Wang, J, Xu, Zhijie and Hoder, M.E.

Fast human detection for video event recognition

Original Citation

Wang, J, Xu, Zhijie and Hoder, M.E. (2009) Fast human detection for video event recognition. In: Proceedings of Computing and Engineering Annual Researchers' Conference 2009: CEARC’09. University of Huddersfield, Huddersfield, pp. 70-75. ISBN 9781862180857

FAST HUMAN DETECTION FOR VIDEO EVENT RECOGNITION

J. Wang[1], Z. Xu[1] and M.E.Hoder[1]
[1] University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

ABSTRACT

Human body detection, which has become a research hotspot during the last two years, can be used in many video content analysis applications. This paper investigates a fast human detection method for volume based video event detection. Compared with other object detection systems, human body detection brings more challenge due to threshold problems coming from a wide range of dynamic properties. Motivated by approaches successfully introduced in facial recognition applications, it adapts and adopts feature extraction and machine learning mechanism to classify certain areas from video frames. This method starts from the extraction of Haar-like features from large numbers of sample images for well-regulated feature distribution and is followed by AdaBoost learning and detection algorithm for pattern classification. Experiment on the classifier proves the Haar-like feature based machine learning mechanism can provide a fast and steady result for human body detection and can be further applied to reduce negative aspects in human modelling and analysis for volume based event detection.

Keywords Human detection; Haar-like features; AdBoosts learning algorithm

1 INTRODUCTION

Digital analysis on human activities from video clips is all based on the accurate calculation of human body recognition and modelling. This paper focuses on a fast human recognition system through a subject of research known as "Volume based event detection" which intends to develop an advanced intelligent video surveillance system. It introduced 3D volume structure for analysing dynamic information in a video. Before translating data from 2D pixels to 3D voxels, useful 2D properties should be abstracted in a frame-by-frame "pre-processing" stage. More information about the whole research project can be referred in a recently publicised paper (Wang et al (2009))

Non-rigid human action detection with different clothes colours and gestures is more difficult than facial detection because the difference between one face and another face changes in a confined range. This paper represents the human features with machine learning algorithm. It adapts and adopts Haar-like features which are not sensitive to colours and detailed gestures but can be effective on objects with uniform shapes. It also introduces the AdaBoost algorithm to learn Haar-like features from different samples.

Experiments detailed later in this paper show distribution of Haar-like features on the human body. The first Haar-like features are easily located on the head, hands and feet of human. This is an important result because these locations also provide key features for human stick models. Stability of the points also affects quality of volume event detection. In order to reduce false positive rate in this experiment, appropriate segmentation and image shape analysis methods are applied after the recognition algorithm. It is an effective and efficient approach to filter error classification generated by AdaBoost.

The remaining sections of the paper will discuss the implementation of the detector, related theory, and experiments. Section 2 provides a brief review on the existing object detection methods. Section 3 introduces the proposed Haar-like features calculation method. Section 4 Highlights the AdaBoost algorithm with programming details. The result is further analyzed and discussed in Section 5. Section 6 covers the conclusions and future work.

2 LITERATURE REVIEW

Based on image segmentation technique, traditional human detection approaches combines certain pixels together and map these image features to a specific mathematic human skeleton model (Huo et al (2007)). It builds pixel blocks with different segmentation approaches and analyses these shapes by using image morphological approaches. The fundamental issue is to locate different pixel parts in one image and translate them into a human feature model, such as a matchstick model, based on shape features. But this human modelling method ignores the relationship of different characters and depends on the value of every single pixel excessively. This disadvantage limits the usability of the approach because it often abstracts the human body from an image with poor resolution, too much noise or serious occlusion problems.

To solve this problem, many research groups have investigated the pattern analysis machine learning technique. Schneiderman et al (2000) described a statistical method for 3D object detection using a product of histograms. Rowley et al (1998) presented a multiple neural networks-based upright frontal face detection system by using bootstrap algorithm, which adds false detections into the training set as training progresses. These methods gained great success in face detection, but they are difficult to transplant directly to solve human body recognition problem due to the wide changes among groups of people such as the colour of clothing or body size which brings much more noise in the feature space. One significant problem is that non-rigid human action brings unstable quantitative feature and machine learning results based on these erratic features always "confuse" the recognition system.

Recently, Di et al (2009) researched elastic motion based on Hidden Markov Models for motion estimation which can provide stable human gesture feature. This feature abstract and learning system contains two correlated issues which emphasize the spatial smoothness and temporal continuity, respectively. This project combines these related aspects into a single 3D model and draws the dependence between their hidden states and adopted Mixture of Transformed Hidden Markov Models (MTHMM) for machine learning algorithm.

