Semantic Attribute Classification Related to Gait

Imen Chtourou[1(✉)], Emna Fendri[2(✉)], and Mohamed Hammami[2(✉)]

[1] MIRACL Laboratory, ENIS, University of Sfax,
Road Sokra km 4, BP 1173, 3038 Sfax, Tunisia
imene.chtourou@gmail.com
[2] MIRACL Laboratory, FSS, University of Sfax,
Road Sokra km 4, BP 802, 3038 Sfax, Tunisia
fendri.msf@gnet.tn, mohamed.hammami@fss.rnu.tn

Abstract. Human gait, as a behavioral biometric, has recently gained significant attention from computer vision researchers. But there are some challenges which hamper using this biometric in real applications. Among these challenges is clothing variations and carrying objects which influence on its accuracy. In this paper, we propose a semantic classification based method in order to deal with such challenges. Different predictive models are elaborated in order to determine the most relevant model for this task. Experimental results on CASIA-B gait database show the performance of our proposed method.

Keywords: Pedestrian analysis · Semantic attributes · Classification

1 Introduction

In the two past decades, surveillance cameras are widespread in many places such as train stations, parking lots, airports, banks, etc. Therefore, surveillance video applications have drawn a large amount of research attention in the world. Several biometric identifiers are used in these applications. Biometrics identifiers used currently include signature, fingerprint, palm vein, face, iris, retina scans, etc. These identifiers are not convenient for busy environments since they all need high quality images from close distances and the subjects cooperation. Human gait as a behavioral biometric identifier has received significant interest in recent years. This is due to its unique characteristics such as unobtrusiveness, recognition from distance and no need of high quality video. However, there are difficulties that face several gait based methods caused by covariate conditions that affect the gait negatively. Examples of these covariate conditions that commonly occur in real life are changes in clothing and the carrying objects (such as carrying a bag). These covariate conditions create problems for practical gait based application and significantly deteriorate their performance. They occlude the appearance of the body shape. In addition, they have an effect on the dynamic pattern of body movements. Whenever there are occluded pixels as

A. Abraham et al. (Eds.): ISDA 2017, AISC 736, pp. 508–518, 2018.
https://doi.org/10.1007/978-3-319-76348-4_49

a result of carrying a bag or wearing a baggy coat, it is normal that the accuracy decreased. A solution for these covariates may be adopted to use parts which are unaffected by these items in re-identification application. In this paper, a new method relying on the detection of these covariates (i.e. clothing variation and carrying objects) is proposed. To this end, we have used semantic attribute as they are human-understandable properties. Semantic attribute classification related to gait may be investigated for several applications such as gait recognition or gait based person re-identification, etc. The rest of this paper is organized as follows. Section 2 summarizes some related works. Section 3 introduces the proposed method. Section 4 describes the evaluation protocol. Section 5 concludes this paper.

2 Related Work

Recently, semantic attributes, which are human-understandable properties, such as male, black eyes, long hair, etc. are gained more and more interests. This interest is due to their ability to infer high-level semantic knowledge. Vaquero et al. [7] identified people by a series of attribute detectors. They introduce an attribute-based people searching system in surveillance environments. Layne et al. [4] definite a set of human-understandable pedestrian attributes such as "longhair", "headphones", "male", "backpacks", "sunglasses", "v-necks" and "clothing" on person re-identification databases like Viper [1] and PRID [2]. The work of [8] proposed an approach to illustrate the appearance of pedestrian with several binary attributes like "is male", "has T-shirt", "glasses", "has jeans", "long hair", etc. They used a set of parts from Poselets [9] for extracting low-level features and perform subsequent attribute learning. Describing clothing appearance with semantic attributes is an appealing technique for many important applications. Yang et al. [11] proposed a clothing recognition system that identifies clothing categories such as suit, T-shirt and Jeans in a surveillance video. Their method was based essentially on color (i.e. colour histogram) and texture features (i.e. histogram of oriented gradient HOG, a bag of dense SIFT features and DCT responses). Linear Support Vector Machines (SVM) classifiers have been trained in order to learn clothing categories. Chen et al. [10] proposed a method that comprehensively describes the upper clothing appearance with sets composed of multi-class attributes and binary attributes. They consider high-level attributes such as clothing categories and deal also with some very detailed attributes like "collar presence", "neckline shape", "striped", "spotted" and "graphics". Liu et al. [3] assemble a large online shopping database and a daily photo database for the research of the cross-scenario clothing retrieval. They define a set of clothing-specific attributes. These can be summarized into three classes, i.e., global, upper-body and lower-body attributes. Lower-body attributes can be further divided into two related attributes classes. Recently, Convolutional neural networks (CNN) has been adopted also in pedestrian attribute classification [15,17,18]. Zhu et al. [17,22] applied the learned CNN for person re-identification task. They used the pedestrian attribute classification by weighted interactions

