Recognition of Human Actions in Video

Hetal Shah
Dept. of Computer Engineering
B.V.M. Engineering College
Anand, Gujarat, India
E-mail: hetal189@gmail.com

N. C. Chauhan
Dept. of Information Technology
A. D. Patel Institute of Technology
Anand, Gujarat, India
E-mail: narendracchauhan@gmail.com

***Abstract*— Recognition and analysis of Human actions is an important task in the area of computer vision. There are many applications of this research which include surveillance systems, patient monitoring systems, human performance analysis, content-based image/video retrieval/storage, virtual reality and a variety of systems that involve interactions between persons or interactions between person and devices, etc. The need for such systems is increasing day-by-day, with the increase in number of surveillance cameras deployed in public spaces. Automated systems are required that can detect, categorize and recognize human activities and request the human attention only when necessary.**

In this paper, important steps of such a system are described that can robustly tracks human in various environments and recognizes their actions through image sequences acquired from a single fixed camera. The overall system consists of major three steps: blob extraction, feature extraction, and human action recognition. Given the sequence of images, a statistical method is demonstrated to extract the blobs and to remove the shadows and highlights in order to obtain a more accurate object silhouette. Shape context is used to extract features in next step and at-last human action is recognized using neural network.

Keywords- surveillance, blob extraction, feature extraction, shape context, neural network.

*****

I. INTRODUCTION

The analysis of human activity by a computer is gaining more and more interest. In recent years, this problem has acquired the attention of researchers from industry, academia, security agencies, consumer agencies and the general populace due to its wide span of applications in different areas.

The Application areas [1] can roughly be grouped under three major titles: surveillance, control, and analysis.

1) *Surveillance:* The surveillance area includes applications where one or multiple subjects are being tracked over time and monitored for special actions. For example the surveillance of a parking lot can be used to detect the theft of car.
2) *Control:* The control area covers applications where the captured motion is used to provide controlling functionalities. For example interface to games, virtual environments, or controlling remotely located implements.
3) *Analysis:* This application area is related with the detailed analysis of the captured motion data. This may be used for clinical studies (diagnostics) of orthopedic patients or to help athletes understand and improve their performance.

According to Turaga, et al. [2] action or activity recognition system can be viewed as a hierarchy of steps, starting from a sequence of images to recognition of complex activities. The steps involved in such system are as follows:

1. *Acquire input:* video or sequence of images
2. *Perform low-level execution:* Extraction of concise low-level features (such as background foreground separation)
3. *Perform mid-level execution:* action descriptions from low-level features are used to perform action-recognition
4. *Perform high-level execution: to get* semantic interpretations from primitive actions and get the attention of operators/humans.

II. BACKGROUND STUDY AND REVIEW

There has been a large amount of research in the area of human activity analysis. Human action recognition can be considered to be part of human activity analysis. A detailed survey of several recent techniques of human action recognition is found in [1] and [2]. For the purpose of clarity, the authors make the distinction between an action and an activity.

'Actions' are referred as simple motion patterns which are executed by a single person and lasts for short durations of time. Examples of actions include bending, walking, swimming, running, hand waving, boxing,etc. In this work the emphasis is on recognizing the simple actions. 'Activities' are referred as complex sequence of actions performed by several humans who could be interacting with each other in a constrained manner and are characterized by much longer time durations. Examples of activities are two persons shaking hands, a football team scoring a goal, a coordinated bank robbery, violation of traffic rues, etc. In

this paper, the terms action and activity are used interchangeably to refer the motion patterns such as walking, jumping etc.

In general, activity recognition task is done on video sequences instead of still images, because the images are not able to provide sufficient information to reliably recognize an activity/action. In most research works, activity recognition is done in three stages as shown in Fig.1: the blob extraction stage, where background is subtracted and shadows and highlights are removed, the feature extraction stage, where the data is processed to extract only relevant distinct features, and the action recognition stage, where various models or classifiers may be used to determine what action has been performed.

Figure 1. Stages of Action Recognition

The review of work done on each of these phases is described below:

A. Blob Extraction

Identifying moving objects from a video sequence is a very critical task in many computer-vision applications [3, 4, 5]. This task may be accomplished by extraction of blob from video frame sequence. The blob extraction process can be viewed as composed of two sub processes: the important process is background subtraction (or foreground extraction), and it is followed by shadow removal which can improve the performance in later phases.

Background subtraction is a computational process in which foreground objects are extracted from the image. A foreground object can be termed as an object of attention, which reduces the amount of data for the task under consideration.

