Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6182654, 12 pages
https://doi.org/10.1155/2020/6182654

Research Article

Sex Classification via 2D-Skeletonization

Miguel Contreras-Murillo,[1] **Sergio G. de-los-Cobos-Silva**,[2] **Pedro Lara-Velázquez**,[2] **Eric A. Rincón-García**,[2] **Román A. Mora-Gutiérrez**,[3] **and Miguel Á. Gutiérrez-Andrade**[2]

[1]*Posgrado en Ciencias y Tecnologías de la Información, Universidad Autónoma Metropolitana-Iztapalapa, Ciudad de México 09340, Mexico*
[2]*Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana-Iztapalapa, Ciudad de México 09340, Mexico*
[3]*Departamento de Sistemas, Universidad Autónoma Metropolitana-Azcapotzalco, Ciudad de México 02200, Mexico*

Correspondence should be addressed to Miguel Contreras-Murillo; miguelcontrerasmurillo@xanum.uam.mx

Received 26 June 2020; Revised 31 August 2020; Accepted 23 September 2020; Published 23 November 2020

Academic Editor: Pablo Gil

Sex classification is a challenging open problem in computer vision. It is useful from statistics up to people recognition on surveillance video. So far, the best performance can be achieved by using 3D cameras, but this approach requires the use of some especial hardware. Other 2D approaches achieve good results on normal situations but fail when the person wears loose clothing and carries bags or the camera angle changes as they rely on calculating borders, silhouettes, or the energy of the person in the image. This work aims to provide a novel sex classification methodology based on the creation of a virtual skeleton for each individual from 2D images and video; then, the distances between some points of the skeleton are measured and work as input of a sex classifier. This improves the results since clothing, bags, and the camera angle affect little the skeletonization process.

1. Introduction

One of the first algorithms to implement biometrics was the Bertillonage system for people identification proposed in 1879 by Alphonse Bertillon, a French police officer. This method was based on the measurement of different parts of the body and the description of unique marks in each person [1, 2]. This strategy is still used usually as support to locate people, as it is the way in which people are described semantically.

Ever since, the use of hard and soft biometrics to identify people has increased as the technology necessary for its automation has been improved. Hard biometrics include features that are unique on each subject and can be physical (fingerprints, iris, face, etc.) or behavioural (voice, movement, etc.); while soft biometrics are those that can be shared among several individuals and can be used to describe a person semantically (sex, height, weight, etc.).

Nowadays, most of the automatic people recognition systems rely on hard physical biometrics given that the effectiveness of the algorithms improves as the differentiation between samples increases. However, they require a very close or even direct interaction with special hardware (fingerprint scanners, 3D cameras, etc.). In addition, these algorithms usually compare the obtained samples with each one of those stored in a database. Therefore, this strategy can have a negative impact on the performance and accuracy of the classification when applied in very large populations, since the number of comparisons required increases with each additional individual.

On the other hand, soft biometrics are very useful to describe people verbally. They are not invasive as no near interaction with the hardware is required and can easily group people that share the same features. However, soft biometrics can be similar in different individuals and their use for identifying people is limited to small populations [3, 4].

Although hard biometrics are more accurate than soft ones, it is not always possible to have direct collaboration from the person whom is being searched or the hardware cannot be near enough to bring good results. For these cases,

the best option seems to be a combination of hard and soft descriptors for the identification of people through no invasive techniques, such as methods based on surveillance camera systems.

The sex (a biological concept) and gender (a social construct) of a person are shown in different forms. In biology, the term sexual dimorphism is a condition where some physical characteristics in individuals from the same specie differ between their sexes. While gender expression is the behaviour or appearance of people which is used to show their gender, and is not always related to their biological sex, gender identity, or sexual orientation.

Some of the most effective noninvasive methods that can be applied to classify sex are based on four main gait identification approaches [5]. The first one compares the contour of the person with some labelled samples and classifies it with the most similar one. The second approach uses a similar strategy; however, in this case, the entire silhouette is used, and not only the contour, therefore, segmentation errors occur less frequently. The third one is based on the gait energy, which is the average silhouette from all the frames in a walking cycle, and this energy silhouette is compared with those of all the people samples stored for the algorithm training; therefore, this methods are ineffective if the silhouette is modified, as a mismatch occurs; moreover, the camera must capture the person in the same camera angle each time [6–8]. The last approach for the identification of people by their gait uses RGB-D images. This technology provides depth information from the image and simplifies its segmentation, achieving a clearer view of the object or person being recorded. This last method is not only useful on gait and sex identification, but also on other applications such as action recognition [8–10] and representation [11]. Sex detection is especially useful to find missing or wanted people, when there is only a semantic description or not enough information to apply other automatic searching techniques, and the process is manual, or manually assisted, as it reduces the searching area by retrieving only some matching samples.

In supervised learning, the machine learns to identify different classes given some input. As the human intervention during the training process is almost null, the algorithm selects some features autonomously to identify. Other sex classifiers bring the whole image or silhouette of a person as input and the classified sex results as output in a single stage, and it makes the machine learn how to identify the most common shapes in the input figure; despite the common emergence of these shapes with a specific gender, they might not even be related to the gender expression of a person. This is the way most algorithms fail to classify sex when the clothing of the subject is very loose, as they learn to identify only the way people dress. Using skeleton information in a two-stage process controls the data, the algorithm learns and avoids problems when the person wears loose clothing.

In this paper we propose a novel two-stage approach for sex detection based on virtual skeleton information [12–16]. Thanks to the use of skeleton information, our algorithm can get good results even when a person carries luggage or dresses differently, as this method does not rely on silhouettes; in addition, it is not necessary to retrieve several frames from a walk cycle as some other methods based on energy and gait would do. The proposed methodology is evaluated on the CASIA B dataset [17, 18], as it has become the main benchmark for other research studies about sex detection. As this database was not originally intended for 2D-skeletization, it was necessary to manually label 17,732 key points in order to train the classifiers, and the resulting body part labels for every individual are provided as part of this investigation, see Supplementary Materials (available here).

