Human Activity Recognition using bag of feature

Sohila Nemati, Azam Bastanfard and Shabnam Asbaghi

Department of computer, Karaj Branch, Islamic Azad University, Karaj, Iran

*Corresponding Author's E-mail: soheylanemati@gmail.com

Abstract

One of the primary challenges in the identification of human behavior is to identify several human behaviors since identification of various human behaviors is very hard and requires a lot of training data. In this study, using bag-of features containing 5 valuable spatial-temporal features and extracting the most valuable features from features collocation, it has been tried to increase the accuracy of identifying human behavior in videos. The accuracy of the solution proposed in this study was 96% based on the standard database of Cambridge University, KTH which had a favorable capacity compared to other similar solutions. In addition, classification of K NearestNeighbor (KNN) has been used. According to the value of bag-of feature made of human behaviors, the algorithm of classification will have high precision to identify the defined classes.

Keywords: Feature Dimensionality Reduction, Feature Vector Extraction, Recognition of Human Behavior, Classification of K NearestNeighbor (KNN), Bag-of-Features, Spatial and Temporal Feature

1. Introduction

Today, recognition and analysis of various human motion patterns following video images are cases which have increasingly been considered by many vision researchers [1]. One of the most important human motion analysis applications is the use of intelligent video surveillance systems whose objectives are to recognize and track certain objects in the sequence of images and generally understand and describe their behavior. Such systems have numerous functions and are widely used in cases such as accessing in specific areas such as military bases and important government institutions, identification to recognize the suspects, analysis of population density and influx in the commercial and entertainment centers and detection of unusual scenarios. The analysis of individuals’ activities at the scene to recognize and classify his normal and anomalous behaviors is an important objective of intelligent video surveillance system. The source of motions can be as movement or change in any part of the body; however, face or hands are usually used. Today, the focus in the field is on recognizing emotions through face and identifying hand and leg movement. The systems interfacing between man and computer can help connect between computer and user via recognizing user’s behavior. In addition, many algorithms have been made to use the computer camera and vision so that they can interpret body language. The recognition of human motion is done based on a recognition-adaption task model [2]. In fact, there are a gallery of templates in which the person’s descriptor in video have been adapted to the descriptors of a gallery of templates to recognize human behavior and the answer to the question is the recognition of the pattern that is most closely corresponds to person’s motion descriptor. These steps are required to recognize a person’s behavior: 1- Segmentation of a person’s image for separating pixels belonging to the background and deleting foreground; 2- Detecting and tracking the person; 3- Using global or local descriptor as an identifier of each person; and 4- Recognition of human behavior according to the database of a collection of known and prominent behaviors. In this study, feature-based methods are investigated to recognize human behavior and it is tried to present a smart way from providing...
spatio-temporal bag-of-features of human body to increase the accuracy of a classifier. In previous methods, one-dimensional spatio features of an image or motion features have been used and the investigation of their results show that these algorithms are unable to provide a desired result if the image is busy or noisy and suffers from lack of light. Therefore, using bag-of-feature technique and weighted features is a very clever solution to enhance the accuracy. Then, in Section 2, previous studies are reviewed and in Section 3, the processes of analyzing human behavior are evaluated. Section 4 have examined the details of human analysis stages and in Section 5, there are assessment and the results of the proposed method and the conclusion is done in the last section.

2. Literature Review

A bulk of previous studies in the field of machine vision science have focused on one of the spaces of spatio-temporal features and tried to separate the behaviors from information in the one of these two areas. In [3], the researchers describe videos through dense routes. They sampled the dense points of each frame and tracked them based on movement information of optical compact field. Local descriptors of HOG, HOF and MBH were calculated around the considered points. In [4], the analysis of spatio-temporal information including spatio information i.e. person’s appearance mode at any time and temporal information i.e. a person’s motion body position at any location were examined. Huang and Wang [5] also performed identification with the help of images of energy yxt page obtained from xyt space.

In [6], the statistical changes of texture features which inherently contains motion information were examined over time and because of the calculation of texture features on video images, information related to the appearance (position, direction and angle of view) were also entered into the problem. Since human movements such as walking also include information on spatio and temporal dimensions, the descriptors of composite texture were used for analysis. In [7], the researchers extracted human schema in every frame by taking into account the points corresponding in each frame and separating background and formed a three-dimensional volume including important spatio and temporal information by putting these schemes together. Then, the considered movement is analyzed with the help of a suitable descriptor. Noise in the efficiency of this method is very effective because the noise disrupts to gain the salient points at the edge of spatio and temporal volume.

