Relative-Distance-Based Soft Voting for Human Attribute Analysis using Top-View Images

Toshihiko Yamasaki (member),†,††, Tsuhan Chen ††

Abstract  This paper proposes a soft voting based bag-of-features (BoF) model considering relative distance of the feature vectors to the nearest-neighbor codeword. The proposed method is more efficient than the kernel distance based soft voting method, which requires brute force parameter optimization. The proposed algorithm is applied to human attribute analysis using top-view images and conventional object classification. The experimental results for the human attribute analysis have demonstrated 100% accuracy for both gender classification and bag possession status classification. It has also been demonstrated that discriminative ability is comparable to that of the fine-tuned kernel distance based soft voting method.

Key words: bag of features, soft voting, human attribute analysis.

1. Introduction

Visual surveillance has been one of the most important areas in computer vision and pattern recognition and a lot of work is summarized in survey papers1)2). A wide range of algorithms are involved in visual surveillance: moving object detection, object classification, counting, tracking, behavior labeling, human identification, abnormal object/event detection, flux analysis, data fusion among multiple cameras, and so on.

Understanding human attribute and behavior, in particular, is getting more attention not only for security and safety purposes but also for better quality of life services. If surveillance systems can recognize gender and age range of the pedestrians, digital signage 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. On the other hand, when systems detect elderly people walking unsupported, the system can automatically arrange pick-up carts for them. In addition, systems can alert person who is dragging a large suitcases widely spread behind him/her, which is dangerous and is becoming a significant safety issue in crowded airports and stations.

In this paper, therefore, we propose a human attribute analysis system that can classify gender of the pedestrians and their bag possession status. The data used in this paper are more challenging than previous works because they were recorded by top-view cameras: pedestrians’ faces and cloths are not clear and the size of pedestrians are relatively small. In the feature generation stage, a relative distance based soft voting method is proposed. It is more efficient than the previous kernel based soft voting method3) because it is parameter optimization free.

The rest of this paper is organized as follows. Related works are summarized in Section 2. Section 3 describes the proposed algorithm. The process flow of the human attribute analysis is described in Section 4. The experimental results for the human attribute analysis is presented in Section 5 followed by concluding remarks in Section 6. The results for the object classification are demonstrated in Appendix A.

2. Related Works

2.1 Human Attribute Analysis

There are a lot of works on surveillance systems for security and safety. On the other hand, high-order understanding of pedestrians is getting more attention these days.

Chen and Hauptmann proposed MoSIFT for activity analysis9). Ozturk et al. proposed body and head orientation detection to analyze what pedestrians are looking at13). Inoue et al. proposed a combination of global image feature and local image features for classifying high-level human activity recognition9).

We have also been developing a human attribute analysis system to classify the gender of the pedestrians
and judge whether they have baggage or not\(^7\(^8\)). The difference from previous works\(^9\)\(19\) is that we only use top view cameras. It is more challenging because the details of the pedestrians such as faces and kinematic parameters are difficult to analyze. It should also be pointed out that most of the previous works are based on some assumptions such as the size of the human is large enough, the human is shot from the side-view so that his/her clothing and appearance is clear, and the environment is well controlled (e.g., lab room, chroma key background). On the other hand, our data were recorded in an airport, where all the pedestrians were ordinary people: no students or those pretending to be ordinary pedestrians were included.

In our previous work\(^7\), for the gender classification, feature vector representation using position, area, and aspect ratio of the pedestrians’ blobs and their temporal changes were considered. For bag possession status classification, on the other hand, p/g/z-type Fourier descriptors\(^20\) with Gaussian mixture model (GMM)\(^21\) and GMM universal background model (GMM-UBM)\(^22\) have been employed. The classification accuracies for the gender and with/without bag were 68.5\% and 78.8\%, respectively. Then, the performances were improved up to about 95\%\(^8\) by the optimized histogram of oriented gradients (HoG)\(^23\).

