symmetry

MDPI

Article

ILRA: Novelty Detection in Face-Based Intervener Re-Identification

Pedro A. Marín-Reyes [1,*,†], Itziar Irigoien [2,†], Basilio Sierra [2,†], Javier Lorenzo-Navarro [1,†], Modesto Castrillón-Santana [1,†] and Concepción Arenas [3,†]

1 Instituto Universitario SIANI, Universidad de Las Palmas de Gran Canaria (ULPGC), 35017 Las Palmas, Spain
2 Department of Computer Science and Artificial Intelligence, UPV-EHU, 20018 Gipuzkoa, Spain
3 Statistics Section: Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, 08028 Barcelona, Spain
* Correspondence: pedro.marin102@alu.ulpgc.es
† These authors contributed equally to this work.

Received: 22 July 2019; Accepted: 8 September 2019; Published: 11 September 2019

Abstract: Transparency laws facilitate citizens to monitor the activities of political representatives. In this sense, automatic or manual diarization of parliamentary sessions is required, the latter being time consuming. In the present work, this problem is addressed as a person re-identification problem. Re-identification is defined as the process of matching individuals under different camera views. This paper, in particular, deals with open world person re-identification scenarios, where the captured probe in one camera is not always present in the gallery collected in another one, i.e., determining whether the probe belongs to a novel identity or not. This procedure is mandatory before matching the identity. In most cases, novelty detection is tackled applying a threshold founded in a linear separation of the identities. We propose a threshold-less approach to solve the novelty detection problem, which is based on a one-class classifier and therefore it does not need any user defined threshold. Unlike other approaches that combine audio-visual features, an Isometric LogRatio transformation of a posteriori (ILRA) probabilities is applied to local and deep computed descriptors extracted from the face, which exhibits symmetry and can be exploited in the re-identification process unlike audio streams. These features are used to train the one-class classifier to detect the novelty of the individual. The proposal is evaluated in real parliamentary session recordings that exhibit challenging variations in terms of pose and location of the interveners. The experimental evaluation explores different configuration sets where our system achieves significant improvement on the given scenario, obtaining an average F measure of 71.29% for online analyzed videos. In addition, ILRA performs better than face descriptors used in recent face-based closed world recognition approaches, achieving an average improvement of 1.6% with respect to a deep descriptor.

Keywords: re-identification; open world scenario; novelty detection; one-class classification; ILR transformation; local descriptors; deep descriptor

1. Introduction

Person re-identification is the process of recognizing an individual over different non-overlapping camera views [1–6]. Usually, probe is used to refer to the image of the individual to be recognized and gallery to the set of images of known people where the probe has to be recognized. Re-identification problems can be classified into different categories depending on the considered dimension [2]: sample set, body model, etc. Bedagkar-Gala and Sha [5] propose a wider taxonomy based on the mandatory presence or not of the probe in the gallery. Thus, a closed world, or closed set, scenario

is similar to the classic matching problem with a fixed size gallery. In an open world, or open set, the probe does not necessarily belong to the gallery, which evolves dynamically, adding new identities as the re-identification process takes place.

In the open world re-identification scenario, firstly, it is necessary to decide whether the probe belongs to the gallery or not. If the probe belongs to the gallery, a matching process is carried out; otherwise, the probe is added to the gallery as a new identity. The first stage in an open world re-identification scenario is very similar to the problem of novelty detection [7–9], which refers to the identification of new or unknown individuals, who were not previously registered in the system. Those individuals are denominated atypicals in opposition to those registered, who are referred to as typicals.

Speaker diarization [10] can be considered a similar problem to person re-identification. In the former, systems try to answer the questions of who spoke when. The difference lies in the scenarios where they are applied. Person re-identification is considered mostly in video surveillance scenarios where there is no audio, and coarse views of the people are obtained, so appearance based methods are widely used [2]. On the contrary, speaker diarization is carried out in video recordings (news, talk shows or television debates) where audio and close views of the participants are available. The availability of audio and images allows the application of techniques that combine both information sources [10,11]. In addition, the intervener views are normally close frontal views that allow information of the face to be extracted, instead of the general appearance of the intervener, allowing the exploitation of the facial features that are almost symmetrical and uniform [12,13].

