USING LOCAL TEMPORAL FEATURES OF BOUNDING BOXES FOR WALKING/RUNNING CLASSIFICATION

Berkay Topcu and Hakan Erdogan

Faculty of Engineering and Natural Sciences
Sabanci University, Orhanlı Tuzla 34956, Istanbul, Turkey
berkayt@su.sabanciuniv.edu, haerdogan@sabanciuniv.edu

ABSTRACT

For intelligent surveillance, one of the major tasks to achieve is to recognize activities present in the scene of interest. Human subjects are the most important elements in a surveillance system and it is crucial to classify human actions. In this paper, we tackle the problem of classifying human actions as running or walking in videos. We propose using local temporal features extracted from rectangular boxes that surround the subject of interest in each frame. We test the system using a database of hand-labeled walking and running videos. Our experiments yield a low 2.5% classification error rate using period-based features and the local speed computed using a range of frames around the current frame. Shorter range time-derivative features are not very useful since they are highly variable. Our results show that the system is able to correctly recognize running or walking activities despite differences in appearance and clothing of subjects.

Index Terms— surveillance, video signal processing, pattern classification, time domain analysis

1. INTRODUCTION

Recognition of human activities has several important surveillance applications. One of these applications is tracking of suspicious activities. This application is directly related to public security, airport security and transportation. Intelligent surveillance is about understanding/recognizing objects or events instead of just tracking an object in a scene. First aim of this kind of systems is the recognition of human presence. An intelligent surveillance system should be able to distinguish the difference between a human and another moving object and more importantly it should distinguish a suspicious act from a usual one [1].

The improvements in object detection and tracking systems in recent studies should be followed by interpreting human behaviour. Activity recognition and classification is the bridge between tracking and behaviour understanding. In this article, we study activity recognition which could be the next step after object/subject tracking.

There are mainly three approaches to interpret the human actions. In generic model recovery, human posture is modeled by 3D models that are used in 3D feature extraction. Appearance based models depend on 2D images and features are extracted from frames directly. Features to be used are usually extracted from silhouettes of subjects in 2D images [2,3]. The extracted features are compared with the training data and then classified [4]. Motion based models depend on the characteristics of human motions instead of static models of human beings [5,6,7].

Our approach is based on the change of the bounding box of a subject in time. One of the most important reasons that made us prefer this method is that it is generally relatively easier to find a box that bounds a foreground object when compared to detailed models and silhouette extraction. Also, dynamic information extracted from bounding boxes is easy to process. In addition, bounding box method is suitable to be used in real-time operations because it is faster to extract the dynamic information as compared to dynamic features extracted from silhouettes and other model based systems.

Our method performs running/walking classification in each frame. A window of 33 frames around the current frame is used to extract dynamic features for that frame. Thus, only 16 frames in the future are used. So the system is able to decide with a delay of less than one second. In this study, hand-drawn bounding boxes are used as an initial study, but in real applications, an automated object detection system should be able to create these boxes accurately.

2. PROBLEM AND THE DATABASE

2.1. Problem

The problem we want to solve is the classification of human activities as running or walking when the human of interest is bounded by a rectangular box in every frame. These bounding boxes are used for feature extraction and features are used for activity classification.

2.2. The Database

The database that is used for the training and test of the planned system consists of videos of two different actions, walking and running, of different people. In addition to walking and running videos of all people, videos of people with walking one hand full, both hands full and carrying a backpack are recorded. These videos are separated into frames and in every frame the subject is bounded by a hand-marked box. Coordinates, width and height of these bounding boxes are used for feature extraction. For the training and testing of the system, thirty videos are used (six running and twenty four walking). From these thirty videos 1931 frames are generated; 194 frames are from running videos and 1737 frames are from walking videos. Clearly, there is a class imbalance problem in our setup due to available data. However,

we deliberately do not use this information in our classification results (in terms of class priors for example) since we would like to measure average performance of such systems.

3. FEATURES

In our study, we worked on six different features that are listed below:

- Features related to the period
  - Period of width/height ratio signal (PERIOD)
  - Swing of width/height ratio signal (SWING)
- Speed of the bounding box (SPEED)
- Temporal derivative features
  - Change of width/height (DERIVATIVE)
  - Percentage change of width (W-DERIVATIVE)
  - Percentage change of height (H-DERIVATIVE)

Similar features were used for "person identification by gait recognition" problems [8], but for the problem we are working on, we have not run into these features in the literature.

