Pedestrian Attribute Analysis Using a Top-View Camera in a Public Space

Toshihiko Yamasaki[1,2,3] and Tomoaki Matsunami[1]

[1] Department of Information and Communication Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[2] School of Electrical and Computer Engineering, Cornell University, 116 Ward Hall, Ithaca, NY 14853, USA
[3] JSPS Postdoctoral Fellow for Research Abroad
{yamasaki,matsunami}@hal.t.u-tokyo.ac.jp

Abstract. In this paper, we propose a method to analyze gender of the pedestrian and whether he or she has a baggage or not in a public space. The challenging part of this work is we only use top-view camera images to protect the pedestrians' privacy. We focused on temporal changes in their position, shape, and contours over the frames because their appearances do not provide much information. We extracted the pedestrians' features using their position, area, aspect ratio, histogram of oriented gradients (HoG), and Fourier descriptors. The temporal information was taken into consideration by employing Gaussian mixture models (GMM), GMM universal background model (GMM-UBM), and bag of features (BoF) model. The attributes were classified by using support vector machines (SVM). We conducted experiments using 60-minute video captured by a top-view camera attached at an airport. Experimental results show that the classification accuracy is 69% for the gender classification and 79% for baggage possession classification.

Keywords: Human attributes, surveillance, gender classification, bag possession classification.

1 Introduction

Visual surveillance has been one of the most active research areas in computer vision [1][2]. Surveillance cameras have been installed in a lot of places in such as stations, airports, or on the streets for security purposes. Visual surveillance data are easy to analyze for humans. On the other hand, analyzing the data by computers requires a wide range of algorithms such as moving object detection, object classification, counting, tracking, behavior labeling, human identification, abnormal object/event detection, flux analysis, data fusion collected from multiple cameras, and so on.

Understanding human attribute and behavior, in particular, is getting more attention not only for security reasons but for better services, marketing, and so on. If surveillance systems can recognize gender and age range of the passengers, digital

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 541–550, 2012.

singnage dedicatedly designed for a particular target can be displayed. If systems detect children who are alone, they might be lost and looking for their parents. In addition, systems can alert person who is carrying a large suitcases widely spread behind him/her, which is dangerous and is becoming a significant safety issue in crowded airports and stations. For activity recognition, Chen and Hauptmann proposed MoSIFT [3]. MoSIFT was an extension of the Scale Invariant Feature Transform (SIFT) [4] features to the temporal domain and showed its superiority to Histogram of Oriented Gradients (HoG) [5] and Histogram of oriented Optical Flow (HoF) [6] based approaches. Vezzani et al. proposed projection histogram features and used Hidden Markov Models (HMM) to classify the human activities [7]. These activity recognition algorithms focused on general activity classification such as walking, running, jumping, and so forth. Ozturk et al. [8] proposed body and head orientation detection to analyze what pedestrians are looking at in a public space. They used omega-like-shape detection for head pose estimation and SIFT-based feature tracking for temporal analysis.

In this paper, we analyze gender of pedestrians in an airport and to judge whether they have bags by using only top-view images. A lot of work on bag detection [9], gender classification [10], and face attribute analysis [11][12] can be found in literatures. Tao et al. proposed general tensor discriminant analysis using Gabor filter based gait analysis to analyze human carrying status [9]. Zhang et al. analyzed the optimal camera angle for the gender classification using SMV classifiers [10], in which only yaw angles were considered. In these approaches, however, the quality of the images was well-controlled: target objects were large enough, taken from the frontal-view, and so on. On the other hand, only top-view images taken by a surveillance camera is used in this work, which could protect the pedestrians' privacy. Another challenging point is the data used in our experiment were "real-life" data taken at an airport, not simulated data.