### 2.2 Feature Representation

A bag-of-features (BoF) model\(^24\) is one of the most successful approaches for efficient and effective feature representation for image/video classification and retrieval. The basic idea of the BoF is to represent the input data as a histogram of code indices of the features regardless of their spatial and temporal orders. In order to quantize the feature vectors and to assign index values to them, k-means clustering is commonly used. The representative vectors (codewords) obtained in the clustering are called “visual words” in analogous to the bag-of-words model\(^35\), which is used in text retrieval. Namely, a collection of features extracted from the input data is clustered to assign an index value to each feature vector and a histogram of such indices is formed as a new feature vector for further processing. For instance, in Ref.\(^24\) which first introduced the concept of BoF in image retrieval, local feature descriptors using scale invariant feature transform (SIFT)\(^26\) were used to form a histogram for each video frame. The same concept can also be extended to the temporal domain\(^27\)(28).

After the great success of the BoF model, a number of techniques to improve the performance have been proposed\(^29\)\(^31\). Philbin et al. investigated the optimal number of clusters\(^29\). In Ref.\(^30\), the spatial distribution of the local feature descriptors were taken into consideration by dividing the input images into sub-regions in a pyramid manner such as 1\(\times\)1, 2\(\times\)2, 4\(\times\)4. A space-time pyramid with adaptive multiple kernel learning was also proposed for robust video event recognition\(^31\).

Another significant progress in generating the feature vector was soft voting. In Ref.\(^24\), for instance, only a single codeword was assigned to each feature vector. Recent studies have shown that soft voting that assign two or more codewords with weights contributes to improving the image/video classification and retrieval performance\(^31\)(32). In Ref.\(^32\), the weight assigned to each feature vector was an exponential function of the distance to the codewords. In the codeword uncertainty (UNC) model\(^3\), although the weights were also based on the exponential function of the distance as in Ref.\(^32\), the weights were further normalized by the sum of the distances to all the codewords. In Ref.\(^33\), the graphical structure among codewords was investigated. Ref.\(^34\) tried to assign a few soft codewords to each feature vector using the sparse coding instead of using soft voting. Then, Ref.\(^35\) proposed a locality-constrained linear coding (LLC) that enhanced the accuracy and the computational efficiency as compared to the sparse coding based method\(^34\).

Although the soft voting method yields better performance than the conventional hard voting method, an extra parameter optimization for the weight value calculation is needed\(^33\)(32) in addition to the codebook size optimization. This paper proposes a soft voting method which considers relative distance. The distance to each codeword normalized by the distance to the nearest neighbor is used as a weight value for the soft voting. The proposed algorithm has been applied to human attribute analysis using surveillance video data, which has been a difficult problem in previous literatures\(^7\)(8). The experimental results demonstrated that the performance for both gender classification and bag possession status classification was 100\% and its performance and robustness is comparable to that of the UCN model. The efficiency in terms of the parameter optimization was also investigated.

### 2.3 Contribution of this paper

The contributions of this paper are as follows. One is soft-voting-based feature representation considering the relative distance between the input and the codewords.
and the other is accurate human attribute analysis. The algorithm was presented in Ref.36). In this paper, we have conducted intensive study to demonstrate the performance of our algorithm. In addition, the proposed algorithm has also been applied to a conventional object recognition problem, showing the sub-optimality of the proposed soft voting method.

As summarized in Section 2.1, there were many works that relied on “well-controlled” experimental environments. On the other hand, the data in our experiments are “real-life” data that were taken in an airport. In addition, pedestrians’ faces and cloths are not visible very well and pedestrians are very small. Nevertheless, our proposed work can classify pedestrians’ attributes very accurately.