In [8], to identify how to walk, the combined descriptors have been used and rather than describing the whole image, the advantages of local method was applied to extract key features and then, written in dictionary after describing the word. In [9], the method proposed on the KTH database was applied to recognize detect human movement and the desired results were achieved. Then, a similar method was used on the data related to the walking in the database for recognition. In [10], a format of Independent Space Analysis (ISA) algorithm was provided to learn the constant spatio and temporal features from not labeled video information in a hierarchical manner. In particular, the features have first been taken using small input pieces to vectors so that they have been convoluted with a larger area than the input data and then, used as input for higher layer. The features of both layers were combined to be classified as local features.

3. The processes of analyzing human behavior

Any human behavior analysis system in video consists of several basic stages so that some are considered pre-processing (detection of background and foreground and tracking individual in the consecutive frames) and others are considered main issue as follows: feature extraction, selection of classifier or an appropriate model, classification, identification and validation based on the extracted features. The first step in the recognition of an object behavior is to recognize moving object in image and its segmentation. The best-known solution is the background subtraction to recognize objects in motion by a sequence of video. This solution is obtained through comparing each frame of
the video with the static background. Figure 1 shows the overview of human behavior analysis system in video. The recognition aims at interpreting human gestures using mathematical algorithms. Source of movement can be considered a change in the mode of each limb or movement while in most cases, hands, feet or face are used. A bulk of algorithms have been designed and made in order to use machine camera and vision to be able to interpret body language. Recognition and identification of posture, walking, neighborliness and human behavior also consist of part of motion detection technology and human identification. Human recognition can be applied in different fields such as counting humans and security check. On the other hand, recognition of entire human body is not always possible because the whole body may not be recognizable in the population or the person is not standing. Thus, the majority of studies have focused on the identification of human upper body, especially face or head and neck rather than the entire human body. Since there is not always the face image from in front, the recognition of human via face is not still very good and the recent studies have turned to recognize head and neck.

![Figure 1: Overview of human behavior analysis system in video](image)

2. The proposed method

The objective is to develop an automated mechanism and interpret video images so that activities and local descriptors recognize video images and use them for more analysis by video surveillance system managers through labeling. Today, human behavior evaluation and recognition systems have usually special equipment and are very expensive with highly specialized technical methods. The suggestion is to use smart video surveillance system for the identification of pre-defined human behaviors via video images. In such applications, extracting objects, removing shadows and avoiding quick or late update of background are needed to reduce the system error.

The proposed method conducted on human movements for the selection of feature has used a combined method including the combination of local and global texture features and motion features of image components and in order to enhance the accuracy of adaption algorithm, the separation of features has been used into two distinctive and similar categories in the form of bag-of-feature. Figure 2 shows the flowchart of the proposed method.
In this method, the key areas are extracted from training video by a spatio and temporal detector which is a feature-based method and described using local binary pattern. At this stage, a feature vector is obtained for each key area of image. Then, feature vectors with a certain weighting coefficient will be written by the words of a bag-of-feature using an appropriate classification algorithm. Finally, each training class with a histogram of the number of word repetitions of the bag-of-feature is modeled. Thus, a unique model is achieved for each training class. For a video sequence of test, after obtaining the feature vector and comparing it with available bag-of-feature, the histogram of test sample is created. Then, with the help of an appropriate classification algorithm, the test sample can be attributed to a good training model. The first step is the detection in which the segmentation of human body or objects is done from the background in a way that the system is prepared for many high-level tasks such as video surveillance and analysis in tracking step. The next step is the adaptive update of background to select the foreground objects. For continuous update of background image with considering the effectiveness rate of images of successive frames, a linear combination equation is used as follows [11].

\[ \text{Background}_k = (1 - \alpha) \text{Background}_n + \alpha \text{I} \]  

(1)
In equation (1), $\alpha$ is updating rate of background and I is the current frame in the video. Through having an appropriate background using the background difference method of current frame, moving objects in surveillance images can be extracted as equation (2).

$$\text{Foreground} = [\text{Background}_{k+1} - I] \quad (2)$$

By selecting a good background image, the segmentation and labeling candidate areas are done with higher accuracy. In so doing, the examination of selected areas is necessary in terms of some geometric and spatial parameters and if they are approved, they will be identified and extracted.