In this paper, a face based open world re-identification approach is presented in a parliamentary debate scenario. This is a challenging scenario because deputies can participate in the debate from different locations: speaker platform (top row in Figure 1), seats (second and third row in Figure 1) and presidential table (bottom row in Figure 1). These locations impose appearance variations in terms of pose and distance to the camera; therefore, a frontal face is not always available for each intervener during the debate. Thus, the main difference between usual speaker diarization scenarios, e.g., TV talk shows, and parliamentary debates, which makes the latter a challenging problem, is that there exists a higher variability in poses, from closeup intervener frontal views, to a general view where not only the intervener appears, but other deputies that are close to her/him (first image of the bottom row in Figure 1). In order to provide a solution to these situations, the contributions of this paper are threefold:

- We present a contextualization of open world re-identification problems.
- We propose a feature vector based on Isometric LogRatio (ILR) transformation of a posteriori probabilities of belonging to a known intervener, applying a previous descriptor calculated only over the intervener face.
- A threshold-less approach is used to solve the novelty detection problem in an open world scenario. Thus, there is not a need for any user defined threshold.

The remainder of this paper is organized as follows: Section 2 presents a review of recent literature in both re-identification and speaker diarization. Section 3 describes our methodology. Section 4 contains the experiment designs to evaluate our proposal and includes the achievements of the experiments. Section 5 deals with the advantages and disadvantages of the proposal, and, finally, conclusions are drawn in Section 6.

Figure 1. Deputy captures of the Canary Islands Parliament. These images show different problematic situations where correct (green) and incorrect (red) intervener matches are presented.

2. Related Work

In recent years, a dual, i.e., audio-visual, methodology in diarization has become popular. Bredin and Gelly [14] use television series to evaluate their diarization method. Their proposal is based on applying a clustering technique over the face images to assign the most co-occurring face cluster with the corresponding audio cluster. The latter is extracted from the linear Bayesian Information Criterion (BIC) clustering of the audio stream. Lastly, regular BIC clustering is used to obtain the final diarization. Unlike the previous authors, a multiple speaker detection approach that uses the position of the audio signals sources was proposed in [15]. Other authors [16] use the LIUM system, to extract the audio diarization and deformable part-based model (DPM) to detect visual faces. Later, a conditional random field based multi-target tracking is adopted to track the interveners. Subsequently, a clustering technique based on the similarity distances and biometric measures is applied. To assign the names, One-to-One Speaker Tagging is computed to maximize the co-occurrence duration between clusters and the names provided by an Optical Character Recognition (OCR). As opposed to previous works, in [17], the authors do not detect the faces. Instead, skin blocks are detected using the chrominance coefficients of the skin-tone in the YUV color space, where motion vectors are obtained. The Mel Frequency Cepstral Coefficients (MFCCs) of the audio stream are combined with the visual representation using a log-likelihood from two Gaussian Mixture Models (GMM).

Given that our proposal is based on a re-identification approach, we summarize some related works. The approach by Bazzani et al. [18] consists of splitting the individual body parts of the pedestrians. Features are extracted from the HSV color space using weighted histograms. Other features are extracted using an agglomerative clustering of the image pixels and the computation of texture patches. Moreover, in recent years, some researchers have introduced the use of metric learning techniques in the field of people re-identification. The aim of these techniques is to project the representation of the individuals in a feature space where those of the same individual are closer and those of different individuals are further apart. Authors in [19] propose the Keep It Simple and Straightforward (KISS) learning, improving the method using a regularization in order to suppress the effect of larger eigenvalues in the covariance matrices. Moreover, in [20], the authors describe a technique to find a common space in different camera views in an unsupervised context. Thus,

a *k*-means is used to cluster the person images from different views. Neural Networks are also commonly used to project the samples in a new sample space. In this sense, authors in [21] split the image into three grids and use this representation as input into a bilinear network to aggregate in a feature vector. These vectors are used to obtain a new embedding feature space using a Siamese network. This architecture is commonly used to verify the input samples. In [22], the authors add also an identification stage to the model.

