Human-area Segmentation by Selecting Similar Silhouette Images Based on Weak-classifier Response

Hiroaki Ando, Hironobu Fujiyoshi
Dept. of Computer Science, Chubu Univ. Aichi, 487-8501 Japan
ah@vision.cs.chubu.ac.jp, hf@cs.chubu.ac.jp

Abstract

Human-area segmentation is a major issue in video surveillance. Many existing methods estimate individual human areas from the foreground area obtained by background subtraction, but the effects of camera movement can make it difficult to obtain a background image. We have achieved human-area segmentation requiring no background image by using chamfer matching to match the results of human detection using Real AdaBoost with silhouette images. Although accuracy in chamfer matching drops as the number of templates increases, the proposed method enables segmentation accuracy to be improved by selecting silhouette images similar to the matching target beforehand based on response values from weak classifiers in Real AdaBoost.

1 Introduction

In recent years, technology for detecting people in video [2, 11] is important for knowing where humans are in the image. To accurately know what areas of the image are occupied by people, human area segmentation to allow body height estimation [6] and human identification between different cameras [4, 10] are important. Conventional human area segmentation techniques are mainly based on background subtraction. Zhao *et al.* achieved highly accurate human detection and tracking by applying human shape models and camera and scene models to foreground areas obtained by background subtraction [9, 8]. However, such methods require that a background image be acquired, which may be difficult to accomplish in the real world because of camera movement, etc. Leibe *et al.* take a part-based approach to human detection, using local features to achieve human area segmentation by applying human silhouette image and chamfer matching to the detection results [1]. That method requires chamfer matching with all of about 400 silhouette images that have been prepared in advance. A problem with chamfer matching, however, is that the matching accuracy becomes unstable as the number of templates (silhouette images) increases, so there is concern that chamfer matching with all of the silhouette images may lower segmentation accuracy.

We therefore propose an approach of pre-selecting silhouette images that are similar to the matching target, and then do chamfer matching on candidates from among them. The proposed method uses HOG local features and the Real AdaBoost statistical learning method to detect people. Next, the real AdaBoost response values of the detected areas and the silhouette images are compared to select the silhouette image that is most similar to the human in the detected area. Finally, chamfer matching is performed on the selected silhouette image and the human detection result area to accomplish the human area segmentation. By doing the chamfer matching with only the similar silhouette images, matching errors due to variation in human shapes can be suppressed and an improvement in matching accuracy can be expected.

2 Proposed method

The processing flow in the proposed method is illustrated in Fig. 1. The proposed method involves off-line processing to train the human detector and to obtain a hash table from the scores of silhouette images that correspond to the training samples, and on-line processing for detecting human areas in the input image by raster scanning and segmentation using chamfer matching. These processes are explained in detail in this section.

2.1 Off-line processing

2.1.1 Construction of a human detector

The proposed method begins by detecting candidate human areas for matching in the input image. The human detector uses Real AdaBoost [7] with HOG (Histograms of Oriented Gradients) [2] as the input feature. HOG features are based on the magnitude of edges in each edge direction within a local area. HOG features