As the tracked person moves, width/height ratio of the box changes and with every step of the person the ratio signal repeats itself. It can be concluded that as the period decreases the object is running. This ratio is examined within a window of 33 frames (16 frames to right, 16 frames to left and the frame we are processing at the moment). In this window, usually there exists more than one period of the ratio signal. To find the period of this signal, its autocorrelation signal is found and the distance between two consecutive peaks of the autocorrelation signal is calculated as the period.

As the second feature, swing amount of the width/height ratio is extracted. For this operation, the same window of 33 frames is used. Difference between maximum and minimum values of our ratio within this window gives us the second feature. We expect that swing amount to be higher when the subject is running due to over-stretching. These two features are demonstrated on a sample width/height signal in Figure 1.

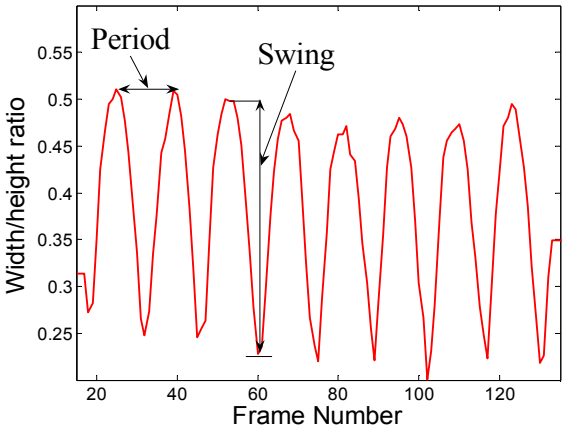

Figure 1 An example width/height ratio signal.

Our third feature is the speed of the bounding box which is calculated by using the displacement of the center point of the boxes. It is obvious that as speed increases a person is more likely to be running. As range of displacement values might vary in different videos due to the distance of the subject to camera, all displacement values are normalized with mean of the height signal within an analysis window of length 33 around the current frame. We fit two lines to the displacement signals in x-direction and y-direction within the analysis window as shown in Figure 2. Slopes of these two lines give us the speed in both directions. Square root of sum of these two speeds' squares provides the overall speed of the bounding box in each frame.

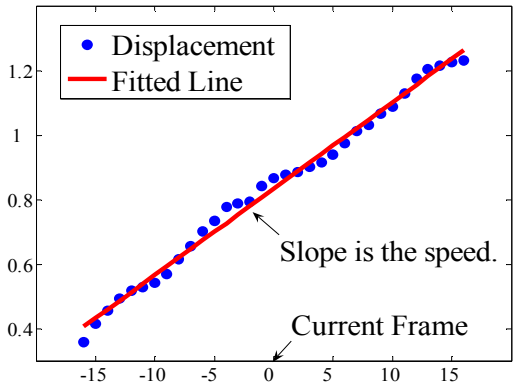

Figure 2 An example displacement signal (in x-direction) and the fitted line around current frame.

Another feature generated by using the width/height ratio is the derivative of this ratio. It can be said that change of this ratio would be faster when object is running. Let's call *Dw* as the half length of the window that is used for derivative operation. So, for every frame, *Dw* frames to right and *Dw* frame to left, together with the frame we are processing, a window of total length 2*Dw*+1 is used. For the derivative operation the formula below is used:

$$\Delta(k) = \frac{\sum_{\theta=1}^{Dw} \theta(v(k+\theta) - v(k-\theta))}{2\sum_{\theta=1}^{Dw} \theta^2}$$

In this equation, *v(.)* indicates the signal value and *Δ(.)* indicates the calculated derivative value.

The remaining two features are the normalized derivatives of the width and height of the rectangular box. Derivatives are calculated regarding to the equation above and the result is divided by the average value within the derivative window in order to normalize the value. The normalization is required because the derivative value should be independent of the scale. During running, change in bounding box's dimension is thought to be faster when compared to walking, so derivative of these changes are selected as the fifth and the sixth features.

