RECOGNITION OF HUMAN ACTIVITIES BASED ON BLOB DETECTION WITH AN EFFICIENT ALGORITHMS

SUHAS R. KOTHAVLE, DUSHYANT VERMA, PRASHANT LAKKADWALA

1 M.Tech Student, 2 Assistant Professor, 3 Assistant Professor

Acropolis Technical Campus, Indore    Acropolis Technical Campus, Indore    Acropolis Technical Campus, Indore

ABSTRACT: This paper deals with blob features of human activities analysis in the video. Human Activity Recognition is an active area of research in computer vision with wide scale of applications in video surveillance, motion analysis, virtual reality interfaces, robot navigation and recognition, video indexing, browsing, HCI, choreography, sports video analysis etc. It analyzes the characteristic features of various human actions and classify them accordingly, which consist of following stages: Background subtraction, tracking, feature extraction and classification. We analyze human activities in the sequential frames because human activities can be considered as a temporal object which contains a series of frames. Firstly, we establish a statistical background model and extract foreground object through background subtraction in the video stream. Then, we use foreground blobs of the current frame and a series of frames before the current frame to form a new feature image in certain rules. Finally, we combine the non-zero pixels in the feature image into blobs using the connected component method. Then each blob corresponds to an activity which is characterized by the blob appearance. By recognizing blob features we can recognize activities. We use Advanced Gaussian Mixture algorithm and HMM model to extract features for each type of human activities and employ minimal Mahalanobis distance to classify the activity.

KEYWORDS: Activity Recognition, Surveillance Blob detection, Video Analysis

1. INTRODUCTION

Human action recognition is a very important component of visual surveillance systems for event based analysis of surveillance videos.

Visual surveillance systems play a very crucial role in the circumstances where continuous patrolling by human guards is not possible like international border patrolling, nuclear reactors etc. Demand for automatic surveillance systems in civilian applications like monitoring a parking lot, shopping complexes etc. is also increasing heavily. Its difficult and manpower intensive to monitor the data collected from various cameras continuously and this gives rise to the necessity for automatic understanding of human actions and building a higher level knowledge of the events occurring in the scene by the computer vision system. Recognition of human movements has also been exploited to a large extent for animation like avatar control, for giving gesture based commands to virtual reality interfaces, human Computer interactions etc. Content based video retrieval, indexing and searching is also becoming popular these days with the concepts like Video Google coming up. These systems require cognitive vision techniques for analyzing videos which in real life scenarios mostly converges to analyzing human actions in the videos. Video annotation of sports videos is an excellent example of this category where complex human sport actions are required to be classified. A good discussion on the promising application scenarios and the suitable approaches in these scenarios can be obtained in [8]. The wide scale
applications of human activity analysis and various challenges involved at different stages in building this system makes it a demanding area of research in automatic human activities recognition in video streams is gaining more and more attention in the video analysis research community due to many video applications such as video content analysis, video retrieval, video summarization, visual surveillance and human-computer interaction. In recent years, many different approaches related to activity recognition have been proposed. The common approaches in human activity recognition can be categorized into two groups based on the methods. The first group of methods regards video as a temporal object which embraces both spatial and temporal information. Human activity is recognized by 3D spatiotemporal volumetric features. In [1], Zelnik-Manor and Irani cluster long sequences of events to detect classes of activities in those sequences. In [2], Bobick and Davis used Motion Energy and Motion History Images (MEI and MHI) to recognize many types of aerobics exercises. In [3], Y. Ke, R. Sukthankar, and M. Hebert employed an approach based on the volumetric analysis of video, where a sequence of images is treated as a three dimensional space-time volume. The second group of methods involves an underlying semantic structure. They model human activity in a long video sequence by using a hidden Markov model (HMM), dynamic Bayesian network (DBN) or stochastic context free grammars (SCFG). In [5], Ivanov and Bobick employed stochastic context-free grammars (SCFG) for recognizing event by combining a HMM at lower level with SCFG. Our approach belongs to the first group. A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM can be considered as the simplest dynamic Bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model parameters are known exactly, the model is still 'hidden'. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical stage. In the feature space, we use Gausscore following, partial discharges and bioinformatics. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. Here we propose an approach for activity recognition, using blob features within short-time intervals. Our method has three basics steps: feature extraction, feature modeling and activity recognition. Given a video sequence, moving objects are detected in each frame by adaptive background subtraction. We use the detected foreground objects in the current frame and the frames in series before the current frame to constitute a new feature image. Then we combine the nonzero pixels in the feature image into blobs using the connected component analysis. Because each blob corresponds to an activity which is characterized by the blob appearance, we can recognize activities by recognizing features of the corresponding blobs. We select some blob features including mean and variance of each blob luminance, the ratio of length and width of the bounding box and 7-hu moments. These features are all rotation, scaling, and displacement invariant. These features can describe not only the luminance information of the blob but
also the shape information of the blob. We use these features to constitute a feature vector. Compared with the common method our method is more efficient because we acquire 3D human activity information in 2D feature using Gaussian Mixture to model features for each type of human activities because the feature vector shows multi-modality. We solve the Gaussian Mixture by using Expectation-Maximization(EM) with automatic model order selection based on modified Minimum Description Length (MDL) principle [11]. In the feature space, a feature point corresponds an activity. After obtaining the parameters of the mixture model for each type of activity, we can recognize the activities in the feature space by checking the Mahalanobis distance between the test feature point and each mean of the Gaussian Mixture in every class. We classified the test feature point as the class which has smallest distance with the test point.