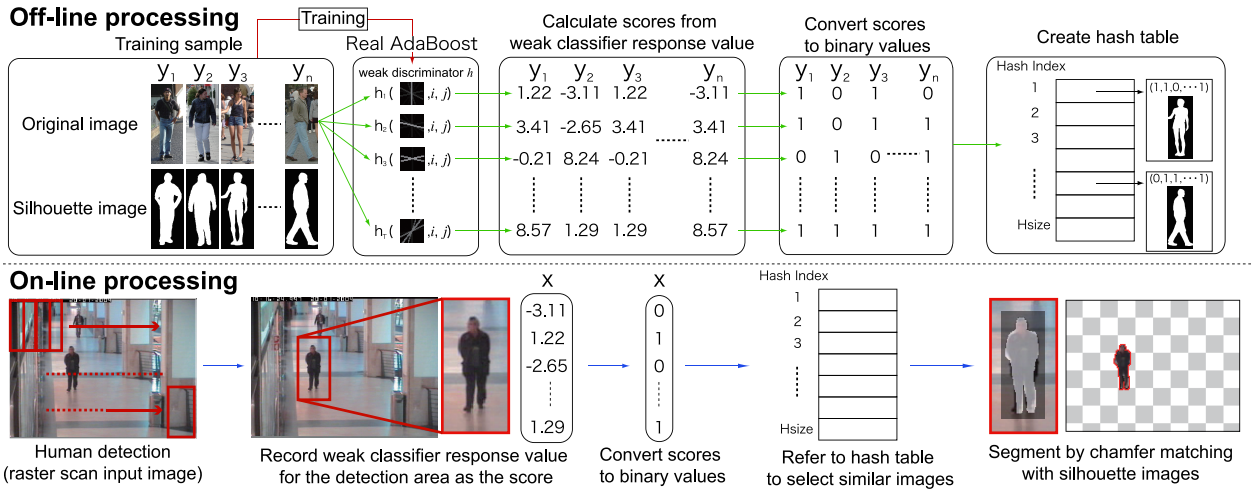

Figure 1. Processing flow in the proposed method

computed from the training samples that were prepared in advance are used to construct a human detector with Real AdaBoost. Real AdaBoost is a boosting algorithm in which the feature vectors calculated from the training samples that are the most effective for discrimination are learned as weak classifiers. The strong classifier, $H(x)$, is represented by a linear combination of multiple weak classifiers as defined in the following equation:

$$H(x) = \mathrm{sign}\left(\sum_{t=1}^{T} h_t(x)\right). \quad (1)$$

2.1.2 Calculation of silhouette image scores

Next, the constructed detector is used to compute the silhouette image score. The silhouette image score represents the response values of the weak classifiers of human detector as a multi-dimensional vector. First, we prepare silhouette images of training samples by manually extracted. Then, the training sample for that silhouette image (positive only) is input to the human detector. The weak classifiers for the input image return results indicating human or non-human. At that time, if the input image is correctly distinguished as human, a positive real number is returned as the weak classifier response value. If the input image is mistaken as non-human on the other hand, a negative real number is returned as the weak classifier response value. Each weak classifier holds one HOG feature vector and the position in detecting window. Accordingly, when a weak classifier response value is a positive real number, it means that the probability that one HOG feature vector obtained at the position selected by the weak classifier is human gradient information is high. Conversely, if the weak classifier response value is a negative real number, the HOG feature vector for that position has a high probability of being background. Finally, T-dimensional feature vector that corresponds to T weak classifiers is obtained. The T dimensional feature vector obtained here serves as the score for the corresponding silhouette image.

2.1.3 Creating the hash table

The calculated silhouette image scores are used to select silhouette images that are similar to the target areas for chamfer matching. In the proposed method, the silhouette image scores are stored in a hash table to allow fast search for similar silhouette images. Here, we use the hash table based nearest-neighbor search proposed by Noguchi *et al.* [5]. We therefore produce the hash table with silhouette image scores as feature vectors in off-line processing. We denote the score based on the response of weak classifiers from a given silhouette image y as the T dimensional feature vector $\mathbf{y} = \{h_1(y), h_2(y), \cdots, h_T(y)\}$. First, the T dimensional feature vector are converted to binary values with the following equation to create bit vector $\mathbf{u} = (u_1, u_2, \cdots, u_T)$.

$$u_j = \begin{cases} 1 & if \quad h_j(y) \geq 0 \\ 0 & otherwise \end{cases}. \quad (2)$$

Next, we use the following equation to obtain the hash table index H_{index} and register the feature vector and silhouette image ID in the hash table.

$$H_{index} = \left(\sum_{i=1}^{T} u_i 2^i\right) \bmod H_{size}. \quad (3)$$

Here, H_{size} is the size of the hash table. This processing is done for all of the silhouette images of positive sample to create the hash table. This is all off-line processing.