Additionally, Wang (2009) represents a video by a novel "bag-of-words" representation and denotes each frame a "word" to learn the Semilatent topic models for human action recognition. The unknown event in that model corresponds to class labels. This feature brings a convenient learning process for labelled training sets

3 FEATURE ABSTRACTION

3.1 Haar-like Features

As mentioned in Section 1, the human detection technique is a pre-analysis step in a volume-based event detection project. The time-consuming feature extraction and pattern analysis approaches should not be adopted in pre-analysis steps. A suitable method is to utilise Haar-like features introduced by Viola et al (2001) which was applied in face detection projects successfully and motivated in part by the work of Papageorgiou et al (1998). It is sensitive to an object with high contrast parts and is also efficient in denoting specific objects with distinct shape features. Compared with other feature abstraction approaches, Haar-like features can be applied even on a simple gray level image. Based on the integral image representation for images, Haar based function extract features uses only simple addition operations. Operating on 384 by 288 pixel images, faces are detected at 21 frames per second on AMD Athlon 2.62GHz GPU with 2G RAM.

More specifically, it uses three kinds of features. As shown in figure 1, Two-rectangle features are shown in (A) and (B). (C) shows a three-rectangle feature, and (D) a four-rectangle feature. The value of a two-rectangle feature is the difference between the sums of the pixels within two rectangular regions. The regions have the same size and shape and are horizontally or vertically adjacent. A three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally a four rectangle feature computes the difference between diagonal pairs of rectangles. These features are calculated on each pixel and also on different rectangle sizes. Because each location also map different rectangle sizes, Haar-like features combine an over-complete feature set. For example, for abstracting features from a 24X24 size gray level image, it generates 86400 horizon and vertical two-rectangle features, 27600 three-rectangle features and 20736 four-rectangle features.

3.2 Integral Image

Computing these features based on the definition directly is a time-consuming task. As well as the relationship of each rectangle feature, Viola also introduced an intermediate representation data format known as integral image. Rectangle features can be computed very rapidly after building an integral image for any images. Integral value on image point (*x,y*) contains sum of all pixels above and to the left of *x,y.* It also can be described in the following expressions:

$$int(x,y) = \sum_{x' \le x, y' \le y} img(x',y') \quad , \qquad (1)$$

where *int*(*x*,*y*) is integral image and *img*(*x*,*y*) is original image. As shown in figure 2, the pixels under the gray area are all counted. Because the larger gray area contains the integral result of the small one, it also can be computed by the following recurrences:

$$s(x,y) = s(x,y-1) + img(x,y), \tag{2}$$

$$int(x,y) = int(x-1,y) + s(x,y), \tag{3}$$

Where *s*(*x*,*y*) is the cumulative row sum. The initial value of *s* defined as *s*(*x*,-1)=0 and *s*(-1,*y*)=0. The integral image can be computed in one pass over the original image.

Haar-like rectangle features can be calculated conveniently after the integral image is built. As shown in figure 3, rectangle "4" with arbitrary location and size can be modelled by four coordinate values from corners (Viola named these value as four array references). A, B, C and D denotes the integral value of these corners. The sum within "4" can be computed by four signed addition operation: D+A-B-C.

4 MACHINE LEARNING FUNCTIONS

The machine learning algorithm intends to minimize the error probability of feature classifier. Haar-like features based human detection simplify the classifier to a dynamic threshold which can minimize detection errors according to feature values of samples. In this project, AdaBoost is used both to select the features and to train the classifier (Freund et al (1995)). It does this by combining a collection of weak classification functions to form a stronger classifier. As shown in figure 4, the blue rectangles (negative sample) and red balls (positive sample) are mixed together. It is hard to draw a straight line (threshold in weak classifier) to separate them. The optimal straight line minimizes the error probability. After the first learning step, the samples which are in the false positive and false negative area should be given more emphasis (weight). The second learning cycle still obeys the "minimum error" rule. This straight line is totally different from first one. If these two lines are combined together then compose a poly-line (strong classifier), the finial classifier can separate these samples accurately.

One practical method for completing this analogy is to restrict the weak classifier to the set of classification functions. In support of this goal, the weak classifier algorithm is designed to select the single rectangle feature which best separates the positive and negative examples. For each feature, the learner determines the optimal threshold classification function, such that the minimum number of examples is misclassified.