In our proposed approach, we detect pedestrians and track them using conventional background subtraction and blob tracking. Then, position, area, contour, shape, and their changes are extracted. In order to analyze the shape feature of the contour images, three kinds of Fourier descriptors [13] and HoG are employed. Features over the multiple frames are considered by using GMM [14], GMM-UBM [15], and BoF [16]. The validity of the proposed algorithm was evaluated using 60 minutes' real-life video taken at Haneda airport, in which 788 pedestrians walked through the view area. The experimental results have demonstrated that the performance for the gender classification was 69% when the change in the position, area, and aspect ratio is directly considered. On the other hand, the P-type Fourier descriptor with GMM-UBM yielded the best score of 79% for baggage possession classification followed by the P-type descriptor with BoF a little behind.

The organization of this paper is as follows. Section 2 describes pedestrian detection and tracking. Feature extraction from each frame and feature vector generation over the frames are explained in Section 3 and Section 4, respectively. Experimental results are demonstrated in Section 5. Concluding remarks are given in Section 6.

2 Detection and Tracking

Since multiple pedestrians could be observed in a frame, detection and tracking of them is mandatory. A conventional approach was employed because detection and tracking themselves are out of scope of this paper. A background image is generated for every frame by averaging the previous 60 frames. After simple background subtraction, graph-cuts [17] based segmentation is applied to extract the pedestrians' silhouettes. The graph-cuts based segmentation is important because some pedestrians wear a white or cream color shirt, whose color is very close to that of the floor and simple background subtraction cannot detect the silhouette properly. The flowchart and some results are shown in Fig. 1.

Since the pedestrians' paths in the view area are rather simple, a simple blob tracking algorithm is employed to save computational time. Once blobs representing pedestrians are detected, blob matching is done between neighboring frames searching for the nearest blob in terms of the position and the size. If there is no correspondence in the previous frame, the blob is detected as a new pedestrian and the same procedure is applied to the pedestrians who are getting out of the view area. In this paper, erroneously detected or tracked blobs were eliminated in advance. And only the pedestrians who existed in the area for more than 20 frames are analyzed because the temporal change is also considered. Some examples of our blob tracking results are demonstrated in Fig. 2.

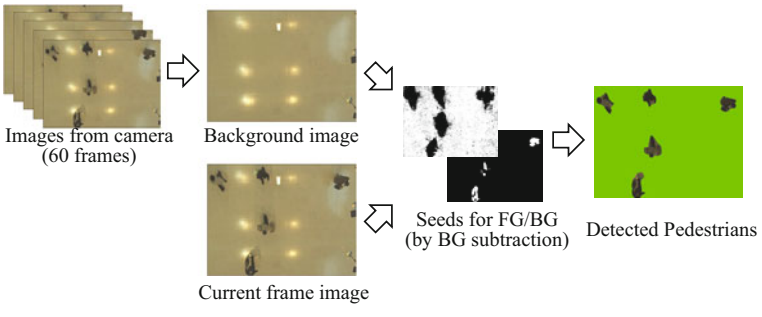


Fig. 1. Pedestrian detection using background subtraction and graph-cuts

Fig. 2. Blob tracking by comparing position and size

3 Feature Extraction from Each Frame

Analyzing human attributes using only top-view images is a challenging task because no face and no details are recorded while it protects the pedestrians' privacy. The underlying assumption is that their silhouette and how they walk would differ depending on their gender and their belongings.

3.1 Position and Area Based Features

The change in pedestrians' position is calculated as follows. We assume that they walk straight in a short distance so the direction of the pedestrian is estimated by the least square linear fitting using five previous positions of the pedestrian. Then, the shift from the previous position in the perpendicular and parallel directions to the estimated moving direction is defined as dx(t) and dy(t), where t is the frame ID. In addition, area, aspect ratio, and their changes from the previous frame are also used.