2.2 On-line processing

2.2.1 Human detection

The on-line processing involves detection of humans in an input image and matching the detection area with

silhouette images. First, a detection window is raster scanned across the input image from the upper left multiple times at different scales. When that is done, the area within the detection window is input to the human detector for detection and each area is discriminated as human or non-human. As the last step, the weak classifier response values is recorded as a T dimensional feature vector. At that time, the weak classifiers that judges the area as human return a positive real number and those that judge it as non-human return a negative real number. Similar silhouette images are selected by comparing the feature vector based on the weak classifier responses for the detected area and the feature vector based on the weak classifier responses for each silhouette image of training samples calculated in the offline processing.

2.2.2 Selection of similar silhouette images

Similar silhouette images are selected by comparing the feature vectors based on the weak classifier responses as described in section 2.1.2 and the feature vectors calculated for the detected window area as described in section 2.2.1. That comparison is accomplished by nearest-neighbor search using the hash table created in section 2.1.3. Denote the feature vector based on the weak classifier responses for the area detected in human detection as $\mathbf{x} = (h_1(x), h_2(x), \cdots, h_T(x))$. For this feature vector, too, we use Eq. (2) and Eq. (3) to obtain a hash index and then search the hash table for silhouette images that have the same hash index. The value for similarity between the feature vectors of the N silhouette images found and the detection area feature vector, s_i, is obtained with the following equation:

$$s_i = \frac{\mathbf{x} \cdot \mathbf{y}_i}{||\mathbf{x}||||\mathbf{y}_i||} \quad i = 1, \cdots, N. \tag{4}$$

The closer are the values of the feature vectors, the higher the probability that the corresponding areas have similar shapes. In this way, silhouette images that are similar to the detection area are selected. Examples of similar silhouette images selected automatically from the feature vectors of the human detection results area are shown in Fig. 2. From Fig. 2, we see that silhouette images that are similar to the detection area shape have been selected. Thus, preselection of silhouette images that are similar to the detection area can improve the accuracy of chamfer matching.

2.2.3 Segmentation by chamfer matching

Human area segmentation is accomplished by using chamfer matching to match the detected human areas and the similar silhouette images selected by the weak classifier response values. The matching is based on the dissimilarity measure. Since chamfer matching is sensitive to changes in scale, the scale of the silhouette image is varied in the matching process. In the proposed method, the chamfer matching is performed with the size of the human detection area varied by a magnification factor of from 0.8 to 1.2 in units of 0.01. In the final step, the position, scale, and silhouette image for which the chamfer matching dissimilarity is the smallest is determined, and human area segmentation is performed by extracting the part of the image that overlaps with the silhouette image.

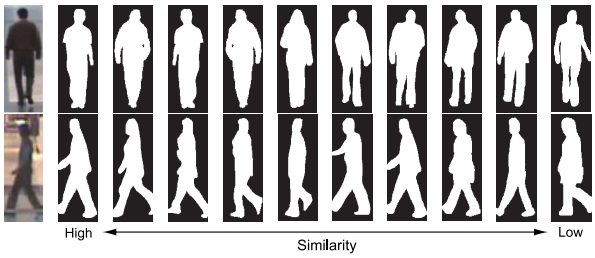

Figure 2. Examples of selected similar silhouette images

3 Experimental Results

3.1 Experiment

To confirm the effectiveness of the proposed method, we conducted an experiment to evaluate segmentation accuracy. To evaluate the effectiveness of pre-selecting silhouette images that are similar to the matching targets based on the weak classifier response values used in the proposed method, we compared the two methods described below:

Conventional method Matching with all of silhouette images

Proposed method Matching of similar silhouette images pre-selected on the basis of weak classifier response values