As mentioned above, recent challenging scenarios in re-identification fields are those related to open world problems, where novelty detection is a must (Figure 2). Novelty detection is used in a large kind of context, such as [23] in wildlife scenes and [24] for temporal series of vital signs with gastrointestinal cancer surgery; in addition, diagnosis of dermal diseases and the analysis of lymphatic cancer have been treated [25] or in robotics scenarios [26]. More related with people re-identification but using audio cues, authors in [27] propose a novelty detection approach in a speaker diarization system. A likelihood ratio thresholding is applied, depending on the speaker gender; and it is normalized using the mean and standard deviation. This thresholding determines typical/atypical speakers. Despite previous approaches, we are focusing on visual based re-identification problems. Authors in [28] propose a novel transfer ranking approach for two types of verification, multi-shot and one-shot verification, in a bipartite ranking problem. They applied RankSVM and probabilistic relative distance comparison to obtain a model, which optimizes a margin parameter based on the typical intra-class and inter-class variations, and inter-class variations between typical and atypical images. Authors in [29] present a supervised subspace learning approach where a linear transformation of the features is learnt by the optimization of a cost function related to the proportion of positive and negative misclassified pairs. In order to determine the presence of a probe person in a gallery, they introduce a margin parameter such that pairs whose distance is lower than the threshold are considered as belonging to the gallery and not belonging to the gallery otherwise. Authors in [30] introduce a new person re-identification search setting where the main features are: a vast probe search population, fast disjoint-view search and sparse training person identities. Over this setting, they obtain a set of features from the cross-view identity correlation and identity discrimination verification. In the same way as previous authors, the novelty detection is based on a threshold over the distance between individual representations.

Open world re-identification problems have dealt with deep learning in recent years; in particular, generative networks are used. For instance, an unsupervised domain adaptation approach that generates samples for effective target-domain learning is presented in [31]. This is done under the assumption that datasets in different re-identification domains have entirely different sets of identities. Thus, a translated image should be of a different identity from any target image. In this way, a Cycle Generative Adversarial Network (CycleGAN) [32] is used to translate images from a source to a target domain. Then, a Siamese network pushes two dissimilar images away and brings similar ones closer, with the aim of classifying a sample as typical or atypical. In addition, authors in [33] take advantage of the benefit of integrating generated people images. On the one side, they use a person discriminator to verify whether the generated image is a person or not. On the other side, a target discriminator identifies if a person belongs to the dataset or not. The feature vector is extracted from the last fully connected layer of the target discriminator and a threshold is used to determine the novelty of the person.

Unlike the previous approaches in which most of them use a margin parameter to detect the novelty of an individual, our approach applies a one-class classifier [34] to determine the novelty of a person, without the need of tuning a threshold. The advantage of this classifier is that only positive samples are needed to train it, unlike other classifiers that make use of positive and negative samples in the training process. Furthermore, we propose the use of a feature vector based on ILR transformation of a posteriori probabilities of belonging to a known intervener, applying a descriptor calculated only over the intervener face that fits with the one-class classifier.