Weak classifier can be described as:

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) > p_j \theta_j \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $h_j(x)$ denotes a weak classifier which is restricted to using a single feature. j denotes feature index and x denotes the whole image. $f_j(x)$ denotes a sigle Haar-like feature value. p is parity indicating. The most important value is threshold (θ_j). It is determined by feature value f_j on all positive and negative samples. Mahajan et al (2002) introduced the lowest exponential error function to solve this problem. AdaBoost learning algorithm can be described in figure 5. It involves two main steps: construction of weak classifier and strong classifier

5 EXPERIMENT RESULT

To assess the devised Haar-like features and AdaBoost learning algorithm, a set of experiments have been designed and carried out. The software tools and APIs used in those experiments include, MATLAB, LabVIEW, and the system prototype is implemented in VC++ on an AMD Athlon 2.62GHz CPU with 2G RAM.

The experiments adopt MIT pedestrian database (Papageorgiou et al (2001)) as positive training data. Some typical pedestrian pictures are shown in figure 6. This database contains 924 pedestrian pictures aligned to a base resolution of 128 by 64 pixels and includes frontal, lateral and rear pedestrian samples. In this experiment, these sample pictures and symmetrical pictures are chosen to train the classifier. Negative training data comes from images downloaded from a random crawl of the World Wide Web. It also contains 924 by 2 pictures (original and symmetrical samples) and they share the same size with positive samples. These samples are hand labelled on the file with prefixes where "P" means positive and "N" means negative.

Figure 7 shows a sample processed by10X10 horizontal two-rectangle Haar-like features. A is the original positive image, B is the related integral image and C is Haar-like feature. It shows that the horizontal Haar-like features are sensitive to the human head, hands and feet. Figure 8 shows more 10X10 horizontal two-rectangle Haar-like features on different positive samples and the same conclusion can also be summarized.

This project selects 11 weak classifiers to compose a strong classifier based on the algorithm introduced in Section 4. The changing of threshold for weak classifier on 7 typical features and the detection rate on strong classifier changes during these 11 cycles are shown in Figure 9 and 10. It reflects the key points of the learning process. Each weak classifier shows approximately the same false rate when the locale threshold is optimized. The false rate is reduced after applied the strong classifier because each weak classifier provides the most credible classification results based on their threshold.

One disadvantage of AdaBoost is the high false positive rate. Viola and Jones introduced a Cascade classifier with 32 layers in their research work. As pre-processing parts in volume-based event detection, making use of the continuity of video frames is a more efficient and easy approach. It adopted the segmentation result that only classified the moving parts in a frame. It also combines the results which are added together. Skin colour detector (Vassili et al (2003)) predicts the probable human location. Based on the research of Peer et al. (2003), the (R,G,B) is classified as skin if R > 95 and G > 40 and B > 20 in 24 bit colour images. It initializes the YCrCb colour space from RGB value because the channel Y is sensitive to a wide range of skin colour in colour space. Every possible human location based on the Y channel is extracted in a region of interest (ROI) by simple image threshold and related morphology operations.

After these "filters", the false positive rate is reduced significantly. The final detection result is shown in figure 11. It is a real-time situation by using a CCTV camera. The detection speed is 21 frames per second which is roughly 15 times faster than the Rowley-Baluja-Kanade detector (Rowley et al (1998)).

6 CONCLUSION AND FUTURE WORKS

During the last few decades, Haar-like features has been successfully applied on the facial detection system. This paper was motivated by the advantage of Haar-like features and adapts this technique to human detection system. It focused on an available fast human body detection approach for pre-processing of volume based event detection.

The first contribution of this paper is research on the sensitivity of Haar-like features on the human body. Experiment results show Haar-like features are sensitive to areas of head, hands and feet which are also key feature points for stick human modelling. The second contribution is an efficient process to reduce the false positive rate. Considering the comparability of video frames series, the new method uses many filters to improve detection results. These image processing based approaches simplify the steps to reduce the error rate but still compose stable results.

The next stage of this research will be carried on extraction of human feature points and composing 3D feature volume for further applications. These subjects are based on the accuracy of human location detection. Furthermore, Haar-like features which are used to detect humans can be used for locating key characters such as head, hands and feet. Volume structure will be established only by these useful key feature points which can reduce the processing time on complicated voxel data and related high-dimensional data structure.

REFERENCES

DI H., RAO L. and XU G. (2009), *A Mixture of Transformed Hidden Markov Models for Elastic Motion Estimation*. IEEE Transactions on PAMI, Vol. 31, No.10, pp.1817-1830.

FREUND Y. and SCHAPIRE R.E. (1995), *A Decision-theoretic Generalization of Online Learning and an Application to Boosting*. In Computational Learning Theory: Eurocolt '95, pp.23-37.

HUO C. and ZHANG H. (2007), *Human Skeleton Reconstruction Technique Based on the Marked-Points.* Second Workshop on Digital Media and its Application in Museum & Heritages, pp.177-181.