A good background subtraction algorithm must be robust against illumination changes; it should not detect non-stationary background objects and shadows which are cast by moving objects. A good background model should also quickly response to changes in background, and adapt itself to accommodate changes occurring in the background. It should also have a good foreground detection rate, and the time required to process background subtraction should be real-time [3].

A detailed survey of background subtraction is provided in [12]. Some of the techniques are, subtraction of a previously captured background frame, application of IIR filter to video frames which incorporates the background model objects that become stationary, use of Gaussian mixture model [4] that models each pixel as a mixture of Gaussian and based on the persistence and the variance of each Gaussian of the mixture, it determines which Gaussian may correspond to background colors and which to foreground.

Normally, the foreground obtained from background subtraction method contains shadow as part of foreground which cause the consequent processes, tracking and recognition to fail. Hence, it needs to be removed before proceeding to the further steps. Use of statistical method [5] removes shadows as well as highlights to good extent in addition to background subtraction.

The resulting silhouettes are used in several different ways for activity recognition.

B. Feature Extraction

In computer vision and image processing the concept of feature is used to denote a piece of information which is relevant for solving the computational task related to a certain application.

Siddiqu, et al. [14] used the watershed transformation as feature extraction, which selects random feature from the image. In [13], the authors have addressed features that encode appearance and motion by means of the Histogram Of Gradients (HOG) and the Histogram Of Flow (HOF) over several video frames. In [15] the authors have represented the activities by feature vectors from Independent Component Analysis (ICA) on video frames, and then based on these features they recognized the human activities by using Hidden Markov Model (HMM) classifier. In [16] the authors divide the image into 8x8 non –overlapping blocks and Mean value of each block is calculated and used as feature. In [17] authors have presented 'Action Bank', a high-level representation of video which is comprised of many individual action detectors sampled broadly in semantic space as well as viewpoint space. Kholgade and Savakis [11] used the shape information in the form of shape context as feature to represent an activity.

C. Action Recognition

In several works classifiers such as nearest neighbors, neural networks and support vector machines (SVMs) are used. For example, the authors of [17] have used SVM for classification purpose and in [11] Neural Network is used as classifier. The authors of [15] have used Hidden Markov Model (HMM). Also Decision Tree is one of such classifiers.

III. Statistical Method For Blob Extraction

Many of the background subtraction algorithms are susceptible to global and local illumination changes such as shadows and highlights. These cause the consequent processes, e.g. tracking, recognition, etc., to fail. These tasks require higher accuracy and efficiency in detection. This problem is addressed by the statistical method [5]. This algorithm is able to handle the local illumination change

problems, such as shadows and highlights, as well as the global illumination changes.
The algorithm [5] is adapted as shown in following steps:

1. Introduction of a new computational color model
2. Background modeling
3. Threshold selection
4. Subtraction operation or pixel classification

A. Computational Color Model

This color model [5] separates the brightness from the chromaticity component. Figure 2 illustrates this color model in three-dimensional RGB color space. In the image, E_i = [ER(i), EG(i), EB(i)] represents the pixel i's expected RGB color in the reference or background image. The line OE_i passing through the origin and the point E_i is called expected chromaticity line. I_i = [IR(i), IG(i), IB(i)] denotes the pixel's RGB color value in a current image that is subtracted from the background. The aim of this model is to measure the distortion of I_i from E_i. This is done by decomposing the distortion measurement into two components, brightness distortion (α) and chromaticity distortion (*CD*). The detailed description about this is given in [5].

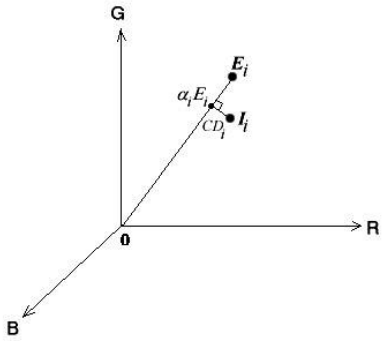


Figure 2. Computational Color Model [5]

B. Background Modeling

In the background training process, the reference background image and some parameters associated with normalization are computed over a number of static background frames. The background is modeled statistically on a pixel by pixel basis. A pixel is modeled by a 4-tuple $<E_i,s_i,a_i,b_i>$ where E_i is the expected color value, s_i is the standard deviation of color value, a_i is the variation of the brightness distortion, and b_i is the variation of the chromaticity distortion the of ith pixel [5]. The detailed description is provided in [5].