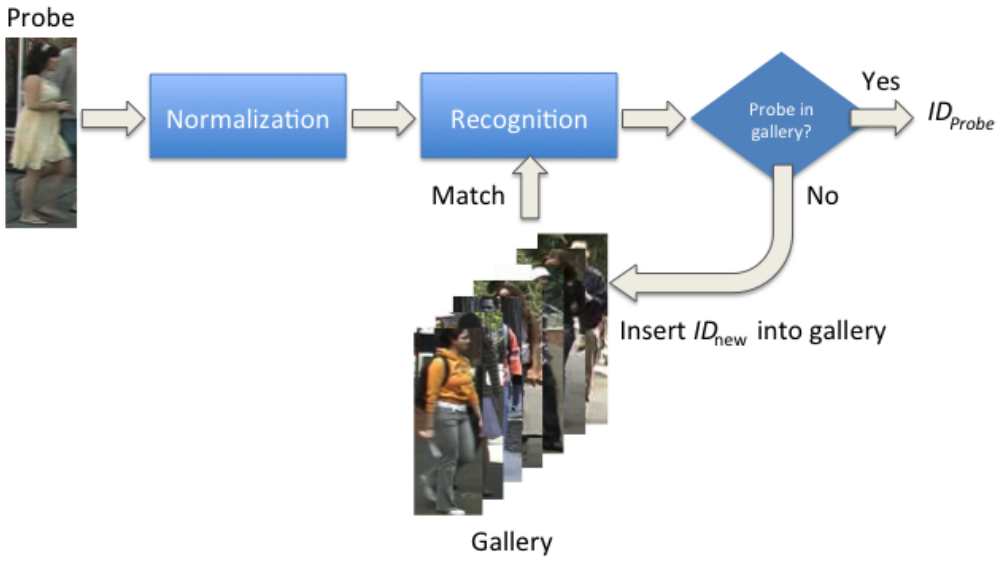

Figure 2. An overview of an open world re-identification system.

3. Method

In this section, firstly we outline the proposed approach, and then we explain in detail its two different stages: initialization and ILR transformation of a posteriori (ILRA) probabilities (see Figure 3). Previously to the initialization stage, the video is pre-processed keeping only frames that contain frontal faces.

A video is composed of a sequence of I shots ($S_1, \ldots, S_I$), where a shot is defined as a sequence of frames with a single intervener—see Figure 3. At the initialization stage, the system assigns an identity ID_1 to the first shot ($K = 1$). Next, shots are processed for novelty detection one by one, as a single intervener is assumed in each shot. Therefore, the system has to recognize whether a current shot intervener has been seen in previous shots (typical) or not (atypical). This stage is finished when an atypical shot is detected. Thus, the system knows two interveners ($K = 2$).

Once the system has registered two interveners, the next shots are processed to solve a new atypical detection problem. To this end, a novel modelling based on a posteriori probability of individuals is proposed. This modelling cannot be implemented in the previous stage because the system needs to have registered at least two interveners. If the new shot is typical, a K-label classification is used to recognize which one of the known interveners corresponds to the current shot. Otherwise, a new identity is assigned to the current shot. This procedure is repeated until no shots are left in the sequence $S_1, \ldots, S_I$. In the following subsections, the different stage details are described.

