RVM-based human action classification through Gabor and Haar feature extraction

S. Maheswari* and P. Arockia Jansi Rani

Department of Computer Science and Engineering,
Manonmaniam Sundaranar University,
Tirunelveli 627012, Tamilnadu, India
Email: jan20mahi91@yahoo.com
Email: jansi_msu@yahoo.co.in
*Corresponding author

Abstract: Human action recognition plays a vital role in surveillance applications. Human action recognition is motivated by some of the applications such as video retrieval, video surveillance systems, human robot interaction, to interact with deaf and dumb people etc. The aim is to analyse the role of Adaboost in the process of recognising the human action by extracting the motion features using optical flow. Adaboost is a supervised learning method used to select the subset of frames with most discriminatory motion features. Saliency point computation is performed to assign a measure of interest to each visual unit. Mean shift algorithm is then used for tracking the objects. Gabor feature is the global feature that includes more detailed information of frequency and orientation. Haar feature is used to show the variation in the pixel. Relevance vector machine classification gives a probabilistic output through Bayesian inference. The proposed system reduces the computation time and provides a higher recognition rate in comparison with existing gentle boost-based recognition system.

Keywords: optical flow; saliency map; mean shift localisation; adaboost; RVM classification.

Reference to this paper should be made as follows: Maheswari, S. and Arockia Jansi Rani, P. (2016) ‘RVM-based human action classification through Gabor and Haar feature extraction’, Int. J. Computational Vision and Robotics, Vol. 6, Nos. 1/2, pp.54–64.

Biographical notes: S. Maheswari is pursuing her PhD in Computer Science and Engineering in Manonmaniam Sundaranar University, Tirunelveli, India. She received her BE in Computer Science and Engineering in Dr. G.U. Pope College of Engineering, Tuticorin, India in 2011 and ME in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli, India in 2013.

P. Arockia Jansi Rani is an Assistant Professor in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India. She received her PhD in CSE from Manonmaniam Sundaranar University, Tirunelveli, India in 2012. She is in teaching profession for the last ten years. Her area of interest includes digital image processing, neural networks and data mining. She has presented her research papers at various national and international conferences.
1 Introduction

Action is any meaningful movement of the human and it is used to convey information or to interact naturally without any mechanical devices. Human action recognition has attracted a great deal of attention due to its potential value and wide usage in a variety of areas, such as video search and retrieval, intelligent surveillance systems, healthcare systems and human-computer interaction. Typically, human action recognition is based on either global or local feature extraction.

Computer vision systems are far behind the capabilities of human vision. For instance, video search in large scale database archives is currently feasible with costly manual annotation. Web search engines commonly rely on textual data, such as description or tags in order to retrieve relevant videos. Since recognising actions in videos is a challenging problem, a lot of approaches have been considered. The human action recognition is necessary in shop surveillance, city surveillance, airport surveillance and in other places where security is the prime factor. The surveillance area covers applications for tracking one or several subjects and detecting unusual behaviour.

Monitoring activities of daily living is gaining interest because of the growing population of elderly people and their need for care. A system that contributes to the safety of elderly at home is therefore more than needed. The human behaviour analysis is essential for the medical professionals to detect emerging physical and mental health problems, before they become critical particularly for elderly.

Section 2 discusses the related work found from the literature. The overall process is described in Section 3. The preprocessing and the optical flow details are described in Section 3.2. The feature selection based on boosting and mean shift localisation which is used in order to extract the most probable hypotheses concerning the spatiotemporal localisation of an activity is described in Section 3.3. Gabor and Haar feature extraction is included in Section 3.4. Section 3.5 describes the relevance vector machine (RVM) process. Section 4 presents the experimental results and finally Section 5 concludes the paper.

2 Related work

Oikonomopoulos et al. (2006) focused on the problem of human action recognition by using spatiotemporal events that are localised at points that are salient in both space and time. The spatiotemporal points are detected by measuring the variations in the information content of pixel neighbourhoods not only in space but also in time. The classification scheme is based on RVMs and on the chamfer distance measure. The classification results are presented for two different types of classifiers, displaying the efficiency for the representation in discriminating actions of different motion classes.

Bay et al. (2008) introduced speeded up robust feature (SURF) method for activity recognition. It is a robust local feature detector, first presented by Bay et al. (2006), that can be used in computer vision tasks like object recognition or 3D reconstruction. It is partly inspired by the SIFT descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. SURF is based on sums of 2D Haar wavelet responses and
makes an efficient use of integral images. It uses an integer approximation to the
determinant of Hessian blob detector.