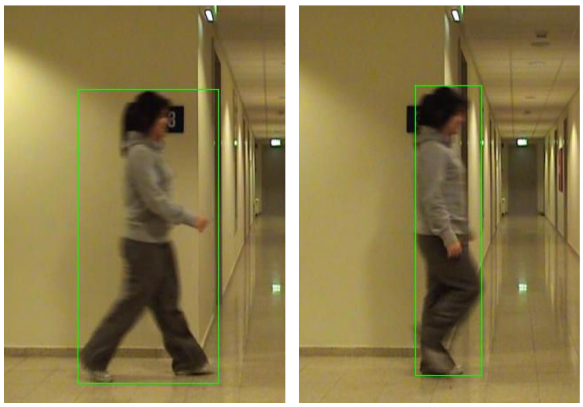

Figure 3 Examples of marked video frames.

Following the extraction of features, we go through the training and testing of the system.

4. TRAINING AND TESTING

The database used for training and testing of the system consists of 30 videos and 1931 frames (194 frames for running videos and 1737 frames for walking). A moving human, within a frame is bounded by a rectangular box that is drawn by hand. Example video frames are shown in Figure 3. Training and testing of the system is done in MATLAB environment. File reading, processing and composition of feature vectors are done via MATLAB. For classification algorithms we are using PRTools [9] which is a MATLAB toolbox. Using PRTools, system can be easily trained and can be classified quickly by different classifiers. For the training and testing of the system, we used n-fold cross-validation.

In this method, certain percentage of the feature vectors is used for training of the system and the rest is used for testing. We divided our database into six parts. Each part consisted of one running video and four walking videos. So, we made sure that training data and test data do not overlap and that they come from different videos. In every test, 5/6 of the videos are used for training and the remaining data as test data. Note that, the videos contain different number of frames, so that each cross-validation test uses different number of test frames. This operation is repeated six times, for every part of the database. At the end of all tests, 6 results are found for every classifier and weighted mean of these 6 results give us the overall classification error rates.

During testing step, every feature is used alone and results are compared. Following this only one feature comparison, we combined the features which provide the least error and trained the system once more. Test results are higher when the system is trained by more than one feature as expected and the error rate decreases significantly. The last test-run is done after combining all feature vectors together.

The error rates vary due to features and also due to classifiers. The classifiers used for our system are listed below:

1. **QDC:** Quadratic classifier assuming normal densities
2. **LDC:** Linear classifier assuming normal densities with equal covariance matrices
3. **NMC:** Nearest mean classifier
4. **PARZENC:** Parzen density based classifier
5. **KNNC:** K-nearest neighbour classifier
6. **TREEC:** Decision tree
7. **LMNC:** Neural network classifier trained by the Levenberg-Marquardt rule

LDC and NMC are simple linear classifiers with few variables. QDC is the quadratic classifier with a more flexible decision boundary. Non-parametric classifiers such as PARZENC and KNNC rely much on the training data and may not generalize well. TREEC and LMNC can be seen as more complex classifiers. We used different types of classifiers to see the effect of using complex (usually over-trained) versus simple but generalizing classifiers in this problem.

Results of the experiments can be seen in Table 1. The most successful three classifiers appear to be QDC, LMNC and PARZENC. When derivative-based dynamic features are used for training and testing of the system, high error rates around 50% are given by the NMC classifier, which is clearly a suboptimal choice.

It can be observed that when features are used separately, the best indicative feature is the speed of the bounding box. Although it has the lowest error rates, when the object walks with an angle to the camera or walks fast, speed cannot provide all the information about the action and classification error rates may increase. A successful feature is the period of the width/height ratio signal. Swing feature does not yield successful classification when used by itself, but when we combine it with the period and speed; we obtain improved classification rates for all classifiers. When we use three features (period, swing and speed), the test results are very close to the result we get by combining all features. It is worth to mention that these three features with QDC classifier yield the best result of 2.38% error. It appears that the derivatives are not as helpful in classification of human actions as the speed and the period based features. We conjecture that the derivatives contain shorter range information which may be misleading in classifying actions. Still, for some classifiers (such as TREEC), it is interesting to see that we gain by including derivative features in our feature set. Thus, more extensive experiments may be needed to confirm the *uselessness* of derivative-based features.

