An Integrated System for Human Action Recognition from Video using Hidden Markov Model

Palwasha Afsar  
ALGORITMI Research Center  
Department of Information Systems  
University of Minho  
4804-533 Guimaraes, Portugal  
Email: palo_afsar77@yahoo.com

Paulo Cortez  
ALGORITMI Research Center  
Department of Information Systems  
University of Minho  
4804-533 Guimaraes, Portugal  
Email: pcortez@dsi.uminho.pt

Henrique Santos  
ALGORITMI Research Center  
Department of Information Systems  
University of Minho  
4804-533 Guimaraes, Portugal  
Email: hsantos@dsi.uminho.pt

Abstract—In this paper, we present an integrated system for real-time automatic detection of human actions from video. The proposed approach uses the boundary of humans as the main feature for recognizing actions. Background subtraction is performed using Gaussian mixture model. Then, features are extracted from silhouettes and Vector Quantization is used to map features into symbols (bag of words approach). Finally, actions are detected using the Hidden Markov Model. The proposed system was validated using a newly collected real-world dataset. The obtained results show that the system is capable of achieving robust human detection, in both indoor and outdoor environments. Moreover, promising classification results were achieved when detecting two basic human actions: walking and sitting.

Keywords—Video data, Human action, Hidden Markov model, Video analysis, Video databases

I. INTRODUCTION

Automatic human behavior from video is a key tool that can be used for decision support in distinct applications, such as surveillance, marketing or health Care. Recent developments in Information and Communication Technologies have increased the interest in this area and currently it is quite common to have digital cameras for monitoring human activities. The area of human action recognition can be linked to many other disciplines that analyze human motion from videos. The recognition of human basic actions (e.g., walking, sitting, jumping, waving hands) from monocular images and videos is an important task in several applications such as human computer interaction, video content retrieval and surveillance. Several methods have been proposed for human action recognition. A detailed survey can be found in [1]. Recently many research work focus on recovery of human pose, which is considered an important step of view-invariant human pose recognition. However, 3D pose reconstruction from a single viewpoint is a challenging problem because of the large number of parameters and the ambiguity caused by perspective projection [2].

For some activities, there is a huge change in performance. For instance, walking movements can vary in rate and step length. Likewise, there are anthropometric contrasts between individuals. Comparable perceptions can be made for different actions, particularly for non-cyclic activities or activities that are adjusted to environment (e.g., guiding towards a certain location, avoiding obstacles while taking a walk). A decent human activity recognition methodology ought to have the capacity to sum up over varieties inside of one class and recognize activities of distinctive classes. For expanding quantities of action classes, this will be additionally difficult as the overlap between classes will be higher. In some domains, dissemination over class labels may be a suitable option. The environment in which the action execution is performed is a vital source in the recording. Individual restriction may demonstrate harder in dynamic or cluttered environments. Besides, parts of the individual may be impeded in the recording. Lighting conditions can further impact the presence of the individual.

Observing the same action from different angles can also lead to different observations. Assuming a known camera perspective limits the utilization to static cameras. At the point when various cameras are utilized, viewpoint issues and issues with impediment can be solved, particularly when perceptions from numerous perspectives can be consolidated into a steady representation. Dynamic backgrounds increase the complication of finding an individual in the image and vigorously watching the motion. At the point when utilizing a moving camera, these difficulties turn out to be considerably harder. In vision-based human action recognition, every one of these issues ought to be tended to expressly.

Hidden Markov Model has been successively used for speech recognition in training based approaches. For action recognition, HMM transforms the problem into pattern recognition. [3] was the first to use HMM for action recognition. In their work, HMM was used to recognize six different tennis strokes among three players. Some recent works [2], [4], [5], [7] have also shown that HMM performs well for action recognition. HMM is a stochastic state transition model that model action recognition by training. For an action recognition, the HMM which best matches an action is chosen. It achieves high recognition rates and requires low processing time.

In this paper, we propose an action recognition method based on HMM using the Bag of Words method (with boundary of humans as the main feature). Time-sequential images of human action are transformed into feature vectors. All of these feature vectors are stored in a codebook where each symbol corresponds to an action by using Vector Quantization (VQ). For the training phase, the model parameters of HMM are tuned well for description of an action to achieve high
results. The Model, which matches to a symbol, is selected as a recognized entity. The framework of the approach can be seen in Figure 1.