3.2 Shape Based Features

The shape feature of the detected pedestrians is analyzed by three kinds of Fourier descriptors: G-type, P-type, and Z-type. Approximating the contour by a closed loop of lines is common to all of them but to what components the Fourier transform is applied is different. In our case, the contour is represented by 100 lines. In the G-type Fourier descriptor, the Fourier transform is applied to the vertex position in a form of $z(i) = x(i) + jy(i)$, ($i = 0, \ldots, 99$) directly:

$$e(k) = \frac{1}{n}\sum_{i=0}^{n-1} z(i)\exp\left(-2\pi j\frac{ik}{n}\right), (\mathrm{k} = 0,\ldots,99)\,. \tag{1}$$

Here, $e(k)$ is the G-type descriptor and i is the ID for the lines. The P-type descriptor is obtained applying the Fourier transform to the length-normalized vector from $z(i)$ to $z(i+1)$ as shown below:

$$c(k) = \frac{1}{n}\sum_{i=0}^{n-1} w(i)\exp\left(-2\pi j\frac{ik}{n}\right), (\mathrm{k} = 0,\ldots,99) \tag{2}$$

where $w(i) = (z(i+1)-z(i))/|z(i+1)-z(i)|$.

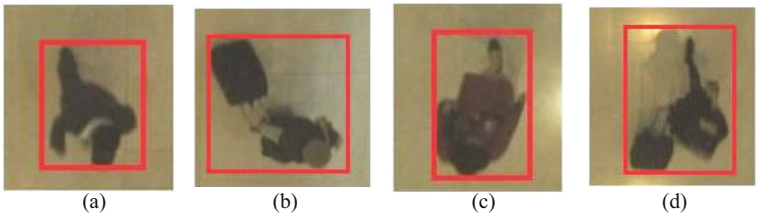
The Z-type is the Fourier coefficients of the angles of the lines instead of using the vectors as in the P-type descriptor.

$$d(k) = \frac{1}{n}\sum_{i=0}^{n-1} \varphi(i)\exp\left(-2\pi j\frac{ik}{n}\right), (\mathrm{k} = 0,\ldots,99) \tag{3}$$

where $\varphi(i)= \theta(i)-\theta(0)-2\pi l/L$.

Table 1. Summary of pedestrians' attributes

	With bag	W/o bag	Total
Male	272	187	459
Female	179	150	329
Total	451	337	788



(a) (b) (c) (d)

Fig. 3. Examples of pedestrians: (a) male w/o bag, (b) male w/ bag, (c) female w/o bag, (d) female w/ bag

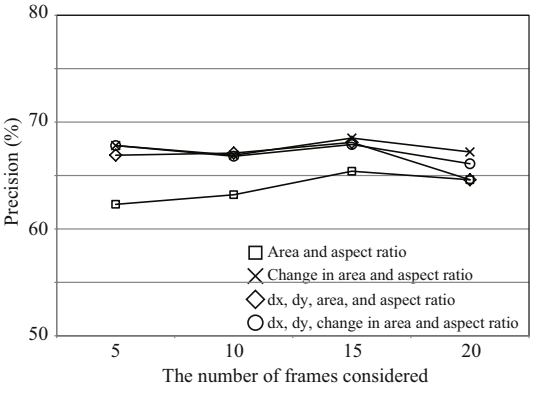
Here, $\theta(i)$ is the angle of the ith line to the x-axis, l is the length of the lines accumulated from the 0th to the ith, and L is the total length of all the lines. The lengths of the lines need be stored for the inverse transform in the P-type and the Z-type descriptors.

Only a limited number of Fourier coefficients from lower frequency components, which describes rough shape of the object, are used for the classification. The performance dependency on the number of coefficients will be discussed in Section 5. HoG-based feature vectors are generated as in the original paper [5].

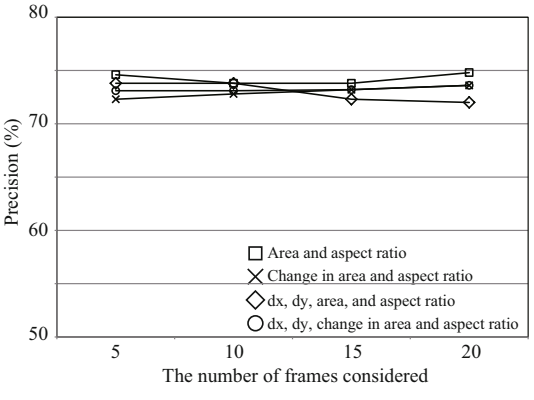
4 Feature Extraction over the Frames

For the position and area related feature, the dimension of the feature description is small. Therefore, they are simply concatenated chronologically. The other features such as Fourier descriptors and HoG descriptors tend to have higher dimension. Therefore, the feature vector distribution over the frames is transformed into a single vector by applying GMM, GMM-UBM, and BoF.