Sexual dimorphism in kids and adolescents is not enough to be accurately classified by this algorithm. It also requires the individuals to be in a standing position and without any major partial obstruction in the input image. Therefore, this research (and the other compared state-of-the-art research studies) focuses on the sex classification of adult pedestrians in a full-body format.

The remaining of this paper is organized as follows. In the next section, some important works are presented. In a later section, we describe the main steps of our methodology. The experimental results are included in the *Results and Discussion* section. In the last section, we include some conclusions.

2. Related Work

There are different methods to recognize people by their body, but they can be classified into two main approaches: based on dynamics models and based on statics models. The first approach analyses the movements that occur during a walking cycle, while the second uses all the information that can be obtained from a static body in only one video frame.

At the beginning of this century, some articles were focused on silhouette comparison and worked as basis of the subsequent research. It is important to analyse them since similar techniques have been used for identifying sex in people.

In [19], the dynamic of walking people is analysed. Since the height of the body varies as every step is taken, these movements can be analysed as unique waves for each person during a walking cycle. First, a clean silhouette of the person is obtained. In order to do so, a picture of the background is taken, and it is subtracted when a person appears in the scene. Next, the centroid of the silhouette is calculated and normalized according to the angle of the person with respect to the camera. Finally, the wave generated on a walking cycle of the person is calculated and compared with those generated by a known person.

In [17], Yu et al. proposed a framework to evaluate the performance of gait recognition algorithms. Their proposal included more than 15,000 sample videos that were taken from 124 volunteers, 93 are men, and 31 are women, with simultaneous shots from 11 camera angles with respect to the camera, ranging from 0° up to 180°, including multiple shots of the background and others from the same person that were taken under different conditions: normal gait, with luggage and with coats. This base of images, included in the

CASIA Gait Dataset, was originally created to identify how much the camera angle and the clothing of the person on each frame affect classification algorithms. However, since it is a very complete database [20], it has been referenced in many investigations of different research areas.

In [21], several silhouettes are calculated during a walking cycle and the average of every one of them generates a unique energy silhouette. Later, fuzzy logic is applied to compare each pixel against its 8 nearest neighbours in order to identify the sex of the person in the scene. Each neighbour pixel is assigned with a binary value; 1 if its value is greater than the pixel evaluated and 0 otherwise. The binary values of the 8 contiguous neighbours are stored as a byte and then converted into a decimal value, and the resulting histogram of the segmented energy silhouettes is calculated and works as input on a classifier. The performance of this algorithm is evaluated in the image database described in the previous paragraph and was able to improve the results obtained so far. However, it keeps the same error tendency when classifying people with luggage, very loose clothing, or the gait angle with respect to the camera changes.

It is also possible to analyse the movements of a person by following the movements of his or her joints or the complete skeleton [8, 22]. In [2], the measurements of different parts of the body, the distance of the strides, the time of each cycle, and the speed of the individual are calculated. Later, this information is compared with that stored in the database with k-NN, but it can become slow to calculate when handling a lot of data. However, in [2, 8], the authors conclude that anthropomorphic features can improve the accuracy of the algorithm.

Most investigations about 2D-skeletization aim to identify the activities that people do on video recording, but as each frame is analysed, it is also possible to retrieve skeleton information from a single image with high accuracy. To achieve this, unlike other research where the whole person is searched, each articulation is searched in the image using convolutional neural networks (CNNs) on a first step [23]. Later, confidence maps improve the results by defining areas where it is highly probable that a body part is found, and the areas with the lowest probability are discarded if they are under a threshold that can be dynamic or a predefined one [24]. The distances between points of interest are calculated to complete the skeleton (part affinity fields). This way, good results are obtained even when there are partial obstructions between the person and the camera [25].

3. Methodology

The methodology proposed in this paper determinates the sex of people based on skeleton key points retrieved from static images. Unlike other methods that rely on silhouettes, our proposal searches body parts of the subject in a single video frame via CNN, measures the distances between them, and utilises this information to make the final classification. By doing so, the person might be in different positions, change clothing, or carry bags without affecting the results. The performance of our methodology was trained and evaluated with a total of 5,456 frames and 17,732 labels from the CASIA Gait Dataset [17], since it is the most suitable and referenced set of images in similar works [21].

In the following sections, we present the main steps of our proposal. In the first stage, a CNN is trained to identify points of interest. In the second stage, we use the trained CNN to generate a virtual skeleton from an input image, and the sex of the person is classified from the distances between some points of interest in the skeleton. The most important steps are described as follows.

3.1. Stage 1: CNN Training. For the purposes of this investigation, it was necessary to train a CNN to identify a set of predefined points to form the virtual skeleton. Therefore, the following steps were applied in each of the eleven camera angles considered in the database. First, some video frames from CASIA B were selected, and some points of interest were manually labelled, cropped, and preprocessed. Finally, the resulting images were used to train eleven CNNs, one for every camera angle, to identify body parts. Each of these steps is described next.

3.1.1. Frame Selection. Unlike other methods, this proposal only requires one frame to classify a person as male or female. To do so, a frame where the person is in the middle of the scene is selected despite of their pose. This way, all samples are taken from a similar distance.

Detecting the exact frame where the individual is in the middle of the scene is achieved by selecting a zone that verifies the difference between two consecutive frames. If there is no difference, no moving objects appear in that zone, but when the person enters the centre of the frame, the difference increases, and the frame is selected.

3.1.2. Labelling. Given that our algorithm relies on supervised machine learning, we selected 1364 frames and manually labelled 17,732 points of interest (13 per image): head, shoulders, elbows, hand, hips, knees, and feet. To do so, we enclosed each body part in a square big enough to fully contain no more than one body part each time, as shown in Figure 1. Given that all images have the same resolution, the size of all labels was set as 28×28 pixels for this dataset. Next, for each square, we save the coordinates of its upper left corner with the name of the body part that it contains. In order to ease the comprehension of our proposal, some figures show only the upper body parts, but the same process is applied to the rest of the body.