\[ \hat{t}(x, y, 0) = \arg\min_{t} \left( \sum_{i} \left( t_i - m_i \right)^2 \right) \]

2. FEATURE EXTRACTION

In this paper, we process gray-level video due to the color information is not reliable in the activity recognition. Color videos are changed to the grey-level videos before processing. Human activities can be considered as a longterm temporal object which contains a series of frames. Firstly, we establish a statistical background model by using adaptive Gaussian mixture proposed by C. Stauffer and W. Grimson in [6]. For each frame, we extract foreground object through background subtraction. Then we use foreground blobs of the current frame and a series of frames before the current frame to form a new feature image. That means we use N frames in series to form the new feature image. The new feature image is generated through the following recursive algorithm: If a pixel belongs to foreground in the current frame:

where \( I(x, y, t) \) is intensity value of a pixel at \((x, y)\) at the current frame, \( 0 < t \leq N \) a is the accumulation factor and \( d \) is the decay factor. \( I_0 \) is pixel intensity value of foreground blobs in the first frame. If a pixel belongs to foreground in the current frame its intensity value increases gradually through the accumulation factor, otherwise its intensity value decreases gradually through the decay factor. The accumulation factor and the decay factor give more flexibility to control change of the pixels. The feature image is equivalent to the Motion History Image when the accumulation factor a is set to 1. Each activity can cause the changes of a group of neighbouring pixels, and the changed pixels are spatially connected. We combine the non-zero pixels in the feature image into blobs using the connected component analysis. Then each blob corresponds to a human activity which is characterized by the blob appearance. The blobs capture many features of the activities including speed of the people, shape of the action. In the feature image, each blob can be described by some blob-level features, such as shape, area, statistic of the pixel luminance in the blob. We need to choose some features that are rotation, scaling, displacement invariant. The chosen features involve mean and variance of each blob luminance, the ratio of length and width of the bounding box and 7-hu moments [10] which are known to yield reasonable shape discrimination. We use these features to constitute a feature vector. Among these features, the mean of blob luminance is to first order statistic and variance of blob luminance is second order statistic. They capture speed and other motion information of the activity. Low mean of blob luminance reflects high speed of the activity and high mean of blob luminance reflects low speed of the activity. The 7-hu moment captures shape of the action. Each type of activities has its corresponding blob shape. The 7-hu moments are rotation, scaling, displacement invariant and they can describe shape information independent of position, size, and orientation. Figure 1 shows some activities and corresponding blobs extracted from the feature image. The activities from left to right on the top row are respectively walk, inactive and fight.
are on the bottom row.

