Human Activity Recognition Based On The Blob Features

Jie Yang, Jian Cheng, Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China 100190
{jyang, jcheng, luhq}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we present a novel approach for human activities recognition in the video. 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 Gaussian Mixture to model features for each type of human activities and employ Mahalanobis distance to measure the similarity.

Index Terms— Activity recognition, video analysis

1. INTRODUCTION

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 spatio-temporal 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.

We propose an approach for activity recognition, using blob features within short-time intervals. Our method has three steps: feature extraction, feature modeling and activity recognition. Given a video sequence, moving objects are detected in each frame by adaptive background subtraction [6]. 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 non-zero 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 [10]. These features are all rotation, scaling, 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 image.

In the feature space, we use 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

978-1-4244-4291-1/09/$25.00 ©2009 IEEE

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.

The remainder of the paper is organized as follows: Section 2 details the blob feature. Section 3 describes how to model the selected feature. Section 4 describes how to recognize the activity. Section 5 presents the experimental result on public video dataset. Section 6 concludes this paper.

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 long-term 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:

$$I(x,y,t) = \min(I(x,y,t-1) + \frac{255}{a}, 255) \quad (1)$$

$$I(x,y,0) = I_0 \quad (2)$$

If a pixel belongs to background in the current frame:

$$I(x,y,t) = \max(I(x,y,t-1) - \frac{255}{d}, 0) \quad (3)$$

$$I(x,y,0) = 0 \quad (4)$$

where $I(x,y,t)$ is intensity value of a pixel at (x,y) at the current frame, $0 < t \le N$, 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 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. The corresponding blobs extracted from the feature image are on the bottom row.

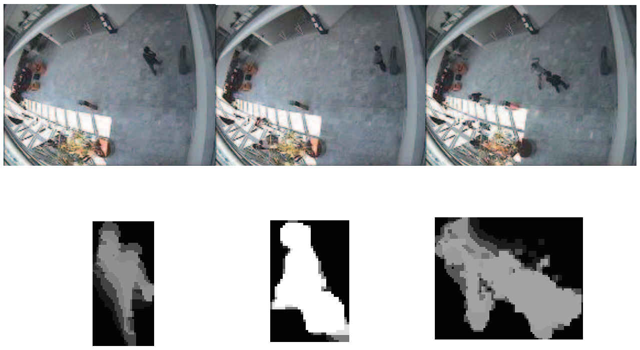

Figure 1: the activities and corresponding blobs from left to right: walk, inactive and fight

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].

3. FEATURE MODELING

Although our selected features are rotation, scaling, and displacement invariant, the features vector still shows multi-modality in the feature space resulting from the activity itself. The same type of activity can show different features. To deal with multi-modality in the feature space we employ Gaussian Mixture to model the features for each type of activity and adopt Expectation-Maximization (EM) to solve the Gaussian Mixture. We choose modified Minimum Description Length (MDL) principle [11] to automatically select model order.

MDL is used to extend maximum likelihood estimation to the model order unknown situation. We have some training data for each type of activity.

Suppose $\{x_1, x_2, ..., x_n\}$ are n independent training data for one type of activity which shows K modality in feature space. We can estimate the model order K by a standard MDL algorithm.

$$\hat{K} = \arg\min\left\{-\sum_{i=1}^{n}\ln f\left(x_i \mid \hat{\theta}(K)\right) + \frac{C(K)}{2}\ln(n)\right\} \quad (5)$$

where $f\left(x_i \mid \hat{\theta}(K)\right)$ is the conditional probability density function, $\hat{\theta}(K)$ are the mixture parameters estimated by maximum likelihood algorithm such as EM and $C(K)$ is the number of parameters which the K-component mixture need.

The first term in the bracket of Equation (5) measures the system entropy and corresponds to the maximized likelihood, while the second term measures the number of bits that are needed to encode the model parameters and serves as a penalty term for too large K.

In the standard MDL, each component in the mixture can only use the data that belong to it. We adopt a modified MDL algorithm in which the whole dataset are used for every component. That means the full covariance matrix is used. If we use the full covariance matrix, then

$$C(K) = K - 1 + \frac{d^2 + 3d}{2}K \quad (6)$$

So the model order is estimated as:

$$\hat{K} = \arg\min\left\{-\sum_{i=1}^{n}\ln f\left(x_i \mid \hat{\theta}(K)\right) + \frac{K-1}{2}\ln(n) + \frac{d^2+3d}{4}K\ln(n)\right\} \quad (7)$$

After we get the estimated model order, we can model each type of activity by Gaussian Mixture in the feature space.

4. 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 σ (σ is standard variance). If more than one class matches that condition, we choose the class which has smallest distance with the test point.