C. Threshold Selection

Threshold selection determines appropriate threshold values used in the subtraction operation to obtain a desired detection rate. One threshold τ_{CD} for chromaticity distortion, and two thresholds $\tau_{\alpha 1}$ and $\tau_{\alpha 2}$ for brightness distortion that define the brightness range, are needed. In this work these three thresholds are selected by trial and error method [5].

D. Pixel Classification or Subtraction Operation

In this step, the difference between the background image and the current image is evaluated. The difference is decomposed into brightness and chromaticity components. Applying the suitable thresholds on the brightness distortion (α) and the chromaticity distortion (*CD*) of a pixel *i* yields an object mask M(*i*) which indicates the type of the pixel. This method classifies a given pixel into four categories [5]. A pixel in the current image is

1) *Original background (B)* if brightness and chromaticity are similar to those of the same pixel in the background image.
2) *Shadow (S)* if chromaticity is similar but brightness is lower than those of the same pixel in the background image.
3) *Highlighted background (H)* if chromaticity is similar but brightness is higher than the background image.
4) *Moving foreground object (F)* if the chromaticity is different from the expected values in the background image.

IV. Feature Extraction using Shape Context

The blobs extracted in the last section are used in several different ways for feature extraction. The feature extraction technique used in this work is to find the shape context.

(a)

(b)

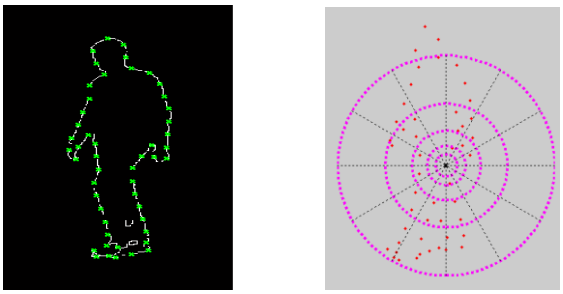
(c)

Figure 3. Results of background subtraction and silhouette boundary extraction: (a) Original image, (b) Silhouette and (c) Boundary of silhouette.

The shape context [11] is a descriptor that provides information about the way boundary points are 'spread out' with respect to one another.

For this, boundary points of each silhouette are obtained using an 8-connectivity criterion, where the perimeter was calculated and divided into 50 equal segments, in order to extract 50 equally spaced points. These constituted the features for further processing needed to obtain the shape context. A count is maintained of all the points that fall in the various segments of the circles drawn out around the central point. This is equivalent to a log-polar distribution, where the r-axis is in log-scale. In this application, 12 angular bins and 5 radial bins were used, with the minimum r being 0.2 units away from the center and maximum being 3.5 units away. The entire histogram was reshaped to a 60-element row vector. This 60-element descriptor was developed for each of the 50 boundary points obtained previously [11]. The detailed mathematical calculations are provided and followed as shown in [18].

Since the shape context is developed by considering the equidistant points, it is invariant to translation. Also, since it refers to number of points in bins and does not care about the displacement values themselves, it is also invariant to scaling. However, the shape context is not invariant to rotation.



(a) (b)

Figure 4. (a) Silhouette divided into 50 equidistant boundary points (b) Log-polar histogram binning

As shape context only refer to the shape information and does not contain any speed information, this descriptor is not able to distinguish among the activities which has nearly same shape information but different speeds, e.g. actions like walking and running.

V. Action Recognition using Neural Network

For training the network, 25 consecutive images were selected from each of the training videos of hand waving and walking with the exception of jumping where 25 alternate images were selected. This resulted in a total of 300 images employed for training. For each image, the shape context matrix was recast into a 3000 element vector, and principal components analysis was used to reduce the dimensionality of this vector to 50. Thus, a 50-by-300 element matrix was generated and was used to train a neural network. The network topology consisted of an input layer of 50 neurons, hidden layer of 10 neurons and an output layer of 3 neurons, where each output neuron corresponded to an activity.

VI. Implementation Results

Table-1 shows the results that were obtained using neural networks, as described in previous Section. The columns indicate actual action available in the video frames, while the actions in the row represent classification of actions according to experiment. These results are obtained on UIUC1 action dataset.

TABLE I. Implementation Results

	Hand Waving	**Jumping**	**Walking**
Hand Waving	99	0	1
Jumping	0	87	13
Walking	0	15	85

Here the shape context descriptor is able to achieve 90.33% accuracy with UIUC1 dataset. As we can see here, the action Hand Waving is mostly correctly classified but Actions jumping and walking are misclassified into each other by few percent. As mentioned in section IV, one of the reasons behind this is lack of speed information.