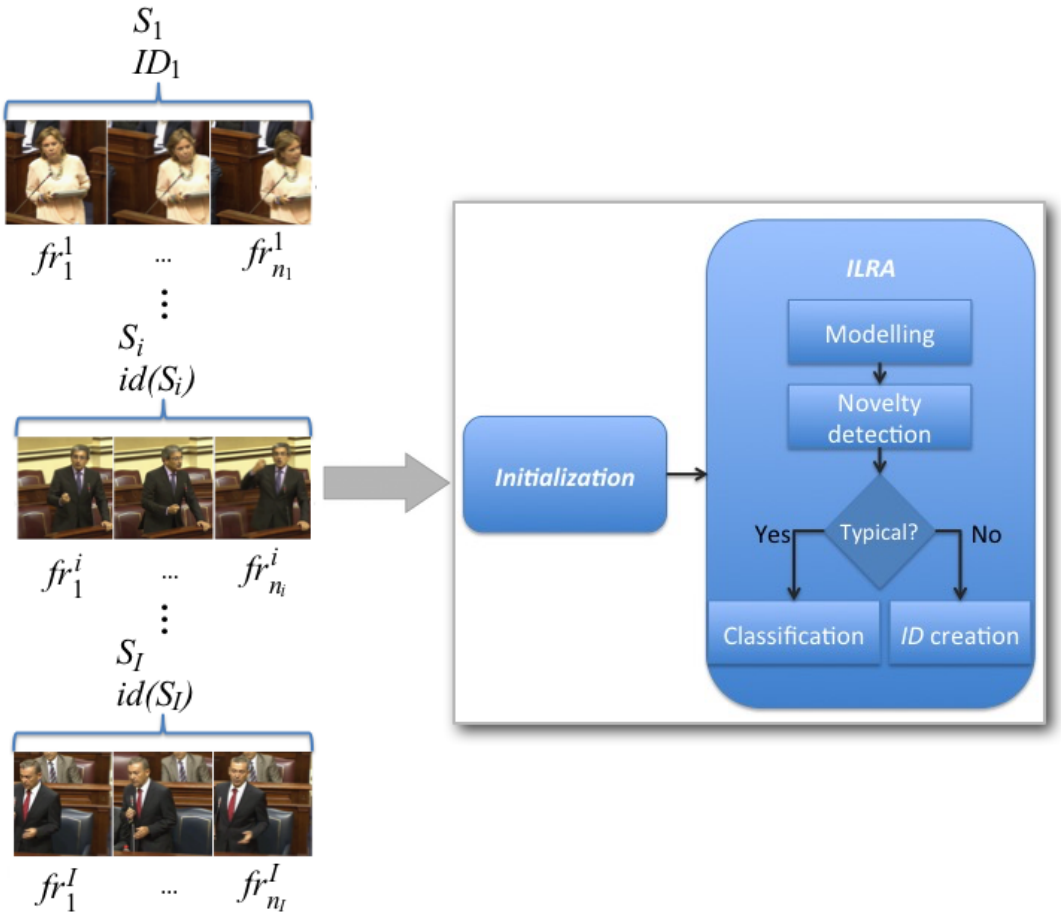

Figure 3. A video is divided into shots, S_i that are composed of frames, fr_i. Each shot contains a single intervener. Each shot is the input data of our proposed system. The system is mainly divided into two stages. The initialization stage is carried out without the modelling approach unlike the ILRA stage.

3.1. Video Pre-Processing

A video is a sequence of $S_1, \ldots, S_I$ shots and each shot S_i is composed of $fr_1^i, \ldots, fr_{n_i}^i$ frames with a detected face; in the case of multiple detected faces, the largest one is selected. Previously to the detection of the faces, each frame has been converted to grayscale because color information is not used by face descriptors [35]. For each shot S_i of the video, a matrix $X_i = [\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i]^\intercal$ is obtained, n_i being the number of frames of shot i-th. The detected face of each frame, fr_j^i, is represented by a descriptor computed on the face region as proposed by [36]. Thus, each row $\mathbf{x}_j^i$ of matrix X_i corresponds to the descriptor of dimension D, $\mathbf{x}_j^i = desc(fr_j^i) \in \mathbb{R}^D$ $(j = 1, \ldots, n_i, \quad i = 1, \ldots, I)$, resulting in a matrix of dimension $n_i \times D$:

$$X_i = \begin{pmatrix} x_{11}^i & \cdots & x_{1D}^i \\ \vdots & \ddots & \vdots \\ x_{n_i 1}^i & \cdots & x_{n_i D}^i \end{pmatrix}. \tag{1}$$

3.2. Initialization Stage

Firstly, the system assigns identity ID_1 to the first shot S_1, obtaining the extended matrix, including the label of the shot intervener:

$$X_1^e = \left(\begin{array}{ccc|c} x_{11}^1 & \dots & x_{1D}^1 & ID_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n_i1}^1 & \dots & x_{n_iD}^1 & ID_1 \end{array} \right). \quad (2)$$

From now on, we refer as identity to the label (ID_x) given to each registered individual. Later, the system has to determine the identity of the intervener in the following shots until the first atypical shot is found. This stage has similarities with a *One Vs. One (OVO) strategy* because, so far, the system knows just one intervener. Therefore, the procedure has to detect whether the intervener in the next shot is the same intervener ID_1 (typical) or if s/he is a different one (atypical). In terms of a classification problem, a one-class Support Vector Machine (SVM) [37] classifier is trained with the extended matrices $X_1^e, \dots, X_{i-1}^e$, and predictions are obtained for input matrix X_i. In this way, for each frame in X_i, a prediction in terms of typical/atypical is obtained. However, all frames do not necessarily have the same predicted labels; and it