Component-based Human Detection

Bhaskar Chakraborty[*], Ignasi Rius[*], Marco Pedersoli[*], Mikhail Mozerov[*] and Jordi Gonzàlez[+]

** Computer Vision Centre,Universitat Autonoma de Barcelona, Edifici O Campus UAB, 08193 Bellatera, Spain, Spain*
E-mail: bhaskar@cvc.uab.es

+Institut de Robotica I Informatica Industrial(UPC-CSIC), Edifici U, 08028 Barcelona, Spain

Abstract In this paper, we present a general framework for human detection in a video sequence by components. The technique is demonstrated by developing a system that locates people in the cluttered scenes where they are performing certain actions like walking, running ect. The system is structured with main three distinct examle-based detectors that are trained to find separately the three components of the human body: head, legs and arms. Some geometric constraints are applied over those detected components to ensure that those are present in the proper geometric configuration. In this way the system ultimstely detects a person. Here we have dveloped the example-based detectors which are view invariant. To achieve this we have designed four sub-classifier for the head and arms taking into account the different positions those body parts can have while a human performing some action. Experimental results shown here can be compared with similar full-body detector. The algorithm is also very robust in that it is capable of locating partially occluded views of people and people whose body parts have little contrast with the background

Keywords: Object detection, people detection, pattern recognition, machine learning, components

1 Introduction

In this paper, we present a general algorithm for detecting human in images by first locating their constituent components and then combing the component detection by applying some geometrical constrains to ensure that the configuration is valid. This method can be applied to the problem of people detection in complex action sequences e.g. walking, running, boxing etc.

Human detection in an action sequence is a difficult task since human movement causes the change in the pose in every video frame sequences and a careful observation can reveal the fact that the human body parts in an action like walking, running etc changes their poses with high variability e.g. if a person performs walking in a circular path then the position of head, legs and arms changes with great diversity and sometimes some parts become occluded so it is quite difficult to detect people in this kind of movement by a single human detector. Further a robust human detection system should be able to detect human in uneven illumination, with some body parts occluded or rotated or mixed with background.

Our focus on the problem of detecting people detecting people in images can be applied in human action detection, surveillance systems, driver assistance systems, and image indexing. The ability to detect people when the limbs are in different relative positions is a desirable trait of a robust person detection system since while performing actions the pose of the body limbs change a lot. The pose variation of different body parts of the people performing different action are illustrate in Fig 1.

2 Previous Work

The methods of detection of objects can be classified in one of three major categories. They are model-based object detection, image invariance methods and example-based learning algorithms [9].

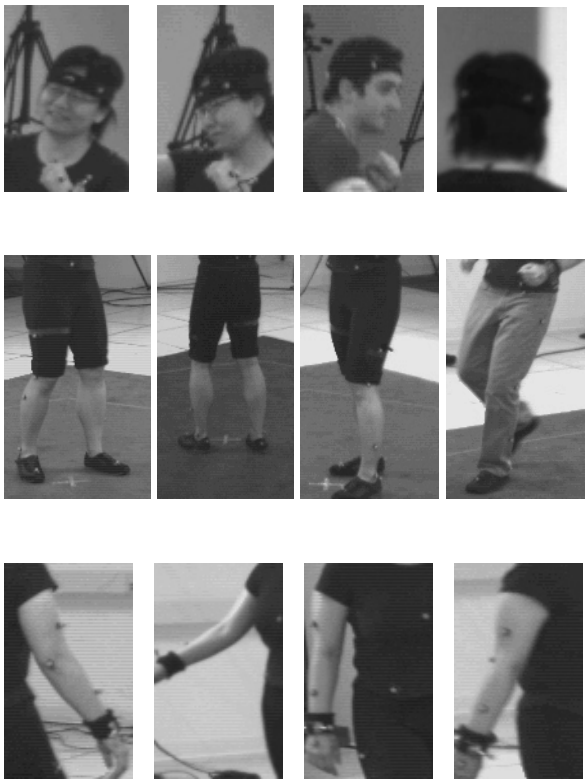

Figure 1. These images demonstrate some of the challenges involved with detecting people in still images where the positions of their body parts changes with great variety while performing some actions like walking, running etc.