from other attributes. In this method, a set of attribute was defined such as "male", "redshirt", "barelegs", "lightbottoms", "notlightdark", etc. Li et al. [16] also propose two deep learning models to learn the pedestrian attributes, one is called as deep learning based single attribute recognition model (DeepSAR) and the other is a deep learning framework which recognizes multiple attributes jointly (DeepMAR). Authors have used attributes like "Age 16–30", "Casual lower", "V-neck", "Sunglasses", "Formal lower", etc. Matsukawa and Suzuki [18] refined CNNs for attribute recognition and employ metric learning for person re-identification. They grouped attributes into 7 groups which are "Gender", "Age", "Luggage", "UpperBody Clothing", "UpperBody Color", "LowerBody Clothing", "LowerBody Color". Lin et al. [15] use person attributes as auxiliary tasks to learn more information. Attributes are composed of groups such as gender (male, female), color of shoes (dark, light), 8 colors of upper-body clothing (black, white, red purple, gray, blue, green, brown), age (child, teenager, adult, old), etc.

However, Attributes like "sunglasses", "headphones", "is male", "Age" and "neckline shape" used in the work of [3,4,8,10,11,15,16,18] may not alter or affect the natural gait appearance and dynamic pattern of body motion. Therefore, we have concentrated in this work on Single Shoulder Bag, Back Pack, Hand Bag and Outerwear attributes. These attributes can influence and occlude the gait based appearance of the body shape and consequently decrease accuracy. We have considered also carrying nothing as an attribute.

3 Proposed Method

We propose to build an automatic semantic attribute classification solution based on machine learning method using a set of manually classified attributes, in order to produce a predictive model, which allow predicting the class of each semantic attribute. To classify each semantic attribute into each one class, we are based on the Knowledge Discovery in Databases (KDD) process for extracting useful knowledge from volumes data [19]. The general principle of the classification method is the following: Let S be the set of image's samples to be classified. To each sample s of S one can associate a particular class of attribute, namely, its class label C. C takes its value in the class of labels (0 for the absence of attribute, 1 for the presence of attribute).

$$C : S \rightarrow \Gamma = \{existence, nonexistence\} \tag{1}$$

$$s \in S \rightarrow C(s) \in \Gamma \tag{2}$$

Our study consists in building a model to predict the attribute class of each persons image. The total process of the detection of pedestrian semantic attributes is composed of two steps: (i) an Off-line step and (ii) On-line step. Figure 1 shows the framework of the proposed method.

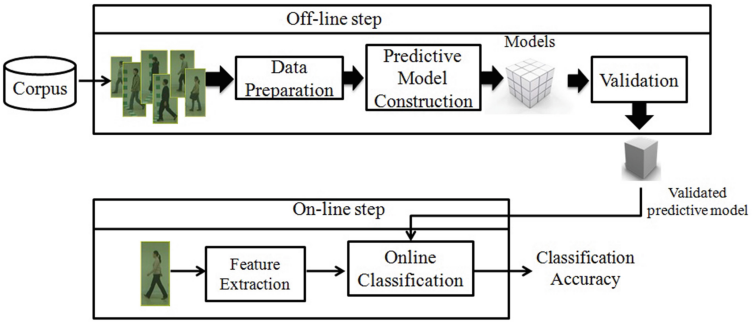

Fig. 1. The framework of the proposed method.

3.1 Off-line Step

The off-line stage involves three major sub-steps. First, we start with a data preparation of the training database related to semantic attribute. Second, a predictive model for each class of attribute is constructed. Third, a validation step is required.