In the GMM approach, a mixture of Gaussians in the feature space is estimated using the expectation-maximization (EM) algorithm and the mean vectors of the estimated Gaussians are concatenated to form a feature vector. In a simple GMM approach, GMM is generated independent of other pedestrians' set of vectors. Therefore, the order of concatenating multiple mean vectors is not consistent among the pedestrians. In the GMM-UBM approach, however, the seed vectors for the EM process are generated by putting all the feature vectors of all the pedestrians first and then the seeds are used in generating a GMM for each pedestrian. Therefore, generated feature vectors are expected to be more robust than those generated by a simple GMM approach. The BoF vectors are generated by clustering all the feature vectors of all the pedestrians and generating a frequency histogram of the cluster IDs for each pedestrian. The BoF vectors are normalized by the number of frames.



(a)

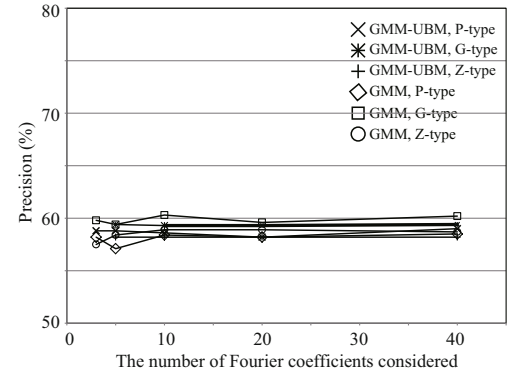


(b)

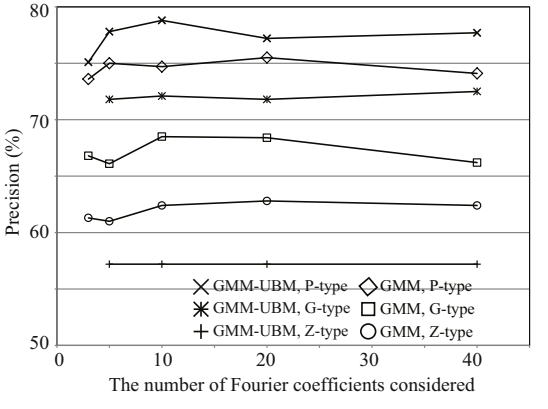
Fig. 4. Classification performance using position, area, and aspect ratio: (a) gender, (b) with/without bag

5 Experimental Results

The data were captured in Haneda airport, Japan, which is one of the top-5 busiest airports in the world. A top-view camera was attached at 12m height. The view area was about 6m x 4.5m. The images were captured at 720 x 540 resolutions and at 6.25 frames per second because the system was originally designed to record images for a long period. Typical pedestrians' sizes are about 100 x 100. Note here is that all the pedestrians were actual travelers; there were no "simulated (or pretended)" pedestrians. Only the pedestrians who were detected for more than 20 frames were used. Erroneously detected/tracked blobs such as those including two or more pedestrians in them were eliminated by hand. Such miss detection and tracking and occlusion/overlap problems are still difficult problems [18] and therefore they are out of focus of this paper. This paper concentrates only on human attribute analysis



(a)



(b)

Fig. 5. Classification performance using Fourier descriptors with GMM and GMM-UBM: (a) gender, (b) with/without bag

assuming that such pre-processing is done perfectly. Ground truth was annotated by the authors. The number of detected pedestrians and their attributes after the pre-processing are summarized in Table 1 and some sample images are shown in Fig. 3. The extracted feature vectors were classified using SVM with the Gaussian kernels optimized for each case. The accuracy was calculated by the 10-cross validation.