In the proposed method, ten similar silhouette images are selected for chamfer matching. The two sequences used in the evaluation are data sets from the CAVIAR Test Case Scenarios [3]. The sequences include 300 frames of a scene that involves movement of humans toward the camera (front direction) and 300 frames of a scene that involves human movement perpendicular to the camera (horizontal direction) In the scene that has movement toward the front, there is a large change in the scale of the human images; in the horizontal movement scene, there is a large change in the shape of the human images. The evaluation method involves computing recall, precision and F-measure values for each frame and then taking the mean value for the entire sequence. Correct areas for all of the evaluation sequence frames were produced manually. Then, 800 typical shapes were selected from the human detection training samples to serve as a silhouette image sample.

Figure 3. Example segmentation results (a), (b), and (c): front direction; (d), (e), and (f): horizontal direction

Table 1. Comparison of segmentation accuracy

		Conventional	Proposed
front	Recall	0.79	0.81
	Precision	0.82	0.90
	F-measure	0.80	0.85
horizontal	Recall	0.72	0.73
	Precision	0.71	0.76
	F-measure	0.71	0.75

3.2 Experimental results

The segmentation accuracy evaluation results for the two sequences are presented in Table 1. Examples of the segmentation results are shown in Fig. 3. The results for the proposed method (with silhouette selection) are from using 10 selected silhouette. The results from the conventional method (without silhouette selection) are from matching with all 800 silhouettes. The results confirm increases in accuracy for the front sequence of 0.02 for recall and 0.08 for precision. An accuracy increase for the F-measure of 0.05 is also confirmed. For the horizontal sequence, accuracy improvements of 0.01 for recall and 0.05 for precision were confirmed. For the F-measure, an accuracy improvement of 0.04 was confirmed. These results show that silhouette image selection can reduce false positives. Furthermore, the mean processing time for segmentation per person with silhouette selection is 2.1 s compared to 352.1 s without silhouette selection, achieving a speed increase by a factor of 167. From the example of segmentation results, we see that relatively good accuracy segmentation is possible for the front sequence in which large scale changes occur. We can also see that segmentation with good accuracy is possible for the horizontal sequence, in which there is more change in human shape than in the front sequence. The above results confirm the effectiveness of the proposed method.

4 Conclusion

We proposed a method of human area segmentation that uses chamfer matching based on preselection of similar silhouette images according to weak classifier responses. The proposed method reduces chamfer matching errors by preselecting silhouette images that are similar to the matching target, thus improving matching accuracy. In future work, we plan to achieve higher accuracy in area segmentation of the target for detection from segmentation results that match the silhouette image shape.

References

[1] E. S. B. Leibe and B. Schiele. Pedestrian detection in crowded scenes, 2005. Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection, 2005. IEEE Computer Vision and Pattern Recognition.

[3] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.

[4] K. Kamimura, Y. Ikegame, K. Shimoyama, T. Tamaki, and M. Yamamoto. Real time person collation system using multiple cameras on the network, 2004. Technical report of IEICE.

[5] K. Noguchiand, T. Nakai, K. Kise, and M. Iwamura. Experimental investigation of relation between near neighbor search methods for feature vectors and efficiency of object recognition, 2006. Technical report of IEICE.

[6] Y. Ogawa and H. Fujiyoshi. Visual tracking system by master-slaving, 2003. The ninth Symposium on Sensing via Image Information (SSII 03).

[7] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions, 1999. Machine Learning.

[8] R. N. T. Zhao. Bayesian human segmentation in crowded situations, 2003. IEEE Computer Vision and Pattern Recognition.

[9] R. N. T. Zhao and F. Lv. Segmentation and tracking of multiple humans in complex situations, 2001. IEEE Computer Vision and Pattern Recognition.

[10] H. Takahashi, K. Nakamura, H. Zhao, and R. Shibasaki. Human identification using laser scanners and image sensors, 2007. Asian Conference on Remote Sensing.

[11] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors, 2007. International Journal of Computer Vision.