MAHAJAN R., BELLOVIN S.M. and FLOYD S. (2002), *Controlling High Bandwidth Aggregates in the Network.* Computer Communication Review, Vol.32, No.3, pp.75-85.

PAPAGEORGIOU C.P., OREN M. and POGGIO T. (1998), *A General Framework for Object Detection*. IEEE Conference on Computer Vision, pp.555-562.

PAPAGEORGIOU C.P.and POGGIO T. (2000), *A Trainable System for Object Detection*. IJCV, Vol. 38, No.1, pp.15-33.

PEER, P., KOVAC, J., and SOLINA, F. (2003), *Human skin colour clustering for face detection.* In submitted to EUROCON 2003 – International Conference on Computer as a Tool. pp.45-51.

ROWLEY H.A., BALUJA S. and KANADE T. (1998), *Neural Network-Based Face Detection*. CVPR, [1]

SCHNEIDERMAN H. and KANADE T. (2000), *A Statistical Method for 3D Object Detection Applied to Faces and Cars*. IEEE Conference on CVPR 2000, Vol.1, pp.746-751.

VIOLA P. and JONES M. (2001), *Rapid Object Detection Using a Boosted Cascade of Simple Feature*. IEEE Conference on CGPR, Vol.1, pp.511-518.

VASSILI V.V., SAZONOV V. and ANDREEVA A. (2003), *A Survey on Pixel-Based Skin Color Detection Techniques*. Graphicon-2003, pp.85-92.

WANG J., XU Z. and XU Q. (2009), *Video Volume Segmentation for Event Detection*. IEEE Conference on CGIV 2009, pp.311-316.

WANG Y.(2009), Human *Action Recognition by Semilatent Topic Models*. IEEE Transactions on PAMI, Vol. 31, No.10, pp.1762-1774.