We want to remark that the number of the frame in the video for each camera angle, the version of the sample, and the selected regions that belong to each label are available as part of this investigation in the Supplementary Materials section (available here).

3.1.3. Image Preparation. After the labelling is finished, each selection is cropped and stored in separate folders, one for each body part, as shown in Figure 2. The resulting label sets are used to train a CNN per camera angle.

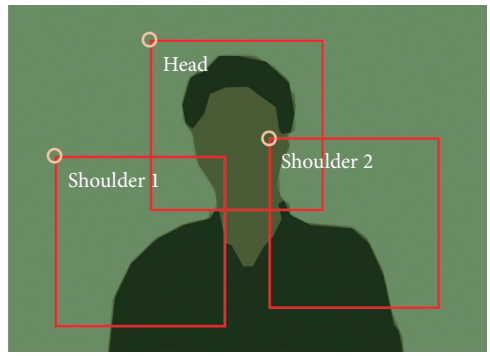
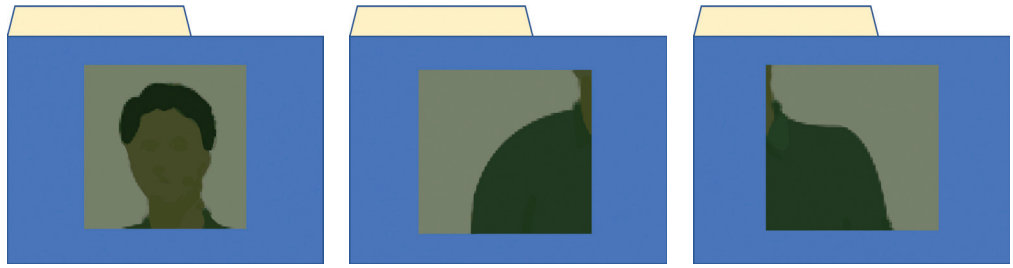


Figure 1: Labels of the head and shoulders.



(a) (b) (c)

Figure 2: Each cropped section is stored in a folder for each body part. (a) Head. (b) Shoulder 1. (c) Shoulder 2.

In order to improve the classification, all the present colours in the samples are corrected via white balancing. This process enhances the contrast on the image and decreases the differences of illumination between the scenes taken on different times of the day. To do so, the lightest pixel on the image is subtracted from the lightest white possible (with the highest illumination and lowest saturation), and the resulting values are added to every pixel.

3.1.4. CNN Training. The frame selection, labelling, and image preparation described above were repeated for every one of the eleven camera angles considered in the CASIA B dataset. In this way, we produced eleven sets of samples used to train different CNNs, which learn to identify body parts from a specific camera angle.

The input of a CNN is an RGB image from the training set, and both need to have equal dimensions ($3x28x28$). After an iteration, the output is a $13x1$ probability vector where each body part is more or less likely to be present in the input sample. The detailed architecture of the CNN is shown in Figure 3.

In this stage, a total of 17,732 manually labelled images were used (available as Supplementary Materials). Given the nature of this algorithm, 70% of the images are used for an iterative training and 15% for testing. To avoid overfitting, the model is validated during training with 15% of the images that are not included in either the training or testing sets.

The main steps of the first stage are included in Table 1 and Figure 4. Also, all scripts are available as Supplementary Materials (available here).

3.2. Stage 2: Sex Classifier. After the CNNs have been trained, they are implemented to find points of interest from a video frame, and the distances between these points are used to determinate the sex of the individual. The main steps of this stage are described next.

3.2.1. Image Preparation. Just like the labelled images used for training, changes in illumination might modify the colours of the scene. Therefore, the entire frame is also white balanced. In Figures 5(a) and 5(b), we show a scene before and after white balancing, respectively. This way, the differences on illumination between the input frames and the trained model are reduced. This strategy is applied to avoid any misclassification due to colour changes.

Next, our algorithm identifies if a zone belongs to the background or the foreground. Hence, it is not necessary to explore the whole image with the CNN to identify the points of interest. By doing so, the algorithm requires less iterations and false positive occur less frequently.

CASIA B includes background samples for this purpose [17]. With the cameras fixed in place, most of the pixels in the sample background and those in the analysed sample scene match and are removed from the image, remaining only those that belong to the foreground. For this purpose, the background is subtracted from the analysed scene, see Figures 6(a) and 6(b). The highest absolute values of the resulting image usually correspond to the most dynamic objects, as shown in Figure 6(c).

Finally, the picture is binarized in order to segment the image. In this fashion, all pixels in the foreground are

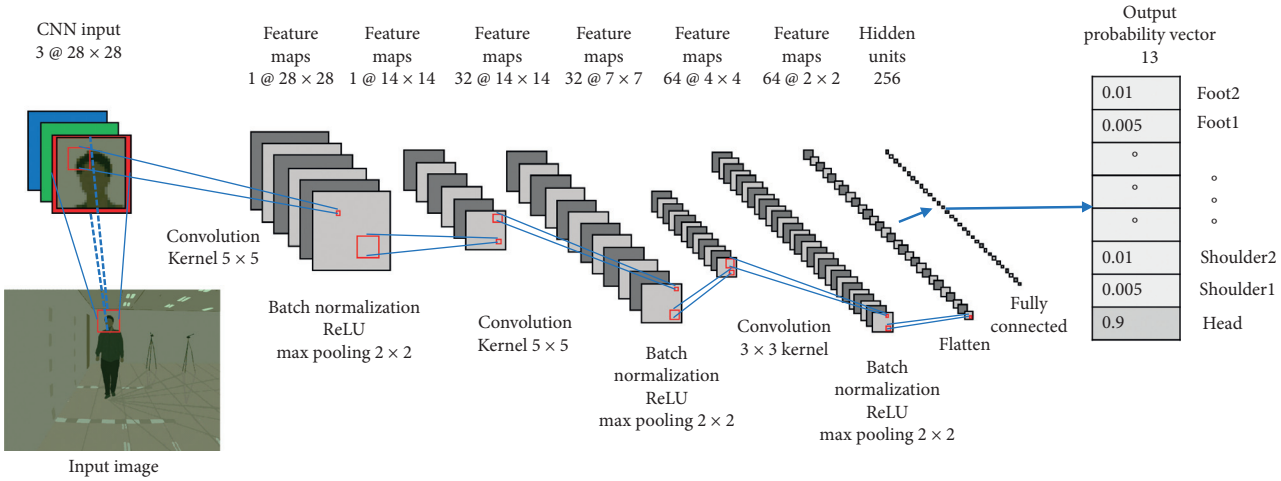


Figure 3: Architecture of the CNN.