The paper is organized as follow: section 2 contains related work on human action recognition. In section 3, a brief introduction of Hidden Markov Model is given. Section 4 covers the dataset created for this research. The methodology of the system is defined in section 5. Experimental results are discussed in section 6 and finally section 7 outlines the conclusion and future work.

II. RELATED WORK

Researchers have explored different methods for action recognition. In 1975, Johansson’s experiment shows that human can recognize activity with extremely compact observers [6]. However, in recent studies [7] proposed to use star skeleton for extracted from silhouettes for motion analysis. They used Hidden Markov based methodology for learning actions. Star skeleton is a fast skeletonization technique achieved by connecting centroid of the target object to contour extremes. For using star skeleton as a feature, the feature vector is defined as a five dimensional vector. An action is composed of a series of star skeletons and hence are transformed into feature vector sequences. A posture codebook is designed where a feature vector is transformed into a symbol sequence. This symbol sequence is used by HMM for recognizing an action. The algorithm achieves robust recognition rates.

[8] Used joint position and pose angles for recognition of simple action like walking, standing and waving hands. The high dimensional 3D data joint space data is decomposed into a feature space where each feature corresponds to motion of single or multiple joints. Hidden Markov model (HMM) is used for learning of actions together with Adaboost classifier. The results obtained on 20 actions shows the effectiveness of the approach.

[4] work focus on posture classification based on projection histogram with HMM for assuring temporal coherence of the posture. For occlusion handling, the single camera postures are passed to multi-camera system to give robust classification. In contrast to [4], a stationary camera was used by [9] for recognizing 15 different continuous human activities in real-time. The activities recognized includes waving hands, raising hands, sitting down and bending down. Activities were described as a continuous sequence of discrete hand postures, derived from affine invariant descriptor. Support Vector Machine was used for classification.

[5] Developed a video surveillance system capable of recognizing human posture from video. The system consists of two modules: human detection and posture classification. In human detection module, human blob are extracted from the background using an adaptive background module based on mixture of Gaussians. For the variation in human pose, pseudo 2D hidden Markov models (P2HMM) is used for representing and recognizing human postures based on its 2D elastic matching property. For classification, observation sequences are extracted from current image and human blobs are classified as the human posture with the highest likelihood.

[10] proposed a system for recognizing human daily life activities. The method utilizes a hierarchical structure of actions and describes it as a tree. The actions are modeled using Hidden Markov model, which output action as a time series feature vectors. The recognition process starts at the root and moves on to the recognition of the child nodes. Hierarchical recognition offers several advantages like simplification of low-level models, recognition of various levels of abstraction, and response to novel data by decreasing the degree of data details. Results show that their proposed algorithm can recognize some of the actions.

A feature-based bottom-up approach is proposed by [3]. HMM was applied to one set of time-sequential images that were transformed into feature vector. These features were converted into symbols using Vector Quantization. The parameters of HMM were optimized well for training actions categories. The HMM which best matches an action category was chosen. They achieve recognition rates higher than 90% on real-time sequential images of sport sequences.

The posture of human is considered important for recognition of human actions. Inspired from [7], we propose an action recognition algorithm using HMM representation of postures extracted from human silhouettes. Posture contour is normalized to achieve a standard size. The observation vectors are extracted from these silhouettes and Vector Quantization (VQ) is used to map features into symbols. A posture codebook representing actions is build and each symbol correspond to a different action stored in the codebook. The existing research works aimed to detect human actions captured against a uniform background with little or no occlusion. Also the test sets often contain just a single subject performing an action. Our method was designed to achieve more robust results in more challenging real-world environments, in the presence of distinct light conditions, with severe occlusion and quite distinct weather conditions, including heavy wind and rain.

III. HIDDEN MARKOV MODEL (HMM)

Hidden Markov Model (HMM) is a statistical Markov Model in which the system being model is assumed to be a Markov process with hidden states as shown in Figure 2. An HMM can also be modeled as dynamic Bayesian network. Each state of the HMM outputs a state transition symbol.
While we can observe the output symbol of HMM, we cannot observe the states of HMM. The formal definition of HMM is as follow:

$$\lambda = \{A, B, \pi\}$$

$S$ is the state alphabet set

$$S = \{s_1, s_2, s_3, ..., s_n\}$$

and $V$ is the observation set

$$V = \{v_1, v_2, v_3, ..., v_n\}$$

HMM performs action recognition in two parts: training the model and computing the probability that the observation sequence was generated by the model $\lambda$. Each model is trained so that it is able to generate the symbols for training data.