Among those the working principal of example-based systems is to learn the relevant features from sets of labelled positive and negative examples which can be successfully used in other areas of computer vision, including object recognition [4]. Relevant work in the most human detection systems either use motion information, explicit models, a static camera, single person in the image, or implement tracking rather than pure detection [1], [2], [5], [6].

To improve the people detection method there come Component-Based Object Detection Systems. This approach searches for an object by looking for its identifying components, rather than the whole object. An example of such a system is a face detection system that finds a face when it locates a pair of eyes, a nose, and a mouth in the proper configuration. These systems have two common features: They all have component detectors that identify candidate components in an image and they all have a means to integrate these components and determine if together they define a face. It is worth mentioning that a component-based object detection system for people is harder to realize than one for faces because the geometry of the human body is less constrained than that of the human face.

Previous work like [9] uses Haar-Like feature and SVM for component wise object detection. But Haar-Like features are too generalized and cannot obtain certain special structural feature that can be useful to design a view invariant human detection. Also in that work there was no method to select best feature for the SVM. If some feature selection technique can be applied then the performance of SVM would be improved greatly since in this way we can minimized the feature vector size for SVM. This present work focuses all this aspect.

The paper is organised as follows. In Section 3 our approach is described in detail. Section 4 reports on the performance of our system. Conclusions and future research are in Section 5.

3 Detection of Human Body Parts

The overall structure of the present system is to detect the human body parts e.g. head, leg and arms and then combining those body parts to detect the full human. The body parts are combined based on the proper geometric configuration. The performance can be improved greatly if one classifier can be used for detection of human after checking the body parts components are in their proper geometric configuration. To unsure the view invariant human detection for each body part more than one detector has been design and the knowledge of each of those body part detectors are combined finally to increase the robustness of the whole system.

The component-based human detection has some inherent advantages over existing techniques. After detection of human body parts some geometric information is used which supplements the visual information present in an image and

thereby improve the overall performance of the system. In contrast, a full-body person detector relies solely on visual information and does not take full advantage of the known geometric properties of the human body. The detection of human body pattern sometimes becomes difficult as a whole due to variation in lighting condition and orientation. The effect of uneven illumination and varying viewpoint on body components (like the head, arms, and legs) is less pronounced and, hence, they are comparatively easier to identify.

The other problem in full human detection is that the system fails to detect the human where body parts are partially occluded. This partial occlusion is accomplished by designing the system, using an appropriate geometric combination algorithm, so that it detects people even if all of their components are not detected.

3.1 System Architecture

This section explains the overall working principal of our system when it is applied to an image. The system starts detecting people in images by selecting a 72 x 48 pixel window from the top left corner of the image as an input for head, 184 x 108 pixel window for leg and 124 x 64 for arms. These inputs are then independently classified as either a respective body parts or a non-body part and finally those are combined into a proper geometrical configuration in a 264 x 124 pixel window as a person. All of these candidate regions are processed by the respective component detectors to find the strongest candidate components.

The component detectors process the candidate regions by applying the modified Histogram of Oriented Gradient (HOG) features and then these features become resultant data vector for respective Support Vector Machine (SVM). Then a standard deviation based feature selection method is applied to take those features where the standard deviations of oriented gradients are less than one predefined threshold. These feature extraction and feature selection methods are described in the following section.

The component classifiers are quadratic SVMs which are trained prior to use in the detection process. The strongest candidate component is the one that produces the highest positive raw output, as the component score, when classified by the component classifiers. If the highest component score for a particular component is negative, i.e., the component detector in question did not find a component in the geometrically permissible area, then it is discarded as false positive.

The raw output of an SVM is a rough measure of how well a classified data point fits in with its designated class and is defined in Section 2.2.1. The each component where the component score is highest is taken to check whether they are in proper geometrical configuration with the 264 x 124 pixel window. The image itself is processed at several sizes. This allows the system to detect various sizes of people at any location in an image.

3.1.1 Feature Extraction and Selection

In our approach for the body part detector the modified HOG feature is used. The feature extraction method is described below. HOG features are extracted from a 8x8 pixel window from top left hand corner of the image. In that window Sobel gradient operator is applied using the mask shown in the Fig 2. After that the gradient is calculated in each of the pixels into that 8x8 mask. The gradients are divided into 6 bins from -900 to +900 and Sobel magnitude is added into corresponding bin where the gradient falls. In this way that 8 x 8 pixel window slides 4 pixels at a time into the total area reserved for head (72 x 48), leg (184 x 108) or arms (124 x 64). In this way for each of this 8x8 pixel window 6 feature vector can be obtained.