Data Preparation. Given a corpus of images taken randomly, pedestrian's bounding box is detected from each image. The corpus of bounding box's images is divided in such a way that the training set contains equal number of positive and negative samples for each attribute. In this step, our goal is to construct a two-dimensional table from our training database. Each table row represents a bounding box's image and each column represents a feature. In the last column, we save the semantic attribute class denoted 1 for the presence of attribute and 0 for the absence of attribute. In our work, we are concerned with five different semantic attributes. Figure 2 gives the list adopted from CASIA-B database. It shows example of positive images for each attribute: Hand Bag, Single Shoulder Bag, Back pack, Outerwear and carrying nothing. The annotation practice was carried out automatically according to the name of images for carrying nothing (Normal), carrying Bags and wearing Coat (Outerwear) and to the image's number for each category of bags. Low-level color and texture features have

Fig. 2. Example positive images for each attribute. From left to right: Hand bag, Single Shoulder Bag, Back pack, Outerwear and Carrying nothing from CASIA-B database.

been shown their robustness in describing pedestrian images [4]. Therefore, we extracted a collection of color features (i.e. color histograms in RGB, HSV and YCbCr color spaces) and texture features (i.e. Gabor and Schmid filters) to model each semantic attribute. Several conventional methods [4,12,13] have used the same features set composed of RGB, HSV, YCbCr, Gabor, Schmid. We extracted a 2784-dimensional feature vector from each bounding box of person image. Once the data preparation step is defined, our task is to perform machine learning using different classifiers in order to prepare model for each class of semantic attribute. Further, trained classes are tested for the classification accuracy and their corresponding results are presented in the experimental result section. There are several algorithms of supervised learning in the literature. Each having its advantages and disadvantages. We used three supervised algorithms from different families like the support vector machines [20], Tree bagger based decision tree [21] and the neural networks [14]. In the end, the best performer predictive model is chosen.

Support Vector Machines (SVM)

Classification by SVM (Support Vector Machines) [5] is performed by constructing a model that iteratively separates the training data into two classes. It is defined over a vector space where categorization is achieved by linear or nonlinear separating surfaces in the input space of the original data set [20].

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \tag{3}$$

$$Subject\ to \left\{ y_i (w^T \Phi(x_i) + b) \geq +1 - \xi_i,\ i = 1, ..., N \xi_i \geq 0,\ i = 1, ..., N \right. \tag{4}$$

ξ_i's are slack variables required to permit misclassifications in the set of inequalities, and $C \epsilon \mathbb{R}^+$ is a tuning hyper parameter, weighting the significance of classification errors to the margin width. Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function Φ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. Furthermore, $K(X_i, X_j) = \Phi(X_i)^T \Phi(X_j)$ is called the kernel function. In our work, we use a SVM using the histogram Intersection (HI) as kernel since our feature vectors are based on histograms, as formulated below.

$$K(X_i, X_j) = \sum_{k=1}^{n} min \left\{ x_i, x_j \right\} \tag{5}$$

where: $X_i = \left\{ x_1^i, ..., x_n^i \right\}$ and $X_j = \left\{ x_1^j, ..., x_n^j \right\}$ are two histograms with n-bins (in $\mathbb{R}^n$). HI kernel has been proved a positive result which makes it suitable as a discriminative classification kernel.

Neural Network (NN)

Neural networks [14] have been used in many image-related applications and exhibited good performances. NN is a network of simple neurons called perceptrons. The perceptron computes a single output from multiple real-valued inputs

for forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. Mathematically this can be written as:

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(wx^T + b) \tag{6}$$

Where w denotes the vector of weights, x is the vector of inputs, b is the bias and φ is the activation function.

In order to train a model based on an Artificial Neural Networks, we use a multilayer perceptron with the back propagation learning algorithm [24]. We have adopted a neural network with a three-layer architecture consisting of input, hidden and output layers for the prediction of each semantic attribute. The activation function for the neurons in the hidden layer and in the input layer is sigmoid function defined as below:

$$f(x) = \frac{1}{1 - exp(-x))} \tag{7}$$

In a back propagation network, a supervised learning algorithm controls the training phase. Then, the input and output (desired) data should be available, thus allowing the calculation of the error of the network as the difference between the calculated output and the desired vector.