Figure 4 shows the classification performance using position, area, aspect ratio and their temporal changes. The changes between frames are more significant for gender classification. The best performance is obtained when the changes in the area and aspect ratio over 15 frames are used and its accuracy is 69%. On the other hand, raw data of area and aspect ratio are better than the others for the with/without bag classification. It can be observed that the number of frames is not so important except for only a few exceptions. Also, it is interesting to see that the pedestrian's gender and bag possession status affects how they walk to some extent and it can be observed in such simple features.

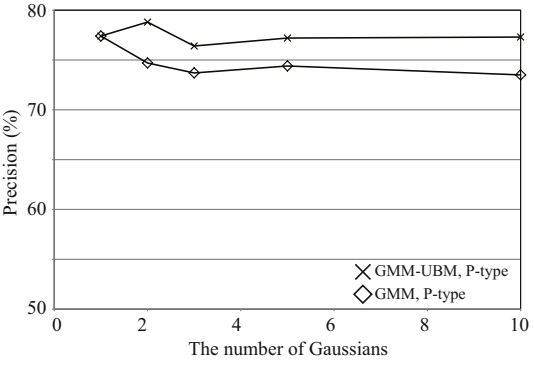


Fig. 6. Performance of with/without bag classification using GMM and GMM-UBM based features as a function of the number of Gaussians

The classification performance using Fourier descriptors with GMM or GMM-UBM is demonstrated in Fig. 5. The number of Fourier coefficients is altered in the x-axis. The number of the Gaussians is set as two. We can see that the gender classification accuracy is around 58-60% for all the cases. Besides, the gender classification accuracy was always within the range of 58-60% for all the experiments hereafter. Therefore, graphs are not shown to save the limited space. For with/without bag classification, the P-type Fourier descriptor performs the best and GMM-UBM yields better results than simple GMM. This tendency coincides with [19], which compared the P-type and Z-type Fourier descriptors and Zernike moments in the context of motion retrieval. The best performance of 79% is obtained when the number of coefficients is 10 for the P-type descriptor with GMM-UBM. 5-20 coefficients out of 100 are enough for the classification, showing that the other coefficients do not contribute to shape description and can be regarded as noise.

The classification performance as a function of the number of Gaussians is shown in Fig. 6. The GMM model performs the best when the number of Gaussians is only one. On the other hand, for the GMM-UBM model, the performance gets the maximum when the number of Gaussian is two. The computational cost and the memory usage for the feature vector storage are almost the same for GMM and GMM-UBM. GMM-UBM is better from the view point of the classification performance.

The BoF representation works well with the P-type Fourier descriptor as shown in Fig. 7. The accuracy is the best (78%) when the number of clusters is set at 200. Since the number of frames for each pedestrian is only 20-40 frames, the generated BoF vector is very sparse. On the other hand, BoF using HoG features performs with less than 75% of accuracy. In addition, gender classification accuracy using HoG was 58% for GMM and 59% for GMM-UBM and that for baggage possession classification was 71% for GMM and 69% for GMM-UBM.

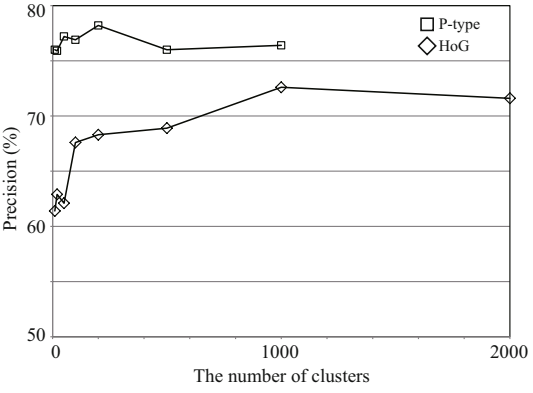


Fig. 7. Performance of BoF-based with/without bag classification using p-type Fourier descriptor and HoG