Table 1: Pseudocode of the first stage.

Input: CASIA Gait Database (original videos)
Output: trained model for body part recognition
(1) Frame selection
(2) Labelling of points of interest
(3) Image preparation
(4) Model training (CNN)

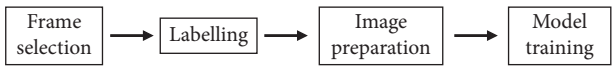


Figure 4: Flowchart of the first stage.

marked as 1, and those in the background as 0. When other objects move between scenes, or the background is similar to the clotting of the person, some noise appears in the image as shown in Figure 7(a). To correct this issue, the segments are dilated and eroded. It means that the border of the foreground is modified twice. First, it is increased to fill holes, and next, it is decreased to disappear smaller zones. The results of this process are shown in Figure 7(b).

3.2.2. Virtual Skeleton Generation. We decided to generate the virtual skeletons using a method similar to the technique described in [25]. We selected this strategy because it has reported the best results in 2D environments.

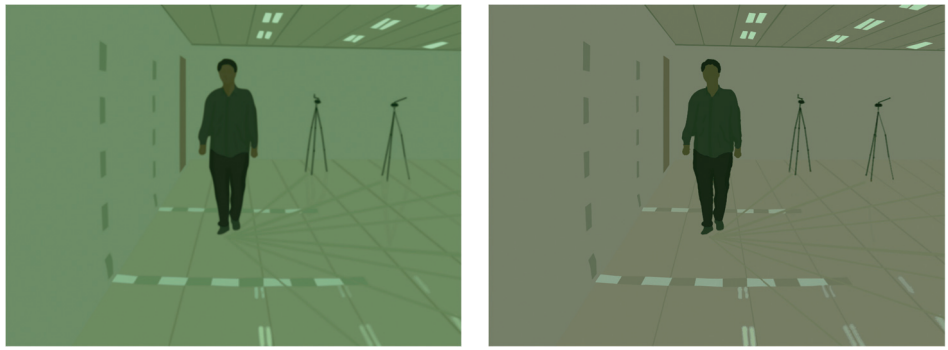
After the image is segmented, only the foreground is analysed through a $28x28x3$ sliding window. It is important to mention that this window needs to be equal in size to the input of the CNN and the labels used during the training stage. After evaluating each window, the CNN returns a probability vector that is used to create a confidence map for each point of interest, as seen in Figure 8.

When the confidence maps are created with the resulting values of the CNN, some false positive classification might occur. As seen in Figure 9, where some areas are misclassified as the head. Therefore, it is necessary to identify the pixels that correctly indicate the position of each part of the body. To do so, the next strategy is applied on every confidence map. First, the highest probability, $p_{\max}$, is identified. Later, all pixels with a value under a threshold, t, are eliminated. Given that $p_{\max}$ is not a fixed value and t needs to be recalculated on every confidence map, after some experiments, we concluded that the best option was $t = p_{\max} - 0.1$. This way, enough information is kept to calculate the coordinates of every skeleton point, while a value lower than $p_{\max} - 0.1$ would introduce noise in the confidence map and reduce its accuracy, and a higher value would eliminate the information needed in the next step.

Finally, equation (1) is implemented to calculate the weighted mean centre of the remaining pixels, where x_i and y_i are the coordinates of the i-th pixel, p_i is its probability, and n is the number of selected pixels. This way, the resulting centroid is placed inside the zone with the highest probabilities. In Figure 10, the resulting points for the head and shoulders are shown on the top of the foreground:

$$\overline{x} = \frac{\sum_{i=1}^{n} (x_i * p_i)}{\sum_{i=1}^{n} p_i},$$
$$\overline{y} = \frac{\sum_{i=1}^{n} (y_i * p_i)}{\sum_{i=1}^{n} p_i}. \quad (1)$$

3.2.3. Sex Classification. After the virtual skeleton of the person in the scene has been obtained, the Euclidean distances between every point of interest are calculated. Next, we used the height in pixels of the analysed person to normalise these distances, because the length varies according to the distance and angle with respect to the camera; otherwise, it would not be possible to tell if the person differs in height with the rest of the samples or is just closer or further in the scene.

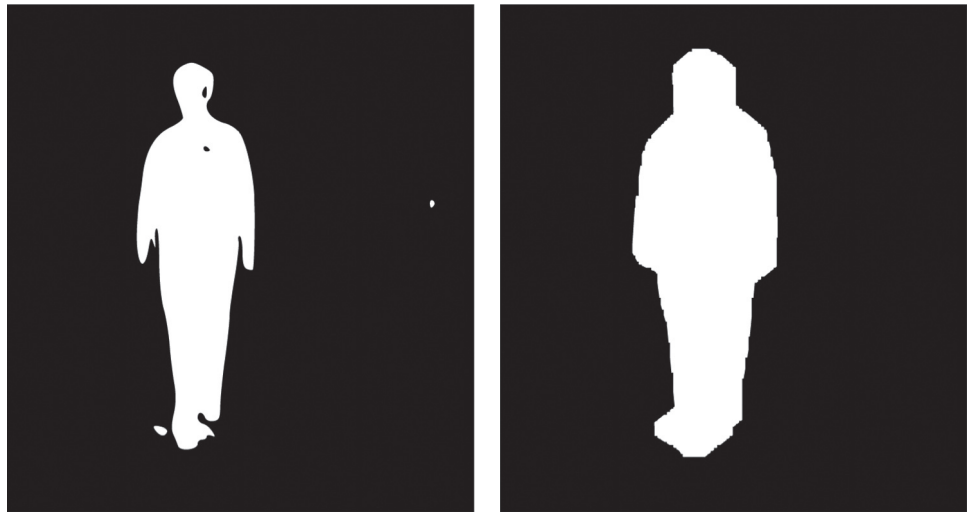


(a) (b)

Figure 5: White balancing.