IV. Dataset

For this research, a purely new video real-world dataset was created at University of Minho. The dataset was recorded from 08h00 to 19h00 during a total of 7 working days. The dataset is challenging as it includes both indoor and outdoor settings under distinct light conditions (e.g., morning vs afternoon, sunny and cloudy weather). Also, in some days there was heavy rain and wind, which particularly affected the outdoor setting (e.g., moving trees and plants due to wind, usage of umbrellas by humans). For capturing the dataset, two cameras HIK Vision and IR Network were installed in order to capture indoor and outdoor environments, as shown in Figure 3. The cameras were installed behind the window glass, therefore, a reflection is always present. The angle of the camera is set in such a way that it always has two or three trees or plants framing the scene. This has been done to make the dataset challenging and test the efficiency of the proposed algorithm. To substantially save disk space, the cameras were configured with motion detection, thus storing only videos when movement is present.

The environment for capturing the video dataset was not controlled. In effect, it corresponds to real life moments related with the daily university activities. Both the outdoor and indoor areas are related with large corridors, where students, professors, researchers and university staff pass by (e.g., entering, leaving or taking a break). Figure 4 shows examples of these two environments, namely (from top left to bottom right): outdoor, outdoor, outdoor, indoor, indoor, indoor and outdoor. The dataset is comprised of hundreds of small videos (with a few seconds to a few minutes each) that correspond to 32 GB. Several actions are present in the dataset, includes walking, running, group interactions, talking on mobile, drinking coffee, standing and shaking hands.

V. Methodology

In current video surveillance system, majority of work is devoted to background modeling and object tracking. However, the recognition of human activities has not gained enough attention. The work in this paper focuses on two modules i.e., human detection in presence of occlusion and classification of activities e.g., walking, running, group interactions, talking on mobile, drinking coffee, standing and shaking hands.

VI. Overview of the System

The detection of moving object is an important component for many computer applications e.g., including action recognition, surveillance systems etc. The system proposed
in this paper is based on two main modules i.e., Human detection module and action recognition module. The human detection module uses a background subtraction algorithm based on Gaussian mixture model. The algorithm compares a color or grayscale video frame to a background model to determine whether individual pixels are part of the background or the moving object. A foreground mask is then computed. Morphological operators are applied to the resulting foreground mask to eliminate any noise present. Blob analysis is performed to look for connected regions or pixels that most likely corresponds to moving objects.

For action recognition, all subsequent frames representing an action are collected. Human boundary is considered one of the best ways to represent an action. Posture contour is extracted from the silhouettes of human and normalization is performed to achieve a standard size. As the boundary of human consists of many neighboring pints that are almost the same, Principal Component analysis is performed for dimensionality reduction. The observation vectors are extracted from these silhouettes and Vector Quantization (VQ) is used to map features into symbols. A posture codebook representing actions is build and each symbol correspond to a different action stored in the codebook. An extracted feature vector is mapped to a codebook symbol and the output is hence a sequence of posture symbols. Hidden Markov model is used for training different actions that optimize model parameters and recognition is performed by choosing maximum likelihood.

A. Background Segmentation

The main concept behind background subtraction is to subtract the image from a reference image that models the background scene. The main steps of the algorithm are as follow:

1) Background modeling: constructs a reference image representing the background.

2) Threshold selection: determines appropriate threshold value to be used in subtraction operation to achieve a desired detection rate.

3) Subtraction operation: It is also called pixel classification and is used for classification type of a given pixel i.e., the pixel is part of the background or the moving object.

B. Blob detection

In blob detection, all of the foreground pixels in are grouped into disconnected blobs. A blob can represent a) no object b) part of moving object c) a single moving object with possible foreground trail and d) multiple moving objects. The first one refers to the presence of ghosts that are caused by shadows. In this work our aim is to remove the entire ghosts to achieve better classification results. The second one is due to aperture problem that are produced when an object is starting to move. These partial blobs are maintained because eventually they will grow to full blobs. Majority of blobs fall into third category. The last case of multiple objects occurs when people start walking in a group and are close to one another. These large blobs finally split into multiple small blobs when a group is dispersed.