Weinland et al. (2011) presented an overview and categorisation of the approaches
used. Feature extraction, action learning, action segmentation, action classification are the
stages involved in action recognition. Feature extraction is used to extract the postures
and the motion cues from the video. Action learning is the process of learning statistical
models from the extracted features. The statistical models are used to classify new feature
observations. Action segmentation is used to cut the streams of motions into a single
action instances that are consistent to set of initial training sequences used to learn the
models.

Tao et al. (2006) developed human gait recognition to validate the GTDA. The
averaged gait images are utilised for gait representation. For image understanding and
object recognition, three different Gabor functions-based image representations are
developed. They are:

1. the GaborD representation is the sum of Gabor filter responses over directions
2. GaborS is the sum of Gabor filter responses over scales
3. GaborSD is the sum of Gabor filter responses over scales and directions.

Deepak et al. (2011) recognises the human actions by tracking the selected object over
the consecutive frames of grey scale image sequences. Initially the background motion of
the input video stream is subtracted and its binary images are constructed. The object
which is needed to be monitored is selected by enclosing the required pixels with
bounding rectangle and by using spatiotemporal interest points. The obtained results after
subtraction are compared with the selected threshold value to predict the type of human
action using linear prediction technique.

Hofmann et al. (1999) eloped probabilistic latent semantic analysis (PLSA) method,
also known as probabilistic latent semantic indexing (PLSI, especially in information
retrieval circles) is a statistical technique for the analysis of two-mode and cooccurrence
data.

Chen et al. (2007) developed a hybrid boost learning algorithm for multipose face
detection and facial expression recognition. The major contribution in this method is
weak hybrid classifier selection based on Harr features and Gabor features.

Carlos et al. (2009) developed a novel unsupervised learning method for human
action recognition. This is achieved by using latent topic models such as PLSA model
and latent Dirichlet allocation. This approach can handle noisy feature points arose from
dynamic background and moving cameras due to the application of probabilistic models.

Liu et al. (2013) developed a method for human action recognition based on boosted
key-frame selection and correlated pyramidal motion feature representations. To
demonstrate generalisability, this method has been systematically tested on a variety of
datasets and shown to be more effective and accurate for action recognition.

Rapantzikos et al. (2011) developed a method to compute visual saliency from video
sequence by counting the actual spatio temporal nature of video. The visual input is
represented by a volume in space-time. The visual input is decomposed into a set of
feature volumes in multiple resolutions. Saliency computation is the problem of assigning
a measure of interest to each visual unit. The saliency measure is produced by taking actual spatiotemporal evolution of the input. This model shows the visual saliency based on spatiotemporal analysis.

Ikizler et al. (2007) developed a ‘bag-of-rectangles’ method for representing and recognising human actions in videos. In this method, each human pose in an action sequence is represented by oriented rectangular patches extracted over the whole body. Then, spatial oriented histograms are formed to represent the distribution of these rectangular patches. In order to carry out this approach four different methods are proposed. They are namely,

1. frame by frame voting, which recognises the actions by matching the descriptors of each frame
2. global histogramming, which extends the idea of Motion Energy Image by rectangular patches
3. SVM classifications
4. adaptation of dynamic time warping on the temporal representation of the descriptor.

Yogameena et al. (2009) introduced real-time video surveillance system capable of classifying normal and abnormal actions of individuals in a crowd. The abnormal actions of human such as running, jumping, waving hand, bending, walking and fighting with each other in a crowded environment are considered. First the foreground blobs are detected using adaptive mixtures of Gaussians which is to be robust to illumination changes and shadows. Then projection is applied to segment an individual in the crowd.

3 Methodology

The proposed RVM-based human action classification through Gabor and Haar feature extraction is described in this section.

In this system, the input video is split into frames and then it is pre-processed to improve the brightness of the image. In any action recognition system, a pre-processing step is carried out to remove the noise caused because of illumination effects, blurring, false contour, etc. Saliency point is detected to know the variation in object in terms of both space and time. This saliency point shows the points in space-time where the motion (walk) abruptly changes the direction, such as stopping or the starting. Boosting method is used for object categorisation. The mean shift localisation is used to detect the variations in the object. Mean shift is a tracking algorithm based on the external features and track the objects in real time. Gabor feature is the global feature that includes more detailed information of frequency and orientation. Haar feature is used to show the variation in the pixel. The RVM classification is used to classify the human action. This process is described in Figure 1.
3.1 Selecting the input and frame split

The input videos are taken from Weizmann dataset. The input videos are recorded in a homogeneous background with a static camera. The Weizmann dataset is downloaded from the website http://www.wisdom.weizmann.ac.il/vision/SpaceTimeActions. The properties of the input videos are

- type: VLC media file (.avi)
- dimensions: $180 \times 144$
- resolution: $180 \times 144$
- frame rate: 25 frames per second.
The input video is split into frames. Frame rate is expressed as frames per second. Frame rate means the number of frames that are projected or displayed per second.