2.1. Making feature vector

For better identification, each candidate images can be entered into the feature vector extraction stage according to the order of the scale of evaluation function and then, the dimensions from feature extraction can be entered into the next stage which is a classification such as nearest neighbor as a vector. According to the vector extraction of rotation and size-independent features for each area, a more simple operation is expected for classifier. SIFT feature is used to build rotation-independent feature vector. The models presented in this study are the hybrid models of feature selection with classifier model. Technical standards as features are the model input which are used to recognize and identify human behavior and operation. So, the performance of hybrid principal component analysis and nearest neighbor classification algorithm are used as a new hybrid method. In general, the selection of features has been designed based on different assessment criteria and are divided into two main categories:

1. Spatio-temporal features: including the feature of texture and color and image energy
2. Motion features: including movement features of image objects

Given that the classifier with the standard data collection has been trained, validation is used to optimize the parameters of each method. The calculation of LBP and HOG histograms for each block are as follows:

In order to calculate LBP for each pixel in each block [12], a neighborhood of the pixel is considered. In order for the operator not to be sensitive to the rotation of image, neighborhood is considered as circular and points whose coordinates are not exactly on the center of pixel are found through interpolation. If we assume that $g_c$ represents the central pixel, $g_i$ neighboring pixels, then, $LBP_{p,R}$ is calculated as equation (3).

$$LBP_{p,R} = \sum_{i=0}^{p-1} s(g_i - g_c)2^i \quad s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3)$$

The method used to extract texture feature is based on the composition of gradient histogram [13] and local binary algorithm. In each image pixel, a neighborhood $w \times w$ called cell is considered. In each cell, a gradient histogram with $n$ between is calculated. The gradient histogram is calculated as (Equation 4).

An image is filtered using Sobel kernels to x and y so that the image gradient is obtained to x and y [10].

$$G_x = I \ast D_x \quad (4)$$

$$G_y = I \ast D_y$$
Where I is original images, Dx and DySobel kernels to x, y, Gx and Gy image gradient to x, y and * convolution action. Then, the size and direction of gradient is obtained at each pixel as equation (5).

\[
|G(i,j)| = \sqrt{(G_x(i,j))^2 + (G_y(i,j))^2} \\
\theta_G(i,j) = \tan^{-1}\left(\frac{G_y(i,j)}{G_x(i,j)}\right)
\] (5)

Where |G| is gradient size and \(\theta_G\) gradient direction and i and j represent the number of rows and columns of the image, respectively. Gradient for each color channel is individually calculated for color images and the largest amount of each pixel is selected as the gradient vector of the pixel. Descriptor SIFT is used to recognize and describe the local features of the image. Using it, the features which are resistant to change in scale and rotation are extracted from the image. For the optimal estimation of motion in optical flow, a preliminary estimation of movement speed and its direction is useful [14]. Maximum and minimum speeds are usually considered between a frame and subsequent frames for the movement of image pixels according to the experimental conditions of the database. Thus, one 10*10 window is first considered in the first frame around each selected feature point and its correlation with one 30*30 window is calculated in the next frame known as the search window with the center of the current frame.

The area for which the maximum value is obtained for the correlation is considered the estimation of the point in the next frame. The way it is done, is shown in the following figure. At this stage, the concept of dimension reduction of data is entered using principal component analysis. Special vectors obtained in the previous stage of features are ordered based on their specific amounts from the largest to smallest values. Thus, the data components are ordered from the most important to least important ones. Here, if the dimensions of data want to be reduced, the least important components can be removed. However, this is associated with losing the qualitative amount of information. What should be done at this stage is to establish feature vector similar to equation (6) which is in fact a matrix of vectors. This matrix contains the feature vector that we want to keep.

\[
\text{FinalData} = (e_1e_2e_3...e_n)
\] (6)

If all feature vectors are put in this matrix, no data will be lost and the original data can exactly be obtained again. To obtain new data in the final stage of PCA, the transposes of feature vector matrix obtained in the previous step should only be multiplied by the transposes of normalized data, as equation (7).

\[
\text{FinalData} = \text{RowFeatureVector} \ast \text{RowDataAdjust}
\] (7)

Where Row Feature Vector is the matrix whose eigenvectors in rows have been from top to bottom according to the eigenvalues and Row Data Adjust is the matrix containing data that the average of each dimension has been declined by the dimension. In this matrix, the data have been stored in its columns and each row is related to one dimension.