(a) (b) (c)

Figure 6: (a) Background, (b) input scene, and (c) foreground.



(a) (b)

Figure 7: (a) Binarized foreground and (b) foreground without noise.

Finally, all the distances are used as input for the sex classifiers. We considered 3 different algorithms: SVM, *k*-NN, and ANN. Their features and results are shown in the next section.

The main steps of the second stage are included in Table 2 and Figure 11. Also, all scripts are available as Supplementary Materials (available here).

4. Results and Discussion

In order to determine the most suitable classifier for sex determination, the resulting distances between the points of interest were evaluated with three different classifiers in in MATLAB®. First, we used both SVM and *k*-NN classifiers included in *Statistics and Machine Learning Toolbox 11.4*.

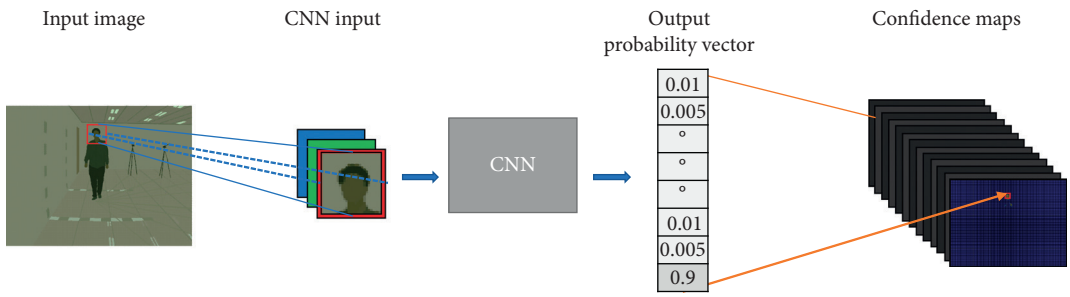
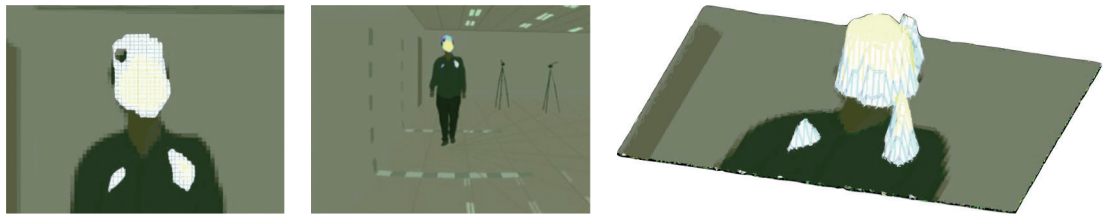


Figure 8: Each CNN generates a pixel in every confidence map.



(a) (b) (c)

Figure 9: Close-up on the confidence map of the head.

Finally, we implemented an ANN with the *Neural Network Start* toolbox. SVM and k-NN were trained and evaluated with *k-fold cross-validation* ($k = 5$), and for ANN, 70% of the data were used for training, 15% for validation, and 15% for the test.

We performed several experiments to find the best configuration for SVM, k-NN, and ANN. Finally, we decided to use the following configurations:

(i) Quadratic SVM : quadratic separation between classes.

(ii) Cosine k-NN : cosine similarity, 10 nearest neighbours.

(iii) ANN : three-layer feedforward backpropagation neural network, with a sigmoid function in the hidden layer with 20 neurons, and SoftMax in the output layer.

For the training of any supervised learning algorithms, it is necessary to provide data on the different scenarios that might occur. For this paper, we consider normal walking, bag carrying, and clothing changes on 11 camera angles. Once these models are trained, and the sex of each individual can be detected using the distances between the points of interest.

In order to evaluate and compare the performance of the results with other known algorithms, we decided to implement the $F1$-score as proposed on [21], where

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}, \quad (2)$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}.$$

In Tables 3–5, we include the $F1$-scores obtained by our proposals when applied to the CASIA B image dataset and those reported on previous investigations: *Local Binary Pattern* (LBP), *Local XOR Pattern* (LXP), *Ternary Pattern* (LTP), *Local Ternary Pattern* (LTP), *Soft LBP* (SLBP), *Fuzzy Local Binary Pattern* (FLBP), and *Local Binary Patterns With Tuned Parameters* (FLBP*).

We can observe that in most cases, our proposal with ANN reported the best scores, and in the few cases where it was surpassed by other techniques, the difference with the best score was small. For example, FLBP* reported better scores than our proposal in some camera angles in normal walking; however, the biggest difference between both algorithms is 0.07 for 72°, as shown in Table 3.

Although the results reported by FLBP* in normal conditions are good, they decay when the person in the scene carries luggage or wears loose clothing, as the generated silhouette is modified by these new conditions. Given that our proposal does not rely on those features, the results of the algorithms are not affected, as shown in Tables 4 and 5.

Similarly, in Figures 12–14, we include the boxplots of the scores reported in Tables 3 to 5. In Figure 12, we can see that FLBP* and ANN get the best performance for normal walking. However, according to Figures 13 and 14, ANN outperforms its counterparts when the person carries a bag or wears loose clothing. We consider that the main advantage of our proposal is the small variation of the distance between the points of interest in the virtual skeleton regardless of the camera angle, clothing, or accessories that the person wears, which helps to improve the performance of the classification. In fact, we must highlight that the scores obtained by SVM, k-NN, and ANN barely change in the three scenarios.