3.2 Preprocessing and optical flow extraction

The preprocessing is performed to improve the brightness of the image. The brightness of the image is improved to provide better results. The Gaussian filter is used to remove the noise in the input. Butterworth highpass filter is used to highlight the fine details of an object. The optical flow is used to extract the motion features. The optical flow is a vector which represents the object velocity in the images. The optical flow is the pattern of apparent motion of objects, surfaces and edges in a visual scene caused by the relative motion between the camera and the scene. It is used to detect the direction of the moving objects.

3.3 Saliency point detection and adaBoost filtering

Saliency map is used to detect the variation of the object in terms of both space and time. Saliency is used for feature point detection in videos and incorporate colour and motion apart from intensity. This saliency point shows the point in space-time where the motion (walk) abruptly changes the direction, such as stopping or the starting.

Boosting method is used for object categorisation. The role of adaboost is feature selection on very large set of features. AdaBoost is an algorithm for constructing a ‘strong’ classifier as linear combination of ‘simple’ ‘weak’ classifiers $h_t(x)$.

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

where $h_t(x)$ is ‘weak’ or basis classifier, hypothesis, ‘feature’ and $H(x) = \text{sign}(f(x))$ is ‘strong’ or final classifier/hypothesis.

3.3.1 Mean shift localisation

The mean shift algorithm can be used for visual tracking. Mean shift (Comaniciu et al., 2002) is a tracking algorithm based on the external features and track the objects in real time.

3.4 Gabor and Haar feature extraction

Gabor feature is the global feature that includes more detailed information of frequency and orientation. Gabor function is used for visual information processing. A Gabor function is the product of an elliptical Gaussian envelope and a complex plane wave.

The core of Gabor filter-based feature extraction is the 2D Gabor filter function $(\psi(x, y))$ (Kamarainen et al., 2006):

$$\psi(x, y) = \frac{f^2}{\pi \sigma_\eta} \exp\left(\frac{-f^2 (x^2/\sigma_x^2 + y^2/\sigma_y^2)}{2\sigma_\eta^2}\right) e^{j2\pi \Psi / \sigma_\eta}$$

$$x' = x \cos \theta + y \sin \theta$$
In the spatial domain [equation (1)] the Gabor filter is a complex plane wave (a 2D Fourier basis function) multiplied by an origin-centred Gaussian $f$. $f$ is the central frequency of the filter, $\theta$ the rotation angle, $\gamma$ sharpness (bandwidth) along the Gaussian major axis, and $\eta$ sharpness along the minor axis (perpendicular to the wave). In the given form, the aspect ratio of the Gaussian is $\eta/\gamma$. This function has the following analytical form in the frequency domain

$$\psi(u, v) = e^{-\frac{1}{f^2}(u^2(w-f)^2 + v^2\eta^2)}$$

(3)

$$u' = u \cos \theta + v \sin \theta$$

$$v' = -u \sin \theta + v \cos \theta$$

In the frequency domain [equation (2)] the function is a single real-valued Gaussian centred at $f$. The Gabor filter in (1) and (2) is a simplified version of the general 2D form devised by Daugman (1985) from the Gabor’s (1946) original 1D ‘elementary function’. The simplified version enforces a set of filters self-similar, i.e., scaled and rotated versions of each other (‘Gabor wavelets’), regardless of the frequency $f$ and orientation $\theta$.

Haar feature is used to show the variation in the pixel. Haar features can be computed rapidly by using an intermediate representation called the integral image (Viola et al., 2001) as

$$ii(x, y) = \sum_{x' < x, y' < y} i(x', y')$$

(4)

where $ii(x, y)$ is the integral image and $i(x, y)$ is the original image.