3. Proposed Algorithm

In the hard voting approaches\(^{24}\), a single codeword is assigned to each input feature vector. Then, a histogram of word frequencies that describes the probability density over the codewords is calculated as in the following equation.

\[
CB(w) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 
1 & (w = \arg \min_{v \in V} D(v, r_i)) \\
0 & \text{(otherwise)}
\end{cases}
\]  

Here, \(n\) is the number of feature vectors, \(r_i\) is ith feature vector, and \(D(v, r_i)\) is the distance between a codeword \(v\) and \(r_i\). \(V\) is the codebook. L2 norm is employed for the distance metric in this paper. The UNC\(^{3}\) is a soft voting model which considers “relevancy” determined by the ratio of the kernel values for all codewords \(v\) in the codebook \(V\):

\[
UNC(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_\sigma(D(w, r_i))}{\sum_{j=1}^{\|V\|} K_\sigma(D(v_j, r_i))}
\]  

where \(K_\sigma\) is the Gaussian-like kernel defined as

\[
K_\sigma(x) = \exp(-\beta x^2).
\]  

The distance distribution depends on the feature representation algorithm and the dataset. Therefore, \(\beta\) value can be decided only by try-and-error. In addition, the optimal \(\beta\) would also change depending on the codebook size. There is also an overfitting problem to be considered. As a result, obtaining the optimal value of \(\beta\) is a difficult problem.

The proposed method is soft voting which considers the relative distance to the nearest codeword:

\[
REL(w) = \frac{1}{n} \sum_{i=1}^{n} \min_{v \in V} \frac{D(v, r_i)}{D(w, r_i)}
\]  

While the UNC model considers the distribution of the distances from the input feature vector to all the codewords, the proposed model only cares the relative distance of all the feature vectors to the closest codeword. Therefore, the weights would not become too large or too small, which can happen in the UNC case when \(\beta\) is not properly set. In addition, since there is no magic number in eq. 4, one has to optimize only the codebook size. The feature vectors are normalized before being fed to machine learning algorithms.

The conceptual difference of the proposed algorithm with the conventional methods is illustrated in Fig. 1. Here, let us assume that we have the codebook consisting of \(a \sim j\) and three input vectors such as a rectangle, a triangle, and a diamond as shown in Fig. 1(a). The dark clouds around the codewords correspond to
the Gaussian function defined in eq. 3. The darker the color is, the larger the value is. As demonstrated in Fig.1(b), if $\beta$ in the UNC model is too large, it is nothing more than conventional BoF. Only the nearest neighbors are considered. On the other hand, if $\beta$ is too small, the generated feature vector is flat and there is no discrimination power. On the other hand, the feature vector generated by our proposed algorithm is close to the optimal case in the UNC algorithm.

4. Process Flow of the Human Attribute Analysis

In this paper, the proposed soft voting method is applied to the bag-of-frames representation of the walking sequence in an airport to classify the gender and bag possession status. Such human attribute analysis is useful not only for the safety and security but also for the digital signage purposes. If we can know the gender, the advertisement dedicatedly designed for the target can be displayed (e.g., ads of whiskeys or ties for male and ads of cosmetics or bags for female). And if we can know the bag possession status, we could know whether the pedestrian in the station is traveling for business or for leisure. Such human attribute analysis has been a difficult problem so far.

As the first step, pedestrians are detected using the background subtraction and graph-cuts. A flowchart is shown in Fig. 2. A background image is generated for each frame by averaging the previous 60 frames. Highly confident background and foreground regions are extracted by the background subtraction. The pedestrians are detected and extracted by using the graph-cuts using the background/foreground seeds. Since the experiments are done in an airport in our experiment, we can assume that the light condition is stable.

Pedestrian tracking is done by blob tracking, which searches for the nearest blob in terms of distance and size difference between the neighboring frames as shown in Fig. 3. If extra number of blobs are detected in the next frame, the blobs that cannot find their corresponding blobs are regarded as new pedestrians. In the same manner, the disappearing pedestrians can also be identified.