Next step is to select the best 6 feature packs obtain from the method described above. This feature selection method is based on standard deviation i.e. σ. For each position of that 8x8 pixel window the σ is calculated for each of the gradient of that 6 bin taking into account the total training image. Now the σ value has been sorted and those 6 feature packs are taken where the σ is less than a predefined threshold value. In this way the feature

size is minimized and those features are fed into the corresponding detector.

-1	0	+1
-2	0	+2
-1	0	+1

G_x

+1	+2	+1
0	0	0
-1	-2	-1

G_y

Figure 2. Sobel Gradient operator for x and y direction

3.1.2 Body-part Detectors

In our approach there are main three body part detector e.g. head, leg and arm. To make the detection view invariant more than one detector has been designed for each of those body part detectors. Constructions of those detectors have been described below.

To identify human in to one particular scale of image each of the individual body part detectors has been applied simultaneously. In the present system there are four head detector one leg detector and four arm detector. The four head detectors are for the view angle 450 to 1350, 1350 to 2250, 2250 to 3150 and 3150 to 450. For arm, there are four classifiers corresponding different position of arms. We use support vector machines (SVM) to classify the data vectors resulting from the feature extraction and feature selection method of the components.

SVMs were proposed by Vapnik [8] and have yielded excellent results in various data classification tasks, including human detection. The SVM algorithm uses structural risk minimization to find the hyper-plane that optimally separates two classes of objects.

This is equivalent to minimizing a bound on generalization error. The optimal hyper-plane is computed as a decision surface of the form:

$$f(x)=\mathrm{sgn}(g(x)), \quad (1)$$

where

$$g(x)=(\sum^{l*}_{i=1} y_i \alpha_i K(x, x^*_i) + b), \quad (2)$$

being K one of many possible kernel functions; yi € {-1, 1} is the class label of the data point xi; and {xi*}l* i = 1 is a subset of the training data set. The values of xi are called support vectors and the points from the data set define the separating hyper-plane. Finally, the coefficients αi and b are determined by solving a large-scale quadratic programming problem. The kernel function K that is used in the component classifiers is a quadratic polynomial and is K(x, x*i) = (x . x*i + 1)2. In (1), f(x) € {-1, 1} is referred to as the binary class of the data point x which is being classified by the SVM.

As (1) shows, the binary class of a data point is the sign of the raw output g(x) of the SVM classifier. The raw output of an SVM classifier is the distance of a data point from the decision hyper-plane. In general, greater the magnitude of the raw output, more likely a classified data point belongs to the binary class it is grouped into by the SVM classifier.

Combining those components which have been detected by classifiers is done based on geometrical configuration. Since the head is most stable part of the body so combination has been done first considering the head then leg component is taken into account after those arms have been combined.

4 Experimental Results

In our system we have developed the human body part component detector using our own developed human component database for head, arm and legs. We have chosen HumanEva Data[1] for

[1] http://vision.cs.brown.edu/humaneva/.

building the database. In that HumanEva Data set there are four action sequences for different agent performing different actions. Here one sequence is chosen for database creation and other sequence for testing the performance of those detectors. Fig 3 shows some of our database samples for head, legs and arms.

For four head classifier we use 10^4 positive and 10^4 negative examples each. For legs, we use 104500 positive and 20000 negative examples. For arms the numbers are same. We developed our own database since there is no proper database for human body component with proper aspect ratio. In Figs 4, 5 and 6, the human detection result are depicted.

Figure 3. The is showing examples of heads legs and arms that were used to train the respective component detectors.

Figure 4. Results from the test image database for head detection.

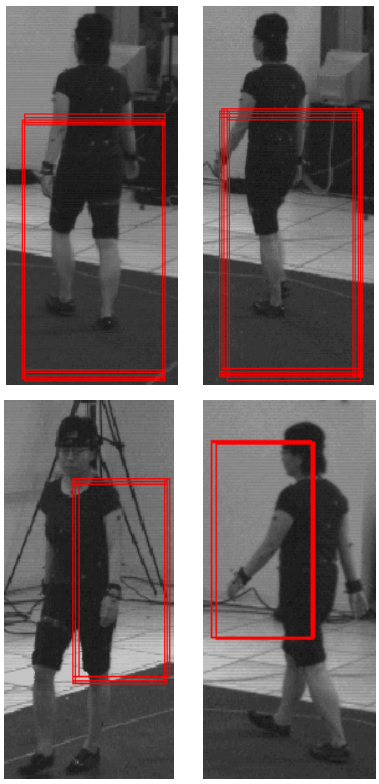

Figure 5. Results from the test image database for legs and arms detection.