2.2. K Nearest Neighbor (KNN)

In the identification of the pattern, KNN is a non-parametric method used for classification and regression. In both cases, the entry includes K-nearest training samples in the feature space.
2.3. Reduction algorithm of Principal Component Analysis (PCA)

PCA is a mathematical method converting a set of correlated observations into a value set of uncorrelated observations called the main component [15]. One of the objectives of combining and reducing feature vector is to increase the richness of feature information while reducing their number. In this study, spatial and location information were collected in a convenient feature space with appropriate dimensions. At this stage, the feature space has been reduced using this technique.

2.4. Condition of implementation and database

To perform the tests, a set of videos were first divided into educational and test videos. In so doing, a standard database called KTH human behavior dataset containing a lot of activities of human behavior set has been used and various images have been collected from the database.

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of training video</th>
<th>Number of test video</th>
<th>Number of all video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast hands and feet motion (boxing state)</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>walking</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Running</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Clapping</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Slow running</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Waving</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>10</td>
<td>32</td>
</tr>
</tbody>
</table>

3. Evaluation Criteria

To evaluate the proposed system and available methods, the precision or accuracy rate of identification have been used. Equation (8) shows the accuracy rate which is the most common criteria identified as the ratio of the number of correctly recognized classes, $N_c$ is defined as the total number of recognition operation classes of $N_{rect}$ [16].

$$P_T = \frac{N_c}{N_{rect}}$$

(8)

By averaging the rate of identification in the test stage, the average rate of identification is obtained via Equation (9).

$$ARP = \frac{1}{N_Q} \sum_{q=1}^{N_Q} Pr(q)$$

(9)
So that $N_q$ and $Pr(q)$ show the number of tests and the rate of identification per class in each test $q$, respectively.

4. The results of the proposed method

The process of identifying the behavior in this study was done by various methods which were different in terms of using motion features and type of classification. Among from classifiers, due to reducing the cost of computing in addition to reducing the complexity for few numbers of features, low cost classifiers such as SVM and KNN can be used.

<table>
<thead>
<tr>
<th>Method to extract feature</th>
<th>Accuracy of retrieve</th>
<th>Recognition accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP [12]</td>
<td>81</td>
<td>85</td>
</tr>
<tr>
<td>Color Hist FFT [17]</td>
<td>73</td>
<td>77</td>
</tr>
<tr>
<td>HOG [13]</td>
<td>89</td>
<td>91</td>
</tr>
<tr>
<td>Sift(Surf) [18]</td>
<td>86</td>
<td>89</td>
</tr>
<tr>
<td>Optical-Flow [19]</td>
<td>87</td>
<td>90</td>
</tr>
<tr>
<td>Proposed method</td>
<td>92</td>
<td>96</td>
</tr>
</tbody>
</table>

Of the compared methods, Color Hist shows the worst performance. This result was expected because this method could not be good in extracting features from images with brightness variations. The operating speed of the proposed method leads to better results compared to other methods in which features are selected based on the values of gray level.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time to extract feature (MS)</th>
<th>Length of the vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP</td>
<td>247</td>
<td>347</td>
</tr>
<tr>
<td>Color Hist FFT</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>HOG</td>
<td>378</td>
<td>952</td>
</tr>
<tr>
<td>Sift(Surf)</td>
<td>683</td>
<td>1132</td>
</tr>
<tr>
<td>Optical-Flow</td>
<td>1100</td>
<td>892</td>
</tr>
<tr>
<td>Proposed method</td>
<td>682</td>
<td>700</td>
</tr>
</tbody>
</table>
Conclusion

Human walking carries important information in two spatio-temporal dimensions. Therefore, in this study, identifying human behavior has been evaluated based on different features and on a combination of features using a more optimal selection of features in KNN classifier. The accuracy of the proposed solution is equal to 96% which enjoys a good capability compared to other similar solutions. In future studies, neural networks, support vector machine or regression may be used to increase the rate of identifying human behavior as a classification model and its combination with new methods to select features extracted from this study. Also given that in the classification, the difference of optimal K should be achieved by various tests which take a lot of time, the methods can be used where the appropriate number of clusters are obtained before any classification.

References