After the tracking, a sequence of detected blobs for each pedestrian is obtained as shown in Fig. 4. The feature vector for each frame is calculated using the HoG algorithm and the feature vector for the sequence is calculated by the bag-of-frames with the proposed soft voting method. The codebook is trained using all the feature vectors (for the sequences) by k-means clustering. The human attribute is classified by a support vector machine (SVM). Our previous works demonstrated that the classification performance was 69% and 76% respectively by the p-type Fourier descriptor with Gaussian mixture model (GMM), and 94% and 96% respectively by the hard voting using HoG features. In the previous works, the performance was also sensitive to the codebook size (please see Section 5.2), therefore the feasibility of the algorithms was still questionable.

5. Experimental Results

5.1 Experimental Setup

The experimental data were captured at Haneda airport (Tokyo International Airport), which is one of the

* http://www.csie.ntu.edu.tw/~cjlin/libsvm/
One of the challenging aspects in our study is that we use only top-view images to protect privacy as shown in Fig. 4. Therefore, no faces or clothes can be observed. The view area was about 6m × 4.5m. The frame rate was 6.25 frames per second. The image size was 720 × 540. Sample images are shown in Fig. 6.

The 60-minute data recorded in the afternoon on Sunday among one-month length data were used for the experiments. The number of detected pedestrians and their attributes are summarized in Table 1. Here, the definition of “bag” includes from small bags such as handbags to large bags such as suitcases. The detected pedestrians’ size is typically 100 × 100. 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 one blob were eliminated by hand. Note that our focus is human attribute analysis and we assume that errors in detection and tracking can be solved by other approaches21. Ground truth was annotated by the authors.

The detected pedestrian regions were resized to 60 × 60 regardless of its original aspect ratio and HoG features were extracted. The k-means clustering was applied to a set of HoG features for all the frames of all the pedestrians to form bag-of-frames feature vectors.

### 5.2 Results using Gaussian Kernel

In this section, a Gaussian kernel was employed based on our preliminary results. The parameters for the SVM classifier were optimized for each case if not mentioned otherwise.

Fig. 7 shows the classification performance as a function of the codebook size. The graphs are the results of ten-cross validation. For the UNC algorithm, β was also varied from 0.001 to 10. In the hard voting approach8), (Fig. 7(a)), the performance is sensitive to the codebook size. Although the best accuracy was 94-96%, it goes down to 80-90% as soon as the parameters shifted from their optimal values. When the UNC is employed, if the proper β is chosen (β=0.1, Fig. 7(c)), the accuracy is very high and stable. However, if β is not appropriate, the accuracy becomes sensitive to the codebook size (β=0.001, Fig. 7(b)) or becomes even worse than the hard voting case (β=10, Fig. 7(d)). On the other hand, the proposed method shows its robustness to the codebook size and the accuracy is very high.

The gender classification performance as a function of the number of training data is shown in Fig. 8. This experiment shows how much discriminative ability the extracted features have. The training samples were picked up randomly and the rest was used for testing. Half of the training samples were positive and the other half were negative. This procedure was repeated 100 times. In this experiment, the cost for the constraints violation in the SVM classifier was set to 128 and the other parameters were set as default. In the hard voting approach8), the accuracy improvement over the number of the training samples is rather gentle, showing that the generated feature vectors are not discriminative and a lot of support vectors are needed. For the UNC model, only several samples are necessary when the codebook size and β are optimized. If the parameters are not optimal, it requires more than 100 training samples (β=0.001, Fig. 8(b)) or the performance is almost same as the hard voting (β=10, Fig. 8(d)). The proposed algorithm shows comparable performance to the optimized UNC model. Only about 10 samples are need if
the codebook size is 100 or larger. In this point of view, our model is easier to optimize while achieving comparative performance with the UNC model. The bag possession classification shows almost the same tendency as shown in Fig. 9.