VII. Conclusion

The described system was implemented in three steps: blob extraction, feature extraction and action recognition. The technique used for feature extraction is shape context in which, shape context for each frame is found out. The technique used for action classification is neural network, which takes a frame as one sample to classify. Three actions Hand waving, jumping and walking are classified with approximately 90% accuracy. This work is also being extended by adding more actions in classification. The accuracy can be further improved by employing other algorithms for feature extraction and using alternative classifiers for action recognition purpose.

REFERENCES

[1] Thomas B. Moeslund and Erik Granum, "A Survey Of Computer Vision-Based Human Motion Capture", Computer Vision and Image Understanding 81, 231-268(2001)

[2] Pavan Turaga, Rama Chellappa, V. S. Subrahmanian and Octavian Udrea, "Machine Recognition of Human Activities: A survey", Circuits and Systems for Video Technology, IEEE Transactions, Nov. 2008, Volume: 18, Issue: 11, Page(s): 1473 - 1488 .

[3] Tarun Baloch, Master of Technology thesis, "Background Subtraction in Highly Illuminated Indoor Environment", Indian Institute Of Technology, Kanpur, January 2010.

[4] T. Bouwmans, F. El Baf, B. Vachon, "Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey", Recent Patents on Computer Science 1, 3 (2008) 219-237.

[5] Thanarat Horprasert, David Harwood, and Larry S. Davis, "A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection", ICCV Frame-Rate WS (1999).

[6] W. L. Khong, W. Y. Kow, H.T. Tan, H. P. Yoong, K. T. K. Teo, "Kalman Filtering Based Object Tracking in Surveillance Video System", 3rd (2011) CUTSE International Conference.

[7] Rachel Kleinbauer, "Kalman Filtering Implementation with Matlab", Helsinki, November 2004.

[8] Thomas B. Moeslund, Adrian Hilton, Volker Kruger, "A survey of advances in vision-based human motion capture and analysis", Computer Vision and Image Understanding 104 (2006) 90–126.

[9] Fahad Fazal Elahi Guraya[1], Pierre-Yves Bayle[2] and Faouzi Alaya Cheikh[1], "People Tracking via a Modified CAMSHIFT Algorithm", [1]Department of Computer Science and Media Technology, Gjovik University College, [2]Department of Computer Science, Universite de Bourgogne.

[10] Ahmad H. Al-Mazeed, Mark S. Nixon and Steve R. Gunn, "Fusing Complementary Operators to Enhance Foreground/Background Segmentation", British Machine Vision Conference 2003, Norwich,BMVA Press, 501-510.

[11] Natasha Kholgade and Andreas Savakis, "Human activity recognition in video using two methods for matching shape contexts of silhouettes".

[12] Massimo Piccardi, "Background subtraction techniques: a review", 2004 IEEE International Conference on Systems, Man and Cybernetics

[13] Plinio Moreno, Pedro Ribeiro, and Jos_e Santos-Victor, "Feature Selection for tracker-less human activity recognition", Proceeding ICIAR'11 Proceedings of the 8th international conference on Image analysis and recognition - Volume Part I Pages 152-160.

[14] Muhammad Hameed Siddiqi, Muhammad Fahim, Sung young Lee, Young-Koo Lee "Human Activity Recognition Based on Morphological Dilation followed by Watershed Transformation Method", 2010 International Conference on Electronics and Information Engineering (ICEIE 2010).

[15] M.Z. Uddin, J.J. Lee, and T.-S. Kim, "Shape-Based Human Activity Recognition Using Independent Component Analysis and Hidden Markov Model, " *Proc. of 21Ist International Conference on Industrial, Engineering, and other Applications of Applied Intelligent Systems*, 2008, pp.245-254, Springer-Verlag Berlin Heidelberg.

[16] Rachana V. Modi, Tejas B. Mehta, "Neural Network based Approach for Recognition Human Motion using Stationary Camera", International Journal of Computer Applications (0975 – 8887) Volume 25– No.6, July 2011

[17] Sreemanananth Sadanand and Jason J. Corso, "Action Bank: A High-Level Representation of Activity in Video", IEEE CVPR 2012

[18] Natasha Prashant Kholgade, "Recognition of Human Activities and Expressions in Video Sequences using Shape Context Descriptor", A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering.