Human Activity Tracking System (Hats)

Ketan Chaudhari, Hemlata Tejwani, Sayali Chandane
BE, Computer Engineering
KC College of Engineering & Management Studies
& Research,Thane

***Abstract*— In this paper, we describe a novel template matching based approach for recognition of different human activities in a video sequence. We model the background in the scene using a simple statistical model and extract the foreground objects present in a scene. The matching templates are constructed using the motion history images (MHI) and spatial silhouettes for recognizing activities like walking, standing, bending, sleeping and jogging in a video sequence. Experimental results demonstrate that the proposed method can recognize these activities accurately for standard KTH database as well as for our own database.**

Keywords-* *Activity recognition; template matching; moments invariant; background modeling

I. INTRODUCTION

Human activity recognition is a popular area of research in the field of computer vision. It is the basis of applications in many areas such as security, surveillance, clinical applications, biomechanical applications, human robot interaction, entertainment, education, training, digital libraries and video or image annotations as well as in video conferencing and model based coding. Recognition of human actions and activities provides important cue for human behavior analysis techniques.

An activity is a sequence of movements generated during the performance of a task. This is a difficult task because the shape of different objects performing an activity can be different and the speed and style with which an activity is performed can vary from object to object. Many approaches have been proposed so far to solve this problem. Template Matching based approaches are very important among them because of their simplicity and robustness. Weinland et al. provided a good survey of different human activity recognition techniques. The template matching based techniques can be broadly classified into three categories: body template based methods, feature template based methods and image template based methods. Body template based methods represent the spatial structure of activities with respect to the human body. In each frame of the observed video sequence, the posture of a human body is reconstructed from a variety of available image features. The action recognition is performed based on these posture estimations. This is an intuitive and biologically-plausible approach for activity recognition and supported by psychophysical work on visual interpretation of biological motion. However, in body model based representations, the resulting interest regions are linked to certain body-parts or even image coordinates. This imposes certain restrictions on recognition of different activities. In feature template based methods, activities are recognized based on the statistics of sparse features in the image. It is a local representation of activities. It decomposes the image/video into smaller interest regions and describes each region as a separate feature. An immediate advantage of these approaches is that they neither rely on explicit body part labeling, nor on explicit human detection and localization. The image template based methods are simple than the above described methods and can be computed efficiently. Motion history images (MHI) and motion energy images (MEI) can be used to determine the location and type of activities in the scene. In outdoor environment, where variations in lighting conditions and change in background produces noise, a robust background modeling is required. MHI and MEI is a good solution of this problem. The approach proposed by Bobick et al., used motion templates for recognizing the activities in a specific environment of aerobics exercise. They used simple frame differencing for obtaining segmented foreground and MHI for obtaining motion information in a view-specific environment. It does not give good activity recognition accuracy in outdoor environment

This paper presents a template based activity recognition method. This approach considers the shape information along with the motion history for performing an activity. For obtaining the accurate foreground segmentation a robust statistical background model is constructed. The technique can recognize the static activities like standing and sleeping as well as dynamic activities like walking, jogging, etc. In the proposed approach, covariance based matching is applied to recognize static activities and moment invariants are used to recognize dynamic activities. The proposed technique has two advantages over the technique described in. One is that the background segmentation is obtained using a robust statistical model which can better adapt the changes in lighting conditions whereas in simple frame differencing is not adaptive to these changes. Second, the proposed method uses motion as well as object shape information to construct the MHIs while on contrary Bobick et al. used only motion to construct the MHIs. So the proposed method can accurately recognize the activities with very less motion such as standing and sleeping.

II. LITERATURE SURVEY

Over the past years extensive research has been conducted by psychophysicists, neuroscientists, and engineers on various aspects of body recognition by humans and machines. Psychophysicists and neuroscientists have been concerned with issues such as whether body perception is a dedicated process and whether it is done holistically or by local feature analysis.

The earliest results on automatic machine recognition of bodys can be traced back to the seminal work of Kanade and Kelly in the 1970's. Early approaches treated body recognition as a 2D pattern recognition problem, using measured attributes of features (e.g. the distances between important points) in bodys or body profiles.

During the 1980's, work on body recognition in the computer vision community remained largely dormant. However, since the early 1990's, research interest in FRT has grown significantly. Body recognition can be classified as holistic approaches, which consider the whole body at a time, or feature-based methods, which look at the interplay between the different features on the body. Among appearance-based holistic approaches, Eigen bodys and Fisher bodys have proved to be effective in experiments with large databases. Feature-based graph matching approaches have also been quite successful. Compared to holistic approaches, feature-based methods are less sensitive to variations in illumination and viewpoint and to inaccuracy in body localization. However, the feature extraction techniques needed for this type of approach are still not reliable or accurate enough.

For example, most eye localization techniques assume some geometric and textural models and do not work if the eye is closed. Recently, much research has been concentrated on video-based body recognition. The still image problem has several inherent advantages and disadvantages. For applications such as drivers' licenses, due to the controlled nature of the image acquisition process, the segmentation problem is rather easy. However, if only a static picture of an airport scene is available, automatic location and segmentation of a body could pose serious challenges to any segmentation algorithm.

On the other hand, if a video sequence is available, segmentation of a moving person can be more easily accomplished using motion as a cue. But the small size and low image quality of bodys captured from video can significantly increase the difficulty in recognition. A thorough review of the literature in body recognition is available in. During the past eight years, body recognition has received increased attention and has advanced technically. Many commercial systems for still body recognition are now available. Recently, significant research efforts have been focused on video based body modeling/tracking, recognition, and system integration. New datasets have been created and evaluations of recognition techniques using these databases have been carried out. It is not an overstatement to say that body recognition has become one of the most active applications of pattern recognition, image analysis and understanding.

In 1975, Johansson's experiment shows that humans can recognize activity with extremely compact observers. Johansson demonstrated his statement using a movie of a person walking in a dark room with lights attached to the person's major joints. Even though only light spots could be observed, there was a strong identification of the 3D motion in these movies. In recent studies, Fuijiyoshi and Lipton proposed to use "star" skeleton extracted from silhouettes for motion analysis.

Yu and Aggarwal use extremities as semantic posture representation in their application for the detection of fence climbing. Zia et al. present an action recognition algorithm using body joint-angle features extracted from the RGB images from stereo cameras. Their dataset contains 8 simple actions (e.g., left hand up), and they were all taken from frontal views. Inspired by natural language processing and information retrieval, bag-of-words approaches are also applied to recognize actions as a form of descriptive action unites. In these approaches, actions are represented as a collection of visual words, which is the codebook of spatio-temporal features.

Schuldt et al. integrate space-time interest point's representation with SVM classification scheme. Dollar et al. employ histogram of video cuboids for action representation. Wang et al. represent the frames using the motion descriptor computed from optical flow vectors and represent actions as a bag of coded frames. However, all these features are computed from RGB images and are view dependent. Researchers also explored free viewpoint action recognition algorithms from RGB images. Due to the large variations in motion induced by camera perspective, it is extremely challenging to generalize them to other views even for very simple actions. One way t address the problem is to store templates from several canonical views and interpolate across the stored views. Scalability is a hard problem for this approach. Another way is to map an example from an arbitrary view to a stored model by applying homography. The model is usually captured using multiple cameras. Weinland et al. model action as a sequence of exemplars which are represented in 3D as visual hulls that have been computed using a system of 5 calibrated cameras.

Parameswaran et al. define a view-invariant representation of actions based on the theory of 2D and 3D invariants. They assume that there exists at least one key pose in the sequence in which 5 points are aligned on a plane in the 3D world coordinates. Weinland et al. extend the notion of motion-history to 3D. They combine views from multiple cameras to build a 3D binary occupancy volume. Motion history is computed over these 3D volumes and view-invariant features are extracted by computing the circular FFT of the volume. The release of the low-cost RGBD sensor Kinect has brought excitement to the research in computer vision, gaming, gesture-based control, and virtual reality.

Shotton et al. proposed a method to predict 3D positions of body joints from a single depth image from Kinect. Xia et al. proposed a model based algorithm to detect humans using depth maps generated by Kinect. There are a few works on the recognition of human actions from depth data in the past two years. Li et al. employ an action graph to model the dynamics of the actions and sample a bag of 3D points from the depth map to characterize a set of salient postures that correspond to the nodes in the action graph. However, the sampling scheme is view dependent. Lalal et al. utilize the Radon transformation on depth silhouettes to recognize human home activities. The depth images were captured a ZCAM. This method is also view dependent. Sung et al. extract features from the skeleton data provided by Prime Sense from RGBD data from Kinect and use a supervised learning approach to infer activities from RGB and depth images from Kinect. Considering they extract features from both types of imageries, the result is interesting but at the same time not as good as one would expect.

III. METHODOLOGY

OpenCV (*Open Source Computer Vision Library*) is a library of programming functions mainly aimed at real-time computer vision, developed by Intel, and now supported by Willow Garage and Itseez. It is free for use under the open source BSD license. The library is cross-platform. It focuses mainly on real-time image processing. If the library finds Intel's Integrated Performance Primitives on the system, it will use these proprietary optimized routines to accelerate itself. History Officially launched in 1999, the OpenCV project was initially an Intel Research initiative to advance CPU-intensive applications, part of a series of projects including real-time ray tracing and 3D display walls. The main contributors to the project included a number of optimization experts in Intel Russia, as well as Intel’s Performance Library Team.

In the early days of OpenCV, the goals of the project were described as Advance vision research by providing not only open but also optimized code for basic vision infrastructure. No more reinventing the wheel. Disseminate vision knowledge by providing a common infrastructure that developers could build on, so that code would be more readily readable and transferable. Advance vision-based commercial applications by making portable, performance-optimized code available for free—with a license that did not require to be open or free themselves. The first alpha version of OpenCV was released to the public at the IEEE Conference on Computer Vision and Pattern Recognition in 2000, and five betas were released between 2001 and 2005. The first 1.0 version was released in 2006. In mid-2008, OpenCV obtained corporate support from Willow Garage, and is now again under active development. A version 1.1 "pre-release" was released in October 2008. The second major release of the OpenCV was on October 2009. OpenCV 2 includes major changes to the C++ interbody, aiming at easier, more type-safe patterns, new functions, and better implementations for existing ones in terms of performance (especially on multi-core systems). Official releases now occur every six months[1] and development is now done by an independent Russian team supported by commercial corporations.

JavaCV first provides wrappers to commonly used libraries by researchers in the field of computer vision: OpenCV, FFmpeg, libdc1394, PGR Fly Capture, video Input, and ARToolKitPlus. The following classes, found under the com.googlecode.javacv.cpp package namespace, expose their complete APIs: opencv_core, opencv_imgproc, opencv_video, opencv_features2d, opencv_calib3d, opencv_objdetect, opencv_highgui, opencv_legacy, avutil, avcodec, avformat, avdevice, avfilter, postprocess, swscale, dc1394, PGRFlyCapture, videoInputLib, and ARToolKitPlus, respectively. Moreover, utility classes make it easy to use their functionality on the Java platform, including Android.

JavaCV also comes with hardware accelerated full-screen image display (CanvasFrame), easy-to-use methods to execute code in parallel on multiple cores (Parallel), user-friendly geometric and color calibration of cameras and projectors (GeometricCalibrator, ProCamGeometric Calibrator, ProCamColorCalibrator), detection and matching of feature points (ObjectFinder), a set of classes that implement direct image alignment of projector-camera systems (mainly GNImageAligner, ProjectiveTransformer, ProjectiveGainBiasTransformer, ProCam Transformer, and ReflectanceInitializer), as well as miscellaneous functionality in the JavaCV class.

HAAR CASCADE CLASSIFIERS :

The core basis for Haar classifier object detection is the Haar-like features. These features, rather than using the intensity values of a pixel, use the change in contrast values between adjacent rectangular groups of pixels. The contrast variances between the pixel groups are used to determine relative light and dark areas. Two or three adjacent groups
with a relative contrast variance form a Haar-like feature. Haar-like features, are used to detect an image. Haar features can easily be scaled by increasing or decreasing the size of the pixel group being examined. This allows features to be used to detect objects of various sizes.

Activity Tracking System

The principle of Activity Tracking System is to capture the users images through the camera firstly, then to analyze the users movements by using detection algorithm as described above, and finally to recognize the activity

The technique used for matching the spatio-temporal templates to recognize the activities is rotation, scale, and translation invariant. The training of actions is performed considering different views of activity performance. For each view of each action a statistical model of the moments using variance and covariance is generated for MHIs. The 7 moment invariants are used as activity descriptors. To recognize an input action, a mahalanobis distance is calculated between the moment description of the input and each of the known actions. The distance matrix so obtained is analyzed in terms of separation distances for different actions.

We have shown results for our own created database. This database contains four static human activities namely, sifting, sleeping, standing, bending and two dynamic activities namely, walking and jogging. These videos are taken in real outdoor environment. From the observation of this figure, it is clear that the proposed method is well capable of recognizing these static and dynamic activities. Moreover, there is some little movement in each activity, i.e. pose of human object does not remain still for all the time. Direction of each human object also changes in different frames. Therefore, the proposed method is pose invariant and frontal view is not necessary for recognition of objects and suits for recognition of objects with frontal as well as side view. In addi*n to this activities (a,b,c,d,e) are performed in outdoor environment whereas activity (f) is performed in indoor environment. But in both scenarios, the proposed method is capable of recognizing these activities. Shadow is also present in scene, but it does not impose any restriction on recognition accuracy

ALGORITHM:

1. Start the Camera
2. Initialize the Page Grabber
3. Capture the image from the Page Grabber Frame
4. Convert the captured image into grey scale image
5. Segment the grey scale images into specified segments according to the grey levels
6. Convert these segments into Haar Cascade format
7. Compare it with Haar Cascade XML history file
8. If comparison is not successful, go to step 3 If comparison is successful, proceed
9. Highlight the region whose comparison is successful
10. Detect the activity
11. Display the detected activity

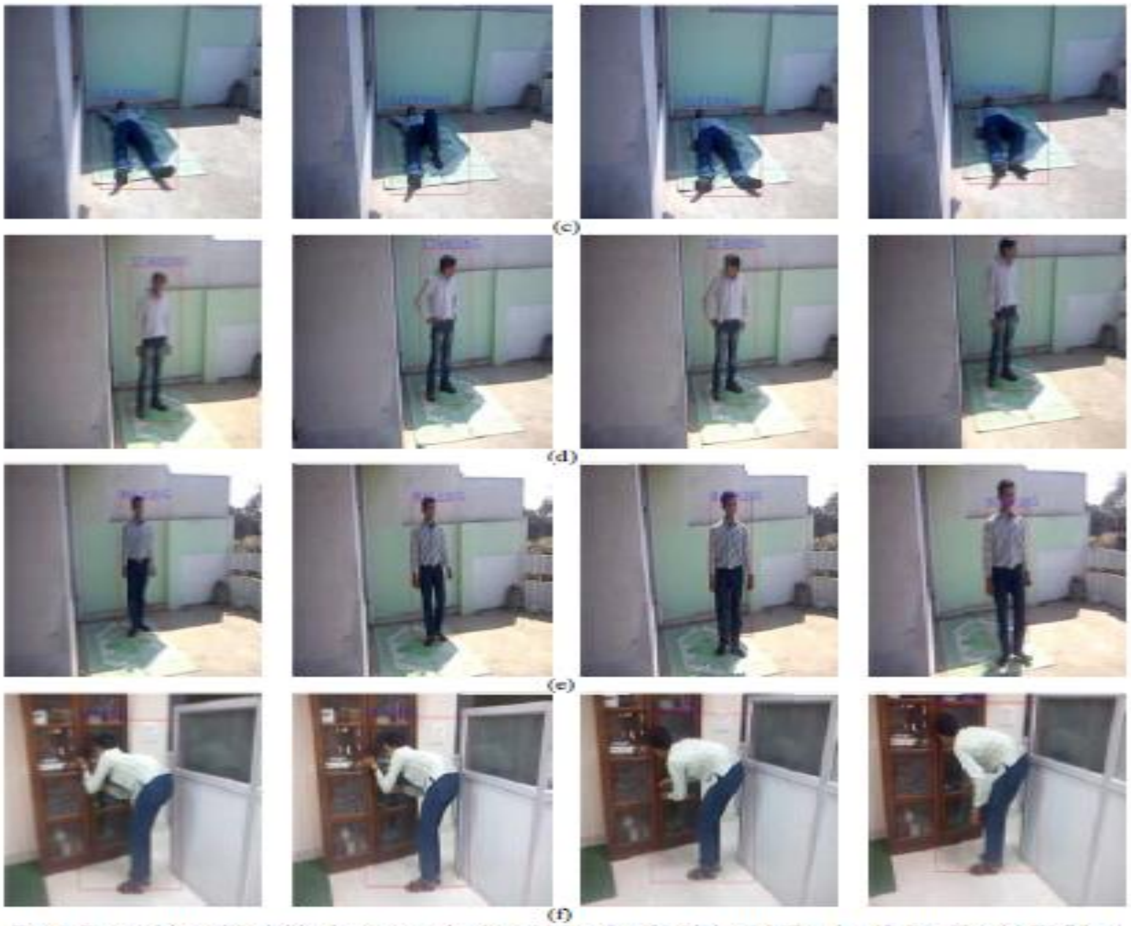
(c)
(d)
(e)
(f)

Fig. 4. Recognition of Activities in our own database (a) Jogging (b) Sitting (c) Sleeping (d) Standing (e) Walking and (f) Bending

IV. ADVANTAGES

1. Quick response time
2. Customized processing
3. Small memory factor
4. Highly secure
5. Less expensive

V. DISADVANTAGES

1. Not more accurate

VI. REFERENCES

1) D. Weinland, and R. Ronfard, "A survey of vision based methods for action representation, segmentation, and recognition," Computer Vision and Image Understanding, vol. 115, no. 2, pp. 529-551, 2011.

2) I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity adapted motion events for spatio-temporal recognition," vol. 108, pp. 207-229, 2007.

3) Y. Ke, R. Sukthankar, and M. Hebert, "Volumetric features for video event detection," Int. J. of Computer Vision, 2010.

4) Technical Report CMU-CS-08-113, "Volumetric features for video event detection," March 2008.

5) A.F. Bobick, and J.W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, 2001.

6) D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, vol. 104, no. 2, pp. 249-257, 2006.

7) A. Fathi, and G. Mori, "Action recognition by learning mid-level motion features," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

8) R. Souvenir, and J. Babbs, "Learning the viewpoint manifold for action recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 118-125, 2008.

9) A. Farhadi, and M.K. Tabrizi, "Learning to recognize activities from the wrong view point," Proc. European Conference on Computer Vision, pp. 154-166, 2008.

10) M. Hofmann, and D.M. Gavrila, "Multi-view 3D human upper body pose estimation combining single-frame recovery, temporal integration and model adaptation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, Miami (USA),2009.