6 Conclusions

In this paper, we have presented the algorithms to analyze the pedestrians' attributes such as gender and whether they have bags or not using top-view images in the airport. After the pedestrian detection and blob tracking, the features for each frame were extracted such as temporal change in position and area, shape feature using HoG and Fourier descriptors. Then GMM, GMM-UBM, and BoF were applied to the feature vectors over the frames to form final feature vectors for the classification. The experiments using 60 minutes' video demonstrated that the gender could be classified with 69% of accuracy. And the accuracy for the with/without bag classification was 79%. It has been shown that simple features such as temporal change in position and area performs well for the gender classification and the P-type Fourier descriptor with either GMM-UMB or BoF is suitable to judge the pedestrian possesses a bag or not. Performance improvement up to 96% is expected by employing a multi-stage classifier framework along with a HoG-based BoF model [20].

The future direction of this work is analyzing more attributes such as age range and group/family detection, which would be moving synchronously, among multiple blobs.

Acknowledgments. The authors would like to thank Prof. K. Aizawa of the University of Tokyo for valuable discussions and comments.

References

1. Hu, W., Tan, T., Wang, L., Mayban, S.: A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34(3), 334–352 (2004)
2. Candamo, J., Shreve, M., Goldgof, D.B., Sapper, D.B., Kasturi, R.: Understanding transit scenes: a survey on human behavior-recognition algorithms. IEEE Transactions on Intelligent Transportation Systems 11(1), 206–224 (2010)

3. Chen, M.Y., Hauptmann, A.: MoSIFT: recognizing human actions in surveillance videos date of original version. CMU Technical Report (September 2009)
4. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) 60(2), 91–110 (2004)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE CVPR, pp. 886–893 (2005)
6. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proc. IEEE CVPR, pp. 1932–1939 (2009)
7. Vezzani, R., Baltieri, D., Cucchiara, R.: HMM Based Action Recognition with Projection Histogram Features. In: Ünay, D., Çataltepe, Z., Aksoy, S. (eds.) ICPR 2010. LNCS, vol. 6388, pp. 286–293. Springer, Heidelberg (2010)
8. Ozturk, O., Yamasaki, T., Aizawa, K.: Estimating Human Body and Head Orientation Change to Detect Visual Attention Direction. In: Koch, R., Huang, F. (eds.) ACCV Workshops 2010, Part I. LNCS, vol. 6468, pp. 410–419. Springer, Heidelberg (2011)
9. Tao, D., Li, X., Maybank, S.J., Wu, X.: Human Carrying Status in Visual Surveillance. In: IEEE CVPR, vol. 2, pp. 1670–1677 (2006)
10. Zhang, D., Wang, Y.: Investigating the separability of features from different views for gait based gender classification. In: Proc. ICPR, pp. 1–4 (2008)
11. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE ICCV, pp. 365–372 (2009)
12. Guo, G., Mu, G.: A study of large-scale ethnicity estimation with gender and age variations. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 79–86 (2010)
13. Uesaka, Y.: Spectral analysis of form based on Fourier descriptors. In: Proc. the First International Symposium for Science on Form, pp. 405–412 (1986)
14. Goldberg, M., Shlien, S.: A clustering scheme for multispectral images. IEEE Transactions on Systems, Man and Cybernetics 8(2), 86–92 (1978)
15. Campbell, W.M., Sturim, D.E., Reynold, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13(5), 308–311 (2006)
16. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proc. IEEE ICCV, pp. 1470–1477 (2003)
17. Boykov, Y., Jolly, M.-P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proc. IEEE ICCV, pp. I-105–I-112 (2001)
18. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. 38(4) Article 13 (December 2006)
19. Kasai, D., Yamasaki, T., Aizawa, K.: Retrieval of Time-Varying Mesh and Motion Capture Data Using 2D Video Queries Based on Silhouette Shape Descriptors. In: Proc. IEEE ICME, pp. 854–857 (2009)
20. Yamasaki, T., Matsunami, T.: Human Attribute Analysis using a Top-View Camera Based on Multi-Stage Classification. In: Proc. 5th ACM/IEEE ICDSC, #61 (2011)