5.3 Results with Different Kernels
In this section, we discuss how the performance of the human attribute analysis would change depending on the kernels. The performance of the gender classification is shown in Fig. 10 and that of the bag possession status classification is shown in Fig. 11. In addition to the Gaussian kernel that was used in Section 5.2, we tried with two more non-linear kernels:

Fig. 7 Mean accuracy of the gender and bag possession status classification over the ten-cross validation. The codebook size was changed. (a) hard voting based\(^8\), (b) UNC\(^3\), \(\beta=0.001\), (c) UNC, \(\beta=0.1\), (d) UNC, \(\beta=10\), (e) proposed.

Fig. 8 Gender classification accuracy over the 100 random sampling as a function of the number of training samples; (a) hard voting based\(^8\), (b) UNC\(^3\), \(\beta=0.001\), (c) UNC, \(\beta=0.1\), (d) UNC, \(\beta=10\), (e) proposed.
The number of training vectors (log-scale)

Accuracy (%)

10  100  1

40  50  60  70  80  90  100

k=100  k=10  k=1000

The number of clusters in k-means (log-scale)

Accuracy (%)

100  10

80  90  85  95  100

Chi-square  Histogram intersection  Gaussian

(a)

(b)

(c)

(d)

(e)

χ² : $k(x, y) = 1 - \sum_{i=1}^{m} \frac{2(x_i - y_i)^2}{(x_i + y_i)}$  (5)

Hist. intersection : $k(x, y) = \sum_{i=1}^{m} \min(x_i, y_i)$  (6)

where $k$ is the kernel, $x$ and $y$ are feature vectors, and $m$ is the dimension of the feature vector. The missing points mean that the classification performances were below 80%. For UNC($\beta = 0.1$) and our proposed model, the performance is almost the same when the codebook size is large enough. When the codebook size is small, the Gaussian kernel works much better than the other two kernels. For the hard voting and the UNC($\beta = 10$) cases, the best performance is almost the same although the codebook size that achieves the
best performance is different from each other. When the codebook size is shifted from the optimal value, the classification performance degrades with the same speed. For the UNC($\beta = 0.001$), the $\chi^2$ and the histogram intersection kernels worked poorly. It is also observed our proposed soft voting method worked fine even with other non-linear kernels. Users are advised to use the Gaussian kernel when it is required to use smaller codebook size to save memory and computational time such as in mobile devices. On the other hand, the $\chi^2$ or histogram intersection kernel would be proffered when the codebook size is large enough because these two kernels do not require the parameter optimization.

5.4 Performance for Object Recognition

Although the main scope of this paper is human attribute analysis using our proposed soft voting method, the proposed algorithm has also been applied to the conventional object classification. Please see Appendix A for the details.

6. Conclusions

In this paper, a soft-voting-based BoF representation has been proposed. The relative distance to that of the nearest-neighbor codeword was used for the soft voting. The proposed algorithm was compared with the hard voting method and one of the state-of-the-art soft voting methods called UNC model. The experimental results have shown that the proposed algorithm worked better and more robustly than the hard voting method and it was also comparable to the optimized UNC, which required time-consuming parameter optimization. The algorithm has been successfully applied to human attribute classification and object classification.

Acknowledgments

The author would like to thank Prof. Kiyoharu Aizawa for valuable discussion. The author also would like to thank Mr. Tomoaki Matsunami for his contribution to data collection and labeling.

References

7) T. Yamasaki and T. Matsunami. Pedestrian attribute analysis
using a top-view camera in a public space. In MMM, pages 541–550 (2012)

### Appendix

#### A. Object Classification using Caltech-101

##### A.1 Experimental Setup

The Caltech-101 is a well-known standard dataset to evaluate the object classification performance. The dataset contains 9,145 images in 101(+1 for background) classes including a variety of objects with large variance in shape. On the other hand, the position, orientation, and size of the objects are roughly aligned. The number of images per category was from 31 to 800. The images were resized to be no larger than 300×300 pixels with preserved aspect ratio. The images were all preprocessed into gray scale. Then, the SIFT descriptors extracted from 16 × 16 pixel patches were densely sampled from each image on a grid with step size of 6 pixels. The code assignment was done in a spatial pyramid manner as in Ref.30). Namely, the input image was divided into 1×1, 2×2, 4×4 and a feature vector is calculated in each sub-block. The final feature vector was obtained by concatenating all the feature vectors.