Tree Bagger (FT)

Decision trees are a popular method for various machine learning tasks. Random forests is a notion of the general technique of random decision forests [21] that are an ensemble learning method for classification, regression and other tasks, that work by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The training algorithm for random forests assigns the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ...x_n$ with responses $Y = y_1, ...y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: After training, predictions for unseen samples $x^{'}$ can be made by averaging the predictions from all the individual regression trees on $x^{'}$:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x^{'}) \tag{8}$$

or by taking the majority vote in the case of decision trees.

Validation of the Predictive Model. The aim of this step is to measure the performance of the learned predictive model in order to guarantee the generality and effectiveness for future samples. Several possible metrics have been proposed in the literature to evaluate the quality of a predictive model. Among

this metrics, we have opted for Correct Classification Accuracy (CCA) which denotes the ratio of correctly classified images with the total number of images. It is defined by:

$$CCA = \frac{Number\ of\ correctly\ classified\ Image}{Total\ number\ of\ images} \tag{9}$$

3.2 On-line Step

Given a new pedestrian's bounding box, we start by extracting features to represent the global information of each image. Labels are generated depending on the semantic information of the image. These feature vectors design the input of the pre-learned predictive models.

4 Experimental Results

In this section, we present two series of experiments: The first serie of experiments is realized in order to show the classification accuracy of each semantic attributes using different supervised algorithms. In this serie of experiments, we used the CASIA-B database [23]. The second serie of experiments concerns comparison of our proposed method with Layne et al. [4]. This experiment shows the independence of our proposed method regarding the database used. In this serie of experiments we have used the VIPeR database [1]. Before presenting the results of the two series of experiments, we present in the next section a description of the two databases used CASIA-B [23] and VIPeR [1].

4.1 Description of Used Databases

CASIA gait database collected by the Institute of Automation of the Chinese Academy of Sciences [23]. Database B [23] is a large multi view gait database collected indoors with 124 subjects and 13640 samples from 11 different views ranging from 0 to 180°. In our experiments, we consider only (90°). This is motivated by the fact that gait information is more significant and reliable in the side view [6]. Each person is recorded six times under normal conditions, twice under carrying bag conditions and twice under clothing variation conditions. Table 1 shows the repartition of train and test sets for the five semantic attributes from the CASIA-B database.

VIPeR viewpoint invariant pedestrian recognition (VIPeR) database [1]. This database contains 632 person image pairs taken from two non overlapping camera views (camera A and camera B). Each image is scaled to 64×128 pixels. Images appearance exhibit significant variation in pose, illumination conditions with the presence of occlusions and viewpoint.

Table 1. Train and test sets repartition of the 5 semantic attributes from CASIA-B database.

Attribute	#train	#test
Carrying nothing	1488	496
Outerwear	744	248
Single shoulder bag	522	174
Hand bag	132	44
Back pack	90	30

4.2 First Serie of Experiments: Performance Evaluation

In this section, we presented the first series of experiment. Data from CASIA-B database is divided in such a way that each attribute had an equal number of positive and negative samples. We have used three supervised algorithms namely Support Vector Machines (SVM), Tree Bagger (FT) and Neural Network (NN). For SVM classifier, we used LIBSVM [5]. For neural network, we have adopted 10 hidden neurons and back propagation algorithm is chosen to evaluate classifier. For Tree bagger, we have used 100 trees. The classification results of the three supervised algorithms are presented in Table 2. Results shows that the neural network gives better accuracy than SVM and tree bagger for 4 semantic attributes (i.e. Outerwear, Single Shoulder Bag, Back pack and Hand Bag) higher than 89%. For the semantic attribute carrying nothing, SVM shows better performance than neural network. This confirms that neural network is more precise and efficient for detecting semantic attributes that alter human shape. Thanks to its effectiveness for the majority of semantic attribute classification, neural network (NN) will be adopted for our proposed method.

Table 2. Experimental results by the three algorithms on CASIA-B database.

Attribute	Classification accuracy (%)		
	Support vector machines	Neural network	Tree bagger
Outerwear	80.645	89.9	73.4
Single shoulder bag	79.885	89.4	78.2
Hand bag	84.09	92.3	65.9
Back pack	73.333	94.4	76.7
Carrying nothing	89.314	86.2	72.0

4.3 Second Serie of Experiments: Comparison with State-of-the-Art Method

We compared our attribute classification results with the popular work related of Layne et al. [4]. Our proposed method was tested using as probe set Camera A from the database VIPeR [1]. It should be noted that images are randomly taken. This serie of experiments shows the performance of our proposed method compared to the popular method of Layne [4] and it also shows that our selected predictive model is independent of the database used. Table 3 shows that classification accuracy for carrying nothing, outerwear (coat), single shoulder bag and hand bag attributes are higher compared to [4] results. Accuracy for back pack attribute is slightly lower. This is due to occlusion that may cover back pack such as arm and may consequently alter the shape.