Figure 2: blob feature in active image. Left to right: input image 1, input image 1 with detected features.

From Figure 1 we can see the blobs that correspond to different activities have different shapes and luminance. The blobs corresponding to the high speed activities have smaller luminance than the blobs that corresponding to the low speed activities. The pictures in Figure 1 are taken from the CAVIAR video dataset [9]. Figure 2 shows two images where blobs have been found.

3. ACTIVITY RECOGNITION

After we obtain parameters of the Gaussian mixture model for each type of activities by using the training data of each type of activities, we can use the parameters of the mixture model to classify the new activities. Because each activity corresponds to a feature point in the feature space, we check the Mahalanobis distance between the test feature point and each mean of the Gaussian Mixture in every class. If one of them is within a threshold, the test feature point is classified as that class. The threshold is chosen as $2.5 \sigma$ ($\sigma$ is standard variance). If more than one class matches that condition, we choose the class which has smallest distance with the test point.

Figure 3 shows the whole process of our method on human activity recognition. It involves two parameters learning process: background modeling learning and parameters of each mixture model learning in the feature space, and contains three steps to recognize human activity: feature extraction, feature modeling in the feature space, and activity recognition in the feature space.

Figure 1: the activities and corresponding blobs from left to right: walk, inactive and lift

Figure 3: Human activity recognition system. From left to right: frame separation, adaptive background subtraction, apply HMM model and Gaussian parameters to feature modeling, apply Mahalanobis distance, output the kind of activity.
4. EXPERIMENTS

We have conducted experiments on Public video datasets with resolution of 180*244 pixels. The data set consists of 3 scenarios, that is Standing, Aerobic Workout and walking appended with ground truth. Firstly, we learn the background model through the long video sequence by using adaptive Gaussian mixture [6]. Then we divide these video sequences into clips which contain the defined activities. The defined activities include Stand, Aerobic Workout and walk from the 3 scenarios in the video dataset. We use a half of these samples in each type for training and the others for testing. In the blobs extraction process, we filter out the blobs whose area smaller than 200 because smaller blobs are mostly generated by noise. The parameters are chosen as: the number of frames that we use to form feature image N =10, background model learning rate α = 0.01, accumulation factor a = 8, decay factor d = 15 and I0 =75. Table 1 shows the confusion matrix for these three activities.

<table>
<thead>
<tr>
<th>Class</th>
<th>Stand</th>
<th>Aerobic Workout</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand</td>
<td>3.3068</td>
<td>3.4745</td>
<td>3.4752</td>
</tr>
<tr>
<td>Aerobic Workout</td>
<td>3.4619</td>
<td>3.4099</td>
<td>3.4105</td>
</tr>
<tr>
<td>Walk</td>
<td>3.4787</td>
<td>3.4030</td>
<td>3.4008</td>
</tr>
</tbody>
</table>

The above Table show the confusion matrix in our method and P. Ribeiro’s method in [12] respectively. The both show the good accuracy.In [12], the author employs two sets of features. One set includes features about velocity of object. The other set includes features about optic flow inside the bounding box of the object. This method extracts motion information more directly, so it is more efficient to the activity that involves a lot of motions (for example, walking), but it can’t capture the shape information of the activity. Our method captures both motion and shape information. It can apply to the more types of activity (for example, running, dancing and gesture).

CONCLUSIONS

Like most of the other computer vision systems, human action recognition is highly context dependent. Activities of interest are also different in different scenarios. Moreover, what constitutes an action is a matter of perception. For example, fast jogging may be considered as running or as jogging. It will be interesting to argue whether we can define a generic set of actions over a broad range of scenarios. The key to any recognition problem is finding robust representative feature sets for the patterns. Three steps are needed in our method: feature extraction, feature modeling and activity recognition. Our method captures motion and shape features of the activity by using spatiotemporal information. We conduct experiments on public video dataset and the results show the accuracy of our method.

REFERENCES