- **Table 1** Classification performance for Caltech-101 dataset. The results with (*) were obtained from the original papers and those without (*) are obtained by our own implementation.

<table>
<thead>
<tr>
<th>Code assignment method</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard voting $^{(*)}$</td>
<td>64.2 ± 0.8</td>
</tr>
<tr>
<td>Hard voting $^{(*)}$</td>
<td>67.0 ± 0.53</td>
</tr>
<tr>
<td>UNC$^{(*)}$, $\beta = 0$</td>
<td>64.4 ± 1.18</td>
</tr>
<tr>
<td>UNC$^{(*)}$, $\beta = 1$</td>
<td>77.5 ± 2.00</td>
</tr>
<tr>
<td>UNC$^{(*)}$, $\beta = 10$</td>
<td>66.7 ± 1.70</td>
</tr>
<tr>
<td>UNC$^{(*)}$, $\beta = 100$</td>
<td>71.3 ± 0.66</td>
</tr>
<tr>
<td>Proposed</td>
<td>69.5 ± 1.19</td>
</tr>
</tbody>
</table>
in the sub-blocks and normalizing it. A simple linear SVM was used as the classifier.

We mostly followed a common testing procedure as in Refs.34,39): 30 samples were randomly picked up from each class and the rest of them were used for testing. This process was repeated 5 times and the performance was measured by using average accuracy over 102 classes (102 accuracy values were averaged). The codebook sizes were set as 1,024.

A.2 Results

The classification performance is summarized in App. Table 1. When the $\beta$ is chosen properly, the accuracy is high (71.3 $\pm$ 0.66% at $\beta = 10$). However, the performance of the UNC model degrades when the $\beta$ value shifts from its optimal value and becomes even worse than the hard voting method. It can also be observed that the optimal $\beta$ is different from that for the human attribute analysis. It indicates that the $\beta$ value needs be optimized for each dataset. Finding the optimal $\beta$ in a brute forth manner is very time consuming. It is possible to conduct $\beta$ optimization with a subset of the data but there is always an overfitting problem.

Note that some works that are more accurate than the UNC model$^{35}$ and our model are already proposed, but the absolute value of the classification accuracy is not very important in this paper. The focus of this paper is the sub-optimality of our proposed method as compared to the UNC model without the time-consuming parameter tuning. Therefore, the relative accuracy to the UNC model is important. If sophisticated algorithms such as affine SIFT with multi-class AdaBoost$^{40}$ or geometric $l_p$-norm feature pooling$^{41}$ is employed, the accuracy would be further improved. However, it is out of scope of this paper.

Toshihiko Yamasaki received the B.S. degree in electronic engineering, the M.S. degree in information and communication engineering, and the Ph.D. degree from The University of Tokyo in 1999, 2001, and 2004, respectively. He is currently an Associate Professor at Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. His current research interests include 3D video processing, 3D computer graphics, analog VLSI design, and so on. Dr. Yamasaki is a member of ITE, IEEE, ACM, IEICE, IPSJ, and so on.

Tsuhan Chen earned his B.S. from National Taiwan University and his M.S. and Ph.D. from Caltech, all in electrical engineering. After working for Bell Labs for several years, he joined the ECE faculty at Carnegie Mellon University (CMU) in 1997. There, in addition to research and teaching responsibilities, he has served as associate department head of ECE and co-director of the Industrial Technology Research Institute (ITRI) Laboratory at Carnegie Mellon, a collaborative research program with ITRI in Taiwan.