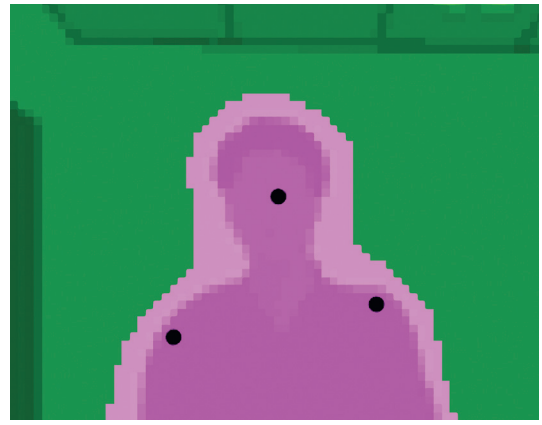


Figure 10: Resulting points for the head and shoulders.

Table 2: Pseudocode of the sex classifier.

Input: background image and frame to analyse
Output: class (male/female) that corresponds to the person in the input frame
(1) Image preparation
(2) Virtual skeleton generation
(3) Sex classification

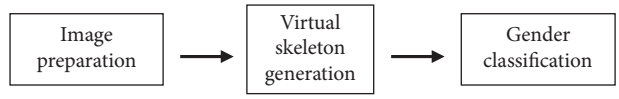


Figure 11: Flowchart of our proposal.

Table 3: $F1$-score in normal walking under different camera angles.

	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
LBP	0.86	0.92	0.91	0.90	0.85	0.86	0.86	0.87	0.92	0.91	0.88
LXP	0.88	0.92	0.93	0.92	0.91	0.84	0.89	0.87	0.90	0.90	0.93
LTP	0.87	0.85	0.88	0.85	0.89	0.80	0.84	0.84	0.88	0.86	0.88
SLBP	0.91	0.89	0.93	0.94	0.93	0.84	0.93	0.92	0.95	0.85	0.95
FLBP	0.93	0.92	0.94	0.97	0.92	0.90	0.91	**0.97**	0.97	0.92	0.96
FLBP*	**0.95**	**0.96**	**0.97**	**0.98**	**0.98**	0.94	**0.95**	**0.97**	**0.98**	0.96	**0.98**
SVM (prop.)	0.93	0.89	0.84	0.89	0.90	0.94	0.92	0.92	0.94	0.95	0.92
k-NN (prop.)	0.90	0.86	0.83	0.87	0.84	0.90	0.91	0.88	0.91	0.90	0.89
ANN (prop.)	0.94	0.94	0.94	0.94	0.91	**1.00**	0.94	**0.97**	0.97	**0.97**	0.94

Finally, in order to compare the overall performance of all the algorithms, we analysed all the scores reported for each technique according to Tables 3 to 5. In Figure 15, we include the boxplots for all algorithms considering their performance under all camera angles and walking conditions. According to this figure, we can assume that ANN has the best performance; however, we decided to confirm this observation using a Wilcoxon test, with a significance level of 5%.

The Wilcoxon test analyses the null hypothesis that the medians of the algorithms compared in pairs are identical. Again, we used all the scores reported in Tables 3 to 5. If the null hypothesis was accepted, a value 0 was assigned to both algorithms, which represents that both strategies exhibit the same behaviour (there is no significant difference between both sets of solutions). If the null hypothesis was rejected, a value of 1 was assigned to both algorithms. In Table 6, we can observe that FLBP* is statistically similar to SVM and k-NN, while ANN is different to all other techniques. If we combine the results of this analysis with the boxplots in Figure 15, we can conclude that our methodology with an ANN can produce the best results, outperforming the remaining

Table 4: F1-score in walking with luggage under different camera angles.

	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
LBP	0.84	0.74	0.73	0.71	0.72	0.79	0.77	0.75	0.82	0.86	0.80
LXP	0.84	0.80	0.77	0.69	0.68	0.73	0.73	0.75	0.85	0.85	0.82
LTP	0.78	0.78	0.81	0.71	0.73	0.70	0.66	0.75	0.73	0.78	0.86
SLBP	0.85	0.86	0.85	0.66	0.80	0.80	0.79	0.80	0.83	0.87	0.85
FLBP	0.83	0.82	0.81	0.74	0.60	0.79	0.83	0.81	0.89	0.92	0.87
FLBP*	0.91	0.91	0.89	0.82	0.86	0.86	0.87	0.81	0.90	0.94	0.93
SVM (prop.)	0.93	0.88	0.90	0.82	0.86	**0.92**	**0.93**	0.90	0.92	0.95	0.92
k-NN (prop.)	**0.94**	0.86	0.88	0.84	0.85	0.91	0.90	0.87	0.91	0.94	0.90
ANN (prop.)	**0.94**	**0.94**	**0.94**	**0.88**	**0.91**	0.91	0.91	**0.94**	**0.94**	**0.97**	**0.97**

Table 5: F1-score in walking with loose clothing under different camera angles.

	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
LBP	0.79	0.75	0.70	0.85	0.75	0.76	0.77	0.79	0.79	0.78	0.80
LXP	0.79	0.84	0.75	0.81	0.79	0.80	0.77	0.83	0.81	0.75	0.80
LTP	0.75	0.83	0.72	0.72	0.79	0.73	0.75	0.77	0.72	0.67	0.78
SLBP	0.74	0.83	0.79	0.78	0.81	0.80	0.79	0.83	0.82	0.83	0.84
FLBP	0.79	0.83	0.82	0.82	0.80	0.80	0.80	0.84	0.82	0.87	0.82
FLBP*	0.85	0.86	0.86	0.85	0.87	0.87	0.87	0.84	0.88	0.87	0.88
SVM (prop.)	0.91	0.87	0.85	0.88	0.90	0.88	0.93	0.92	0.91	0.93	0.92
k-NN (prop.)	0.91	0.86	0.86	0.88	0.86	0.90	0.91	0.93	0.90	0.90	0.91
ANN (prop.)	**0.94**	**0.91**	**0.91**	**0.91**	**0.91**	**0.97**	**0.97**	**0.94**	**0.97**	**0.97**	**0.94**

