Weizmann dataset). For more examples of temporal saliency map video sequences, please visit: http://tiny.cc/demohar

B.1- Feature extraction

The proposed descriptor is obtained by applying the histogram of oriented gradient (HOG) on the salience regions of the temporal salience maps leading to HOG of the temporal salience (HOG-S) such in Fig. 3. The HOG-S descriptor has been improved by proposing a time down-sampling to divide the temporal salience maps of each sequence into two groups: even and odd time-based vectors. The final feature vector $\tilde{v}$ at time instant, t, is computed as follows:

Normalised HOG-S

$$\tilde{v}_t = \left| \sum_{k=0}^{t-1} \hat{v}_{t-2k} - \sum_{k=0}^{t-1} \hat{v}_{t-(2k+1)} \right|, \quad \text{Where } t-k>0.$$

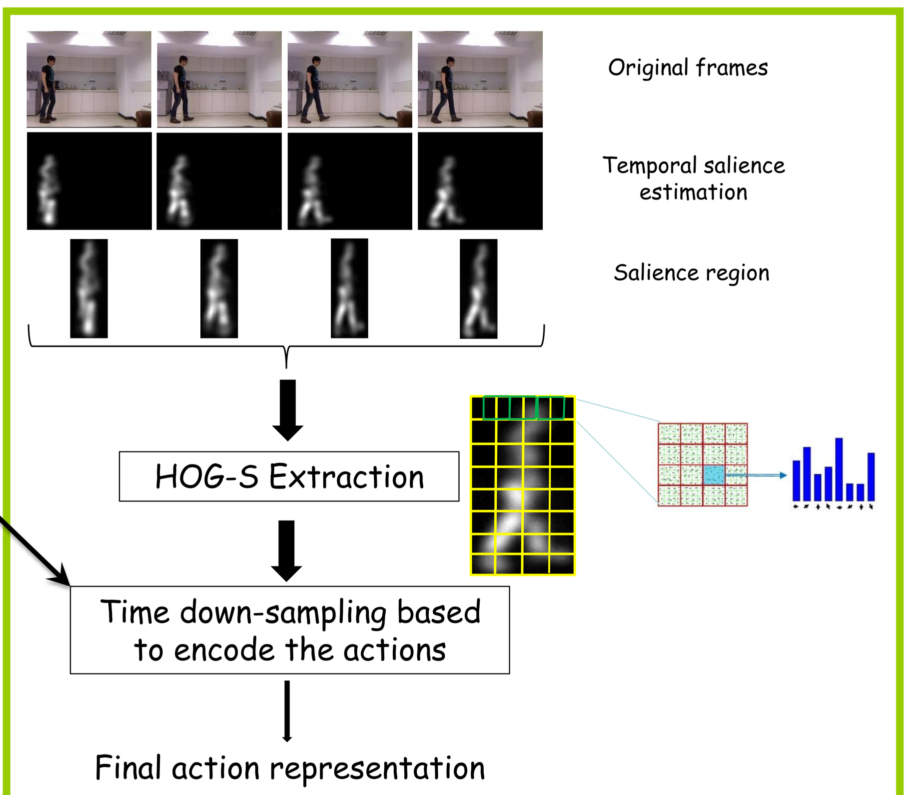

Fig. 3: Proposed approach for temporal saliency estimation and feature extraction

B.2- Classification

Since, the above feature extraction method reduces the variation of the feature vector within a give action and increases the discrimination among the actions set, the discriminant analysis classifier is expected to work well.

Several experiments were conducted to find the optimal classifier and the KNN classifier with 1 neighbouring shows better performance.

Results

Three datasets:

1- Weizmann (Gorelick et al. 2007): 93 low-resolution (144x180, 50 fps) video sequences showing 10 actions achieved by 9 actors.

2- Depth-included Human Action (DHA) (Lin et al. 2012): 23 action categories performed by participating 21 different individuals (12 males and 9 females).

3- KTH (Schuldt et al. 2004): a multi-view scenario dataset including 6 actions with several variations.