Table 3. Comparison with state-of-the-art method.

Attribute	Classification accuracy (%)	
	Proposed method	Layne et al. [4]
Outerwear	45.8	–
Single shoulder bag	56.9	56.0
Hand bag	76.9	54.5
Back pack	60.0	68.6
Carrying nothing	70.0	69.7

5 Conclusion

In this paper, we have investigated the classification of semantic attribute that can not only influence and occlude the appearance of the body shape, but also have an impact on the dynamic pattern of body motion and consequently on accuracy. Different supervised algorithms were proposed, we have proved that using neural network as a supervised algorithm improves the attribute classification performance compared to the state-of-the-art method. Inspired by the promising performance, we will further explore how to adopt this semantic attribute classification solution in gait based task.

References

1. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, no. 5, pp. 1–7. Citeseer, October 2007
2. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis, pp. 91–102. Springer, Heidelberg, May 2011

3. Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S.: Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3330–3337. IEEE, June 2012
4. Layne, R., Hospedales, T.M., Gong, S.: Attributes-based re-identification. In: Person Re-identification, pp. 93–117. Springer, London (2014)
5. SchIlkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
6. Bashir, K., Xiang, T., Gong, S.: Gait recognition without subject cooperation. Pattern Recognit. Lett. **31**(13), 2052–2060 (2010)
7. Vaquero, D.A., Feris, R.S., Tran, D., Brown, L., Hampapur, A., Turk, M.: Attribute-based people search in surveillance environments. In: Workshop on Applications of Computer Vision (WACV), pp. 1–8. IEEE, December 2009
8. Bourdev, L., Maji, S., Malik, J.: Describing people: a poselet-based approach to attribute classification. In: IEEE International Conference on Computer Vision (ICCV), pp. 1543–1550. IEEE, November 2011
9. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: IEEE 12th International Conference on Computer Vision, pp. 1365–1372. IEEE, September 2009
10. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Computer Vision – ECCV 2012, pp. 609–623 (2012)
11. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: 18th IEEE International Conference on Image Processing (ICIP), pp. 2937–2940. IEEE, September 2011
12. Nguyen, N.B., Nguyen, V.H., Duc, T.N., Duong, D.A.: Using attribute relationships for person re-identification. In: Knowledge and Systems Engineering, pp. 195–207. Springer, Cham (2015)
13. Umeda, T., Sun, Y., Irie, G., Sudo, K., Kinebuchi, T.: Attribute discovery for person re-identification. In: International Conference on Multimedia Modeling, pp. 268–276. Springer, Cham, January 2016
14. Yegnanarayana, B.: Artificial Neural Networks. PHI Learning Pvt. Ltd., New Delhi (2009)
15. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220 (2017)
16. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 111–115. IEEE, November 2015
17. Zhu, J., Liao, S., Yi, D., Lei, Z., Li, S.Z.: Multi-label CNN based pedestrian attribute learning for soft biometrics. In: International Conference on Biometrics (ICB), pp. 535–540. IEEE, May 2015
18. Matsukawa, T., Suzuki, E.: Person re-identification using CNN features learned from combination of attributes. In: 23rd International Conference on Pattern Recognition (ICPR), pp. 2428–2433. IEEE, December 2016
19. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Commun. ACM **39**(11), 27–34 (1996)
20. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. ACM, July 1992
21. Zighed, D.A., Rakotomalala, R.: Graphes d'induction: apprentissage et data mining. Hermes, Paris (2000)

22. Zhu, J., Liao, S., Lei, Z., Li, S.Z.: Multi-label convolutional neural network based pedestrian attribute classification. Image Vis. Comput. **58**, 224–229 (2017)
23. Zheng, S.: CASIA gait database (2005). http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp
24. Shih, F.Y.: Image Processing and Pattern Recognition: Fundamentals and Techniques. Wiley, Hoboken (2010)