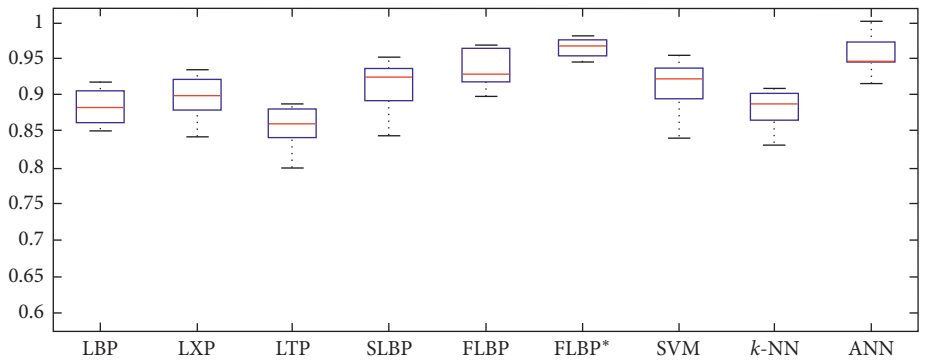


Figure 12: Normal walking.

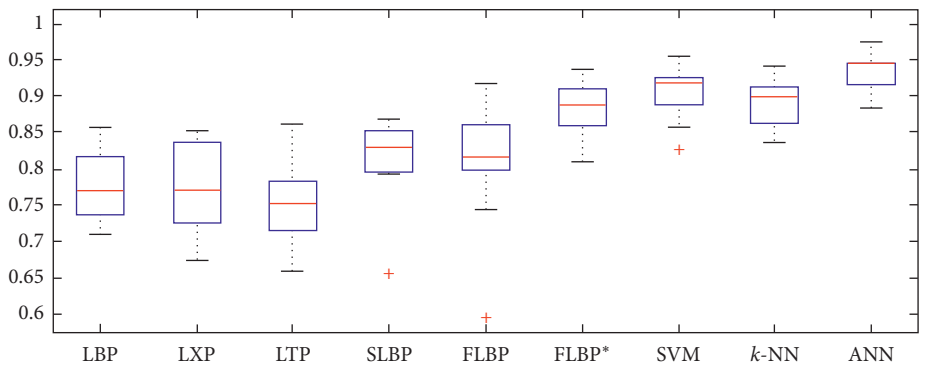


Figure 13: Walking with luggage.

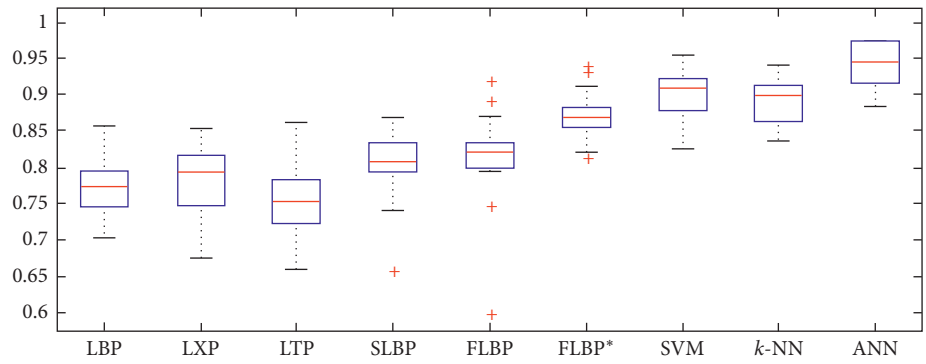


Figure 14: Walking with loose clothing.

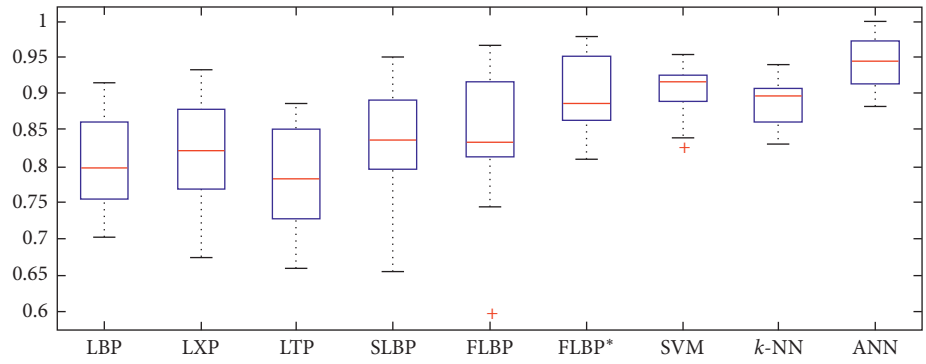


Figure 15: Boxplots considering all scenarios.

Table 6: Wilcoxon test binary results.

	LBP	LXP	LTP	SLBP	FLBP	FLBP*	SVM (prop.)	*k*-NN (prop.)	Ann (prop.)
LBP	1	1	1	0	0	0	0	0	0
LXP	1	1	1	1	1	0	0	0	0
LTP	1	1	1	0	0	0	0	0	0
SLBP	0	1	0	1	1	0	0	0	0
FLBP	0	1	0	1	1	0	0	0	0
FLBP*	0	0	0	0	0	1	1	1	0
SVM (prop.)	0	0	0	0	0	1	1	0	0
k-NN (prop.)	0	0	0	0	0	1	0	1	0
Ann (prop.)	0	0	0	0	0	0	0	0	1

strategies considered in this work. On the other hand, our methodology with SVM or k-NN, instead of ANN, can be considered the second-best option along with FLBP*.

5. Conclusions

In this work, we proposed a novel methodology for sex classification. First, we trained a convolutional neural network to identify the head, shoulders, elbows, hands, hips, knees, and feet of a person in a frame. Next, the distances between these points were used to classify the person as male or female. For this stage, we used three algorithms based on SVM, *k*-NN, and ANN. Finally, the performance of our methodology was evaluated on the CASIA B dataset and compared against 6 previously reported algorithms.