Experimental parameters:

1- Seven user- defined thresholds with values values= 4, 8, 16, 32, 64, 128, 256.

2- The size of the overlapped block is 3x3.

3- The size of the salience region is 168x72 resolution.

4- This bounding box produces a 23040 dimensions HOG-S feature vector.

Table 1. Recognition accuracy (%) of the proposed HOG-S and the state of the art for Weizmann.

Method	Accuracy
Wu and Shao (2013)	97.98
Xu et al. (2017)	99.1
Rodriguez et al. (2017)	98.9
Angelini et al. (2018)	100
Proposed	**99.65**

Table 2. Recognition accuracy (%) of the proposed HOG-S and the state of the art for DHA.

Method	Accuracy
Yang et al. (2012)	86.5
Gao et al. (2015)	95
Liu et al. (2017)	95.45
Zhang et al. (2017)	96.69
Proposed	**99.39**

Table 3. Recognition accuracy (%) of the proposed HOG-S and the state of the art for KTH.

Method	Accuracy
Liu et al. (2013)	94.8
Rodriguez et al. (2017)	94.2
Shi et al. (2017)	96.80
Tong et al (2018)	98.16
Proposed	**99.06**

Table 4. Confusion matrix of KTH dataset using KNN classifier (Overall accuracy: **99.06 %**)

Actual \ Predict	Boxing	Hand waving	Hand clapping	Jogging	running	Walking
Boxing	99.8	0.05	0.1	0	0.05	0
Hand Waving	0.2	99.6	0.1	0	0.1	0
Hand clapping	0	0.2	99.8	0	0	0
Jogging	0.5	0.4	0	95.7	2.4	1
Running	0.05	0.05	0.1	0.3	99.4	0.1
Walking	0.7	0.4	0.6	0.6	0.7	97

Table 5. Confusion matrix of Weizmann dataset using KNN classifier (Overall accuracy: **99.65 %**)

Actual \ Predict	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	100	0	0	0	0	0	0	0	0	0
Jack	0	99.9	0	0.1	0	0	0	0	0	0
Jump	0	0.5	97.9	1.2	0.2	0	0	0.2	0	0
Pjump	0	0.2	0.2	99.4	0	0.2	0	0	0	0
Run	0	0	0	0	100	0	0	0	0	0
Side	0	0	0	0	0	100	0	0	0	0
Skip	0	0.2	0	0	0	0	99.3	0.2	0	0.2
Walk	0	0	0	0	0	0	0	100	0	0
Wave1	0	0	0	0	0	0	0	0	100	0
Wave2	0	0	0	0	0	0	0	0	0	100

Table 6. Confusion matrix of DHA dataset using KNN classifier (Overall accuracy: **99.39%**)

Actual \ Predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	98.7	0	0.3	0	0	0	0	0	0	0.2	0.2	0.2	0	0	0	0	0	0	0	0.2	0.2	0
3	0	0	99.5	0	0.2	0	0	0.1	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0.3	0.3	0	97.3	0	0	0	0	0	0	0	0	0.3	0.5	0	0	0.4	0	0	0	0.3	0.3	0.3
5	0	0	0.3	0	99.5	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0.2	0	0	0	99.5	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0.2	0
7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0.4	0.6	0.2	0.6	0.2	0	95.7	0.2	0.2	0	0.2	0	0.4	0	0.2	0.2	0	0.2	0	0.2	0.5	0
9	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	99.8	0	0	0	0	0	0.2	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
18	0	0	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	99.4	0	0	0.3	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
21	0	0.5	0	0.2	0	0	0	0.2	0	0.5	0	0	0.7	0.2	0	0	0	0	0	0	97.7	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	99.5	0
23	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99.7

Conclusions

- Exploring the temporal saliency maps addresses the problem of extracting an accurate representation for HAR.
- The HOG-S descriptor for HAR increases the discrimination for human actions.
- The proposed method outperforms the state-of-the-art by 2:7% with DHA, 1% with KTH and it is comparable in the case of Weizmann.