5. DISCUSSION

Test results demonstrate that the performance of our walking/running classification system is satisfactory. Although the error rates are low, it should be noted that data used for training and testing of the system are different activities of the same people recorded according to a scenario. As a result, a limited set of data is used in this study. Scenarios recorded in videos are scripted and they are not complicated as real videos. Any actions other than walking and running are not included in the videos. There are no moving objects in the background and the bounding box that surrounds our target object is drawn by hand.

	QDC	LDC	NMC	PARZENC	KNNC	TREEC	LMNC
PERIOD	9.84	9.99	5.96	**4.51**	5.54	10.20	6.84
SWING	10.05	11.13	28.22	18.18	16.93	16.93	13.98
DERIVATIVE	10.05	10.05	49.72	10.05	10.05	17.97	10.05
W- DERIVATIVE	10.05	10.05	49.87	10.20	10.15	16.57	10.46
H- DERIVATIVE	10.41	10.25	52.67	9.79	9.94	11.91	10.41
SPEED	4.97	4.97	5.80	3.47	4.09	7.82	**3.31**
PERIOD+SPEED	4.51	4.19	5.96	4.51	3.42	4.19	**2.54**
PERIOD+SWING+SPEED	**2.38**	2.95	5.96	4.30	3.94	7.72	4.82
ALL FEATURES	**2.59**	3.06	5.96	4.30	3.83	4.92	3.06

Table 1 Percentage Classification Error Rates (%)

In real applications, as the environment diverges from ideal case, results will degrade. Thus, classification error rates much higher than 2-3% should be expected. It should be noted that box extraction is a difficult task because noise, illumination, occlusion and view point changes lead to wrong foreground estimation causing the error rates to increase. As our future work, we will work on non-ideal scenarios and automatically marked videos.

In addition, classification of actions is done on a frame-by-frame basis. This is important in real-life applications because a tracked person might walk for a period of time then start running or vice versa. If it is required to classify activity in a video (or within a range of frames) majority voting can be applied. Temporal information can also be used by employing a hidden Markov model that uses classifier posterior probabilities as features for activity classification.

6. CONCLUSION

In this study, dynamic bounding box based features are studied to classify human actions as walking and running.

As far as we know, these features are not used in order to classify human actions and they are innovative. In testing a database of videos of running and walking, high success rates are achieved. In the future, these algorithms will be applied and developed on a more realistic and non-ideal database.

7. ACKNOWLEDGEMENTS

We would like to thank Ahmet Tüysüzoğlu and Sabanci University PROJ102 students who helped in collection and arrangement of the database and hand-marking of the bounding boxes used in this study.

REFERENCES

[1] Masoud, O. and Papanikolopoulos, N., "Recognizing Human Activities", *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2003.

[2] Han, X., Liu, J., Li, L. And Wang, Z., "Gait Recognition Considering Directions of Walking", *IEEE Conference on Cybernetics and Intelligent Systems*i 2006.

[3] Wang, L., Ning, H., Tan, T. And Hu,. W., "Fusion of Static and Dynamic Body Biometrics for Gait Recognition", *IEEE Transactions on Circuits and Systems for Video Technology,* vol.14, no.2, February 2004.

[4] Robertson, N. and Reid, I., "Behaviour Understanding in Video: A Combined Method", *IEEE International Conference on Computer Vision,* 2005.

[5] Gavrilla, D.M., "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding,* vol.73, no.1, pp.82-98, 1999.

[6] Elbaşı, E., Zuo, L., Mehrotra, K., Mohan, C. and Varshney, P., "Control Charts Approach for Scenario Recognition in Video Sequences", *Turk J Elec Engin,* vol.13, no.3, 2005.

[7] Efros, A., Berg, A., Mori, G. and Malik, J., "Recognizing Action at a Distance", *IEEE International Conference on Computer Vision,* 2003.

[8] Amit Kale, et al, "Identification of Humans Using Gait", *IEEE Trans. Image Processing*, Vol. 13, pp. 1163-1173, Sept. 2004.

[9] Van der Heijden, F., Duin, R., De Ridder, D. and Tax, D. Classification, parameter estimation and state estimation - an engineering approach using Matlab, , John Wiley & Sons, 424 pages, ISBN 0470090138, 2004.