The numerical results show that our methodology performs better when used with ANN instead of SVM or k-NN. However, if all the available instances in CASIA B are considered, the two strategies have the best or the second-best behaviour among all the algorithms proposed in this work. Also, all labels and trained classifiers, used in this proposed methodology for sex classification, are available for any other similar research that might require them.

We consider that the main advantages of our proposal can be summarized in two points. First, we can use only one frame, instead of a section of a video, to get a high precision classification. Second, the small variations of the distances between the previously mentioned points in the virtual skeleton improve the overall performance of the classification algorithms. In fact, in Figure 15, we can see that the

range of scores for SVM, k-NN, and ANN is small, at least when compared with the remaining algorithms. Therefore, we can say that the algorithms had approximately the same behaviour regardless of the camera angle or the specific characteristics of the person in the frame. Finally, we conclude that our methodology can lead to improve the performance of classification algorithms for this kind of problems.

Data Availability

The models data used to support the findings of this study are included in the Supplementary Materials section. The CASIA B data supporting this research are from previously reported studies and datasets, which have been cited. The processed data are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Portions of the research in this paper use the CASIA Gait Database collected by the Institute of Automation, Chinese Academy of Sciences. During the elaboration of this research, the correspondence author received a scholarship by the CONACYT (the National Council of Science and Technology), CVU/Becario: 634201/340258.

Supplementary Materials

In S1–S4, we provide the scripts and files needed to try the proposed algorithm with the CASIA B dataset. The S5 contain the labels needed to retrain the algorithm with the scripts S6 and S7. (*Supplementary Materials*)

References

[1] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? A survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2016.

[2] V. O. Andersson and R. M. Araujo, "Full body person identification using the kinect sensor," in *Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, pp. 627–633, Limassol, Cyprus, November 2014.

[3] C. Prathap and S. Sakkara, "Gait Recognition using skeleton data," in *Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2302–2306, Kochi, India, August 2015.

[4] W. Min, M. Fan, J. Li, and Q. Han, "Real-time face recognition based on pre-identification and multi-scale classification," *IET Computer Vision*, vol. 13, no. 2, pp. 165–171, 2019.

[5] I. Rida, N. Almaadeed, and S. Almaadeed, "Robust gait recognition: a comprehensive survey," *IET Biometrics*, vol. 8, no. 1, pp. 14–28, 2019.

[6] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Automatic learning of gait signatures for people identification," in *Advances in Computational Intelligence*, pp. 257–270, Springer, Berlin, Germany, 2017.

[7] P. Arora, M. Hanmandlu, and S. Srivastava, "Gait based authentication using gait information image features," *Pattern Recognition Letters*, vol. 68, pp. 336–342, 2015.

[8] J. Kovač and P. Peer, "Human skeleton model based dynamic features for walking speed invariant gait recognition," *Mathematical Problems in Engineering*, vol. 2014, Article ID 484320, 15 pages, 2014.

[9] V. A. Chenarlogh and F. Razzazi, "Multi-stream 3D CNN structure for human action recognition trained by limited data," *IET Computer Vision*, vol. 13, no. 3, pp. 338–344, 2019.

[10] M. Camplani, A. Paiement, M. Mirmehdi et al., "Multiple human tracking in RGB-depth data: a survey," *IET Computer Vision*, vol. 11, no. 4, pp. 265–285, 2017.

[11] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Learning to recognise 3D human action from a new skeleton-based representation using deep convolutional neural networks," *IET Computer Vision*, vol. 13, no. 3, pp. 319–328, 2019.

[12] B. Dikovski, G. Madjarov, and D. Gjorgjevikj, "Evaluation of different feature sets for gait recognition using skeletal data from Kinect," in *Proceedings of the 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1304–1308, Opatija, Croatia, May 2014.

[13] A. H. Bayat, M. M. Arzani, M. Fathy, A. Matinnejad, B. Minaei-Bidgoli, and R. Entezari, "A probabilistic graphical model approach for human activity recognition using skeleton data," in *Proceedings of the 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, pp. 1–5, Tehran, Iran, December 2016.

[14] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from RGBD sensors," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 4351435, 14 pages, 2016.

[15] Q. Wu and G. Guo, "Gender recognition from unconstrained and articulated human body," *The Scientific World Journal*, vol. 2014, Article ID 513240, 12 pages, 2014.

[16] M. L. Anjum, S. Rosa, and B. Bona, "Tracking a subset of skeleton joints: an effective approach towards complex human activity recognition," *Journal of Robotics*, vol. 2017, Article ID 7610417, 8 pages, 2017.

[17] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, pp. 441–444, Hague, Netherlands, September 2006.

[18] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *Proceedings of the 2011 18th IEEE International Conference on Image Processing*, pp. 2073–2076, Beijing, China, September 2011.

[19] L. Wang and T. Tan, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, p. 14, 2003.

[20] D. López-Fernández, F. Madrid-Cuevas, A. Carmona-Poyato, M. Marín-Jiménez, and R. Muñoz-Salinas, "The AVA multi-view dataset for gait recognition," *International Workshop on Activity Monitoring by Multiple Distributed Sensing*, Springer, Berlin, Germany, 2014.

[21] E.-S. M. El-Alfy and A. G. Binsaadoon, "Automated gait-based gender identification using fuzzy local binary patterns

with tuned parameters," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 7, pp. 2495–2504, 2019.

[22] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta, "Abnormal gait detection with RGB-D devices using joint motion history features," in *Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 7, pp. 1–6, Ljubljana, Slovenia, May 2015.

[23] S. Yan, Y. Xia, J. S. Smith, W. Lu, and B. Zhang, "Multiscale convolutional neural networks for hand detection," *Applied Computational Intelligence and Soft Computing*, vol. 2017, Article ID 9830641, 13 pages, 2017.

[24] F. Poiesi, R. Mazzon, and A. Cavallaro, "Multi-target tracking on confidence maps: an application to people tracking," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1257–1272, 2013.

[25] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.