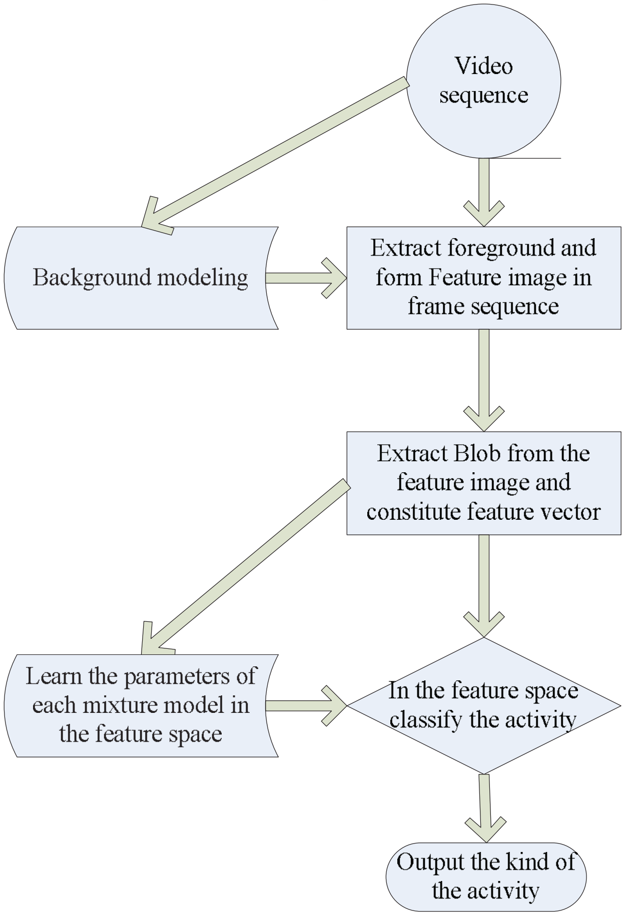
Video sequence
Background modeling
Extract foreground and form Feature image in frame sequence
Extract Blob from the feature image and constitute feature vector
Learn the parameters of each mixture model in the feature space
In the feature space classify the activity
Output the kind of the activity

Figure 2: flowchart of human activity recognition

Figure 2 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.

5. EXPERIMENTS

We have conducted experiments on CAVIAR video dataset [9] with resolution of 384*288 pixels. The data set consists of 28 video sequences in 5 scenarios, that is, Walking, Browsing, Collapse, Leaving objects, Meeting and Fighting 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 inactive, active, walk, run and fight from the 5 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 are smaller than 400 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 I_0 = 75. Table 1 shows the confusion matrix for these five activities.

Table 1: Confusion matrix using our method.

Class	Inactive	Active	Walk	Run	Fight
Inactive	100	0	0	0	0
Active	0	93.2	4.8	2.0	0
Walk	0	2.0	92.2	4.0	1.8
Run	0	2.0	3.7	94.3	0
Fight	0	2.2	2.0	0	95.8

Table 2: Confusion matrix using P. Ribeiro's method.

Class	Inactive	Active	Walk	Run	Fight
Inactive	95.2	4.8	0	0	0
Active	5.8	84.7	0	0	9.5
Walk	0	0.2	98.1	0.8	0.9
Run	0	0	0	100	0
Fight	0	2.7	5.3	0	92.0

Table1 and Table2 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, running), 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, dancing and gesture).

6. CONCLUSIONS

In this paper, we have introduced a novel approach to recognize human activities in the video sequences. 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 spatio-temporal information. We conduct experiments on public video dataset and the results show the accuracy of our method.

7. ACKNOWLEDGEMENT

This work is supported by National Science Foundation of China (NSFC) under Grant No. 60835002, 60605004.

8. REFERENCES

[1] L. Zelnik-Manor and M. Irani, "Event-based video analysis," In *Proc. CVPR*, 2001.

[2] Aaron F. Bobick, James W. Davis, "The Recognition of Human Movement Using Temporal Templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, March 2001.

[3] Y. Ke, R. Sukthankar and M. Hebert, "Efficient visual event detection using volumetric features," In *Proc. ICCV*, 2005.

[4] E. Shechtman and M. Irani, "Space-time behavior based correlation," In *Proc. CVPR*, 2005.

[5] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Analysis and Machine Intelligence,* 22(8):852–872, 2000.

[6] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," In *CVPR,* volume 2, pages 246–252, 1999.

[7] Xiang, T., Gong, S. and Parkinson, D, "Autonomous visual events detection and classification without explicit object-centred segmentation and tracking," In *British Machine Vision Conference*, pp. 233–242, 2002.

[8] T. Xiang and S. Gong, "Model selection for unsupervised learning of visual context," *International Journal of Computer Vision*, Vol. 69, No. 2, pp. 181-201, 2006.

[9] CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[10] M. Hu, "Visual Pattern Recognition by Moment Invariants," *IRETrans. Information Theory,* vol 8. no. 2, pp. 179-187, 1962.

[11] M. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. PAMI*, 24(3):381–396, 2002.

[12] P. Ribeiro, J. Santos-Victor, "Human Activities Recognition from Video: modeling, feature selection and classification architecture." In *British Machine Vision Conference*, pp. 61-70, 2005.