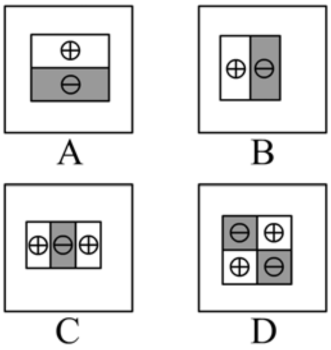

Figure 1. Haar-like features

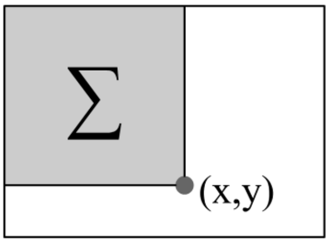

Figure 2. Integral image

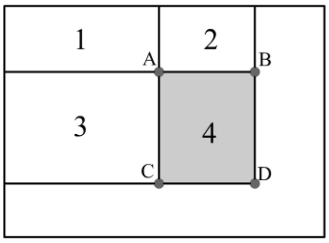

Figure 3. Rectangle feature

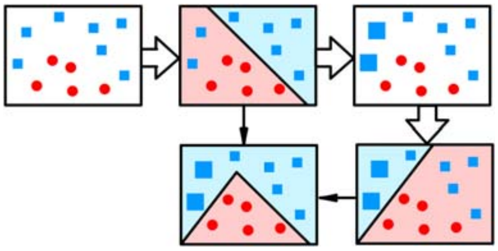

Figure 4. AdaBoost learning algorithm

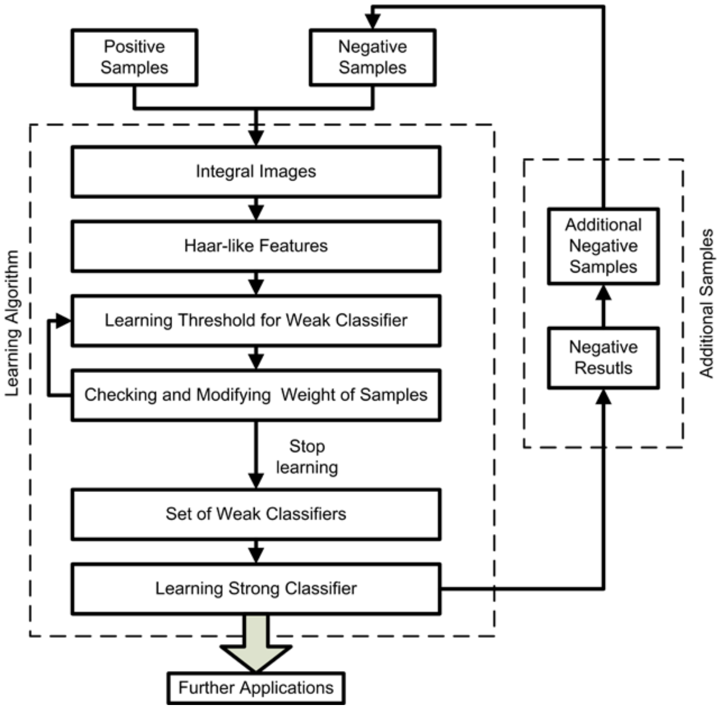

Figure 5. AdaBoost learning flowchart

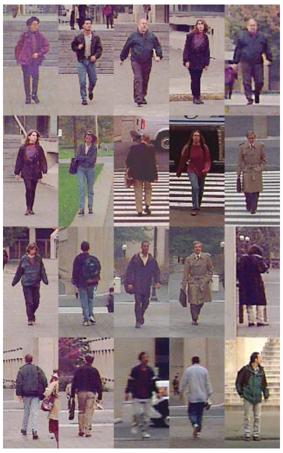

Figure 6. Positive samples

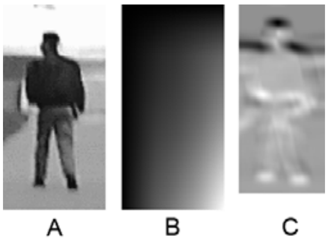

Figure 7. 10X10 horizontal two-rectangle Haar-like features.

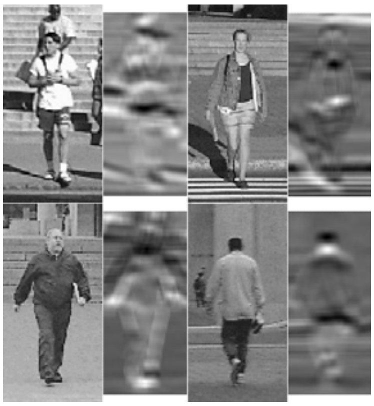

Figure 8. Sensitive area of Haar-like features

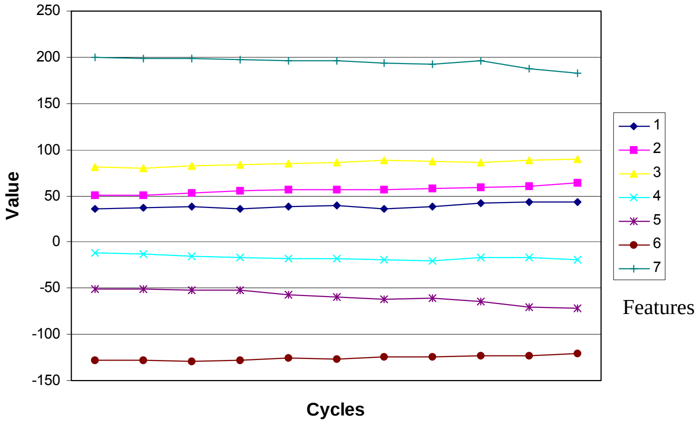

Figure 9. Threshold changes during learning cycles

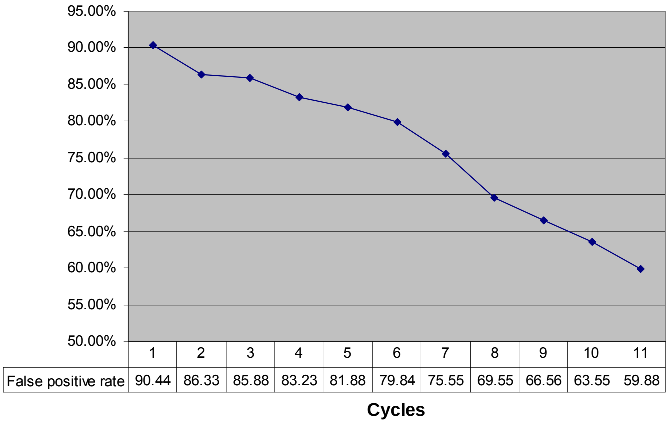

Figure 10. False positive rate changes during learning cycles

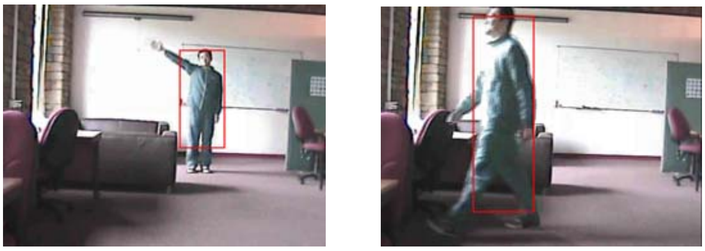

Figure 11. Real time detection