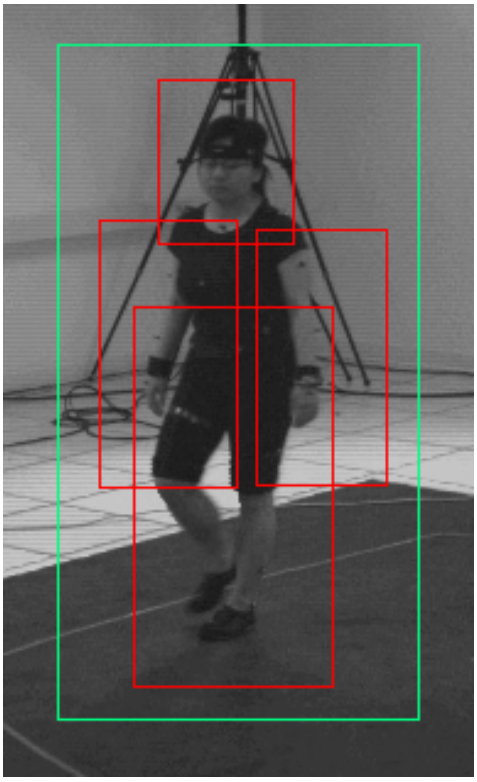

Figure 6. These results demonstrate the capability of the system in detecting people performing some actions, i.e. people moving in a circular path etc. The detection of isolated component detectors shown in Figs. 4 and 5 allows to apply geometrical constrains to perform full human-body detection.

5 Conclusions

In this paper, we have presented a component-based human detection system that is able to detect view invariant and partially occluded people in scenes without assuming any a priori knowledge concerning the image.

A component-based detector looks for the constituent components of a person and if one of these components is not detected, due to an occlusion or because the person is rotated into the plane of the image, the system can still detect the person if the component detections are combined using an appropriate hierarchical classifier.

The framework described here will be applicable to other domains for human action detection. Since human action can be viewed as a composite action of various parts of human body, motion of the human body parts are combined here for detecting moving actions like running, walking etc. Thus, a component-based approach handles variations in lighting and noise in an image better than a full-body person detector and is able to detect partially occluded people and people who are rotated in depth, without any additional modifications to the system.

References

[1] A. Micilotta E. Ong, R. Bowden, "Detection and Tracking of Humans by Probabilistic Body Part Assembly", *In Proc. BMVC05. Oxford UK. Sept 2005. Vol 1, pp429-438*

[2] D. Ramanan, D. Forsyth, A. Zisserman, "Tracking People by Learning Their Appearance", *IEEE Transaction on PAMI, vol. 29, no. 1, pp. 65-81, 2007.*.

[3] X. Lan and D. Huttenlocher, "Beyond Trees: Common-Factor Models for 2D Human Pose Recovery", *In Proc. Of IEEE ICCV, pp. 470-77, 2005.*

[4] H. Schineiderman, "Learning a Restricted Baysian Network for Object Detection", *IEEE CVPR, vol. 2,pp. II-639-46, 2004.*

[5] D. Lowe, "Object Recognition From Local Scale-Invariant Features", *In Proc. ICCV, pp. 1150-57, 1999.*

[6] J. Sullivan and S. Carlsson, "Recognizing and Tracking Human Action," *Proc. European Conf. Computer Vision, 2002.*

[7] M. Hua and Y. Wu, "Learning to Estimate Human Pose with Data Driven Belief Propagation," *IEEE CVPR, vol. 2, pp. 747-54, 2005.*

[8] V. Vapnik, "The Nature of Statistical Learning Theory", *Springer Verlag, 1995.*

[9] A. Mohan, C. Papageorgiou, and T. Poggio, Member, IEEE, "Example-Based Object Detection in Images by Components, IEEE TPAMI", *vol. 23, No. 4, April 2001*