3.5 RVM classification

A RVM is a machine learning technique that uses Bayesian inference to obtain parsimonious solutions for regression and classification (Tipping and Michael, 2001). The RVM has an identical functional form to the support vector machine, but provides probabilistic classification. It is actually equivalent to a Gaussian process model with covariance function:

$$k(x, x') = \sum_{j=1}^{N} \frac{1}{\alpha_j} \Psi(x, x_j)(x', x_j)$$

(5)

where $\Psi$ is the kernel function (usually Gaussian), and $x_1, \ldots, x_N$ are the input vectors of the training set. The main advantage of RVM is that it allow sparse sets of highly relevant features or training examples to be selected for training of separation functions.

3.6 Training and testing phase

Action instance for each run, jump, walk, bend and wave action is considered for analysis. To create the training set the subset of action instances for each class (run, walk etc.) are selected manually. The training sets are registered in terms of both time and
frame rate. The training sets have different time and frame rate for different human actions. Each action instance is trained. During training a threshold value is obtained for each action.

For testing, all action instances for each walk, run, jump, bend and wave actions from the Weizmann dataset is given as input and a threshold value is obtained during the testing phase. If the value obtained in testing phase falls within the threshold values of any class of action that is determined during the training phase, then that class of action is recognised as the output. Figure 2 describes the implementation results of the proposed work for the input action ‘walk’.

**Figure 2**  (a) Input video (b) Frame split (c) Preprocessing (d) Optical flow (e) High pass filter (f) Saliency map (g) Adaboost (h) Mean shift localisation (i) Gabor feature (j) Haar feature (k) RVM classification (l) Recognition (see online version for colours)
4 Experiments and discussions

Weizmann action dataset is used for analysing the results. The performance of the system is analysed by testing different human actions. A common quantitative analysis is performed to assess the overall performance of recognition process. Recognition rate is calculated on the basis of number of true matches and false matches.

4.1 Recognition rate

- number of correctly matched actions = 19
- total number of tested actions = 20

\[
\text{Recognition rate} (\%) = \frac{\text{Number of correct match}}{\text{Total number of tested actions}} \times 100
\]

- recognition Rate = (19/20)*100 = 95%.

Table 1 Experimental results for recognition rate

<table>
<thead>
<tr>
<th>Actions</th>
<th>No. of actions tested</th>
<th>Recognition rate in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>Run</td>
<td>5</td>
<td>80</td>
</tr>
<tr>
<td>Bend</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>Wave</td>
<td>5</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 3 Chart for recognition rate (see online version for colours)

Table 2 Experimental results illustrating computation time

<table>
<thead>
<tr>
<th>Input videos</th>
<th>Adaboost (time taken in seconds)</th>
<th>Gentleboost (time taken in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>42.4279</td>
<td>43.6948</td>
</tr>
<tr>
<td>Run</td>
<td>41.0041</td>
<td>43.9780</td>
</tr>
<tr>
<td>Bend</td>
<td>41.0155</td>
<td>43.0155</td>
</tr>
<tr>
<td>Jump</td>
<td>42.1467</td>
<td>42.7319</td>
</tr>
<tr>
<td>Wave</td>
<td>41.3623</td>
<td>42.6880</td>
</tr>
</tbody>
</table>
The time taken for adaboost is compared with the time taken for gentle boost (Oikonomopoulos et al., 2011). The results are tabulated in Table 2. The goal of gentle boost is to select feature ensembles that appear with high likelihood in the positive and with low likelihood in the negative examples. Time taken for the proposed adaboost-based recognition system is lesser than that of the gentle boost-based recognition system.

### 4.2 Confusion matrix

To assess the accuracy of an image classification, it is necessary to create a confusion matrix. In a confusion matrix, the classification results are compared to truth information. The strength of a confusion matrix is that it identifies the nature of the classification errors.

<table>
<thead>
<tr>
<th>Actions</th>
<th>Walk</th>
<th>Wave</th>
<th>Run</th>
<th>Bend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Wave</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Run</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Bend</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### 5 Conclusions

In this system localisation and recognition of actions is performed. The localisation is done in terms of both space and time. This system uses boosting method and mean shift localisation. Boosting method is used for object categorisation. Adaboost is a type of boosting method. Adaboost method determines the variation points in the entire input video. Mean shift is a tracking algorithm that tracks the objects in video. Gabor feature is the global feature that includes more detailed information of frequency and orientation. Haar feature is used to show the variation in the pixel. RVM classification is used to classify the human actions. Then the action is recognised. The results are presented for the Weizmann dataset. Other datasets like KTH dataset, HOHa dataset can also be used. This system provides an efficient result in recognising the actions.

### References