Fig. 5: Examples of preprocessed frames with boundary of human regions for the walking (left) and sitting (right) actions

VII. EXPERIMENTAL RESULTS

All experiments were conducted using a common MacBook laptop with a 2.4 GHz Intel Core i5 processor. The proposed system was implemented as a prototype, using code written in Matlab1 (version 2014b). To evaluate the performance of our approach, the algorithm was tested on two types of basic actions: walking and sitting. We selected these particular actions since they are more prevalent in the dataset when compared with other actions (e.g., talking on mobile, drinking coffee). Figure 5 shows examples of the selected actions. A manual analysis was conducted in order to label the walking and sitting actions on all frames of the collected videos. The videos were recorded by HIK Vision and IR Network cameras at 1080x1920 HD pixel resolution. The size of the image was reduced to 576x720 resolution and preprocessing was performed to remove any kind of noise present. The holdout train and test split validation method was used to measure the action detection generalization capabilities of the proposed system [11]. Thus, the dataset was randomly split into training, with 234 examples, for training the human action recognition system (e.g., HMM training and codebook storage), and test, with 500 examples. In both training and test datasets the walking and sitting classes are balanced, i.e., each action contains 50% of the examples. This holdout method was applied under two modes: batch and iterative. The former mode performs an action prediction for the whole test frames under a single pass, thus the internal state of the HMM model is not changed during this process. The latter mode detects the action for each single frame under a sequential procedure with 500 iterations, where each iteration changes the HMM internal state.

Our system was able to detect human in cluttered environments. As shown in the examples of Figure 6, for both outdoor (top left frame) and indoor (middle and bottom left frames) settings, the Human Detection module correctly identified the human body contours (right frames). The overall system results are measured in terms of the recognition predictions for the walking and sitting actions, as shown by the confusion matrices of Tables I and II. For both batch and iterative modes, the proposed system achieved high quality action detection results, with classification accuracies higher than 90% for both walking and sitting classes. In terms of computation effort, the system prototype required 188 seconds for the training and 94 seconds

1http://www.mathworks.com/products/matlab/
for the testing, resulting on an average of 0.188 seconds per each test image detection. This is also an interesting result, confirming that the developed prototype is capable of real-time action detection. We note that the written Matlab code can be further enhanced (e.g., conversion into the C language) in terms of its computational effort.

TABLE I: Confusion matrix for the batch mode and test data (the true walking and true sitting accuracies are in bold; the global accuracy is 97%)

<table>
<thead>
<tr>
<th></th>
<th>Sitting</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Walking</td>
<td>6%</td>
<td>94%</td>
</tr>
</tbody>
</table>

TABLE II: Confusion matrix for the iterative mode and test data (the true walking and true sitting accuracies are in bold; the global accuracy is 95%)

<table>
<thead>
<tr>
<th></th>
<th>Sitting</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>Walking</td>
<td>9%</td>
<td>91%</td>
</tr>
</tbody>
</table>

VIII. CONCLUSION

In this paper, we present an integrated system that combines state-of-the-art methods for automatically detection human actions from videos. In particular, we focus on human posture and use the boundary of human as a main feature of recognizing actions. For obtaining the foreground image, background subtraction was performed using Gaussian of mixture model. Then, feature vectors are extracted from silhouettes and Vector Quantization (VQ) is used to map features into symbols. Finally, action recognition is done using Hidden Markov Model (HMM). Moreover, a newly real-world video dataset was collected with indoor and outdoor scenarios and distinct weather conditions. The dataset was then manually labeled in terms of two prevalent actions, namely walking and sitting.

Using an holdout validation method, with 234 images for training and 500 for testing, our approach has shown promising action detection results, with global classification accuracies of 97% (batch mode) and 95% (iterative mode). Moreover, the proposed system can detect the walking and sitting actions in real-time. In future work, we aim to check the robustness of the proposed work by including more actions (e.g., talking on mobile, drinking coffee). We will be also improving the human action detection system by extracting more advanced image features, such as human skeletons.

ACKNOWLEDGMENT

This work is funded by the Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia) within the Project Scope UID/CEC/00319/2013 and research grant SFRH/BD/84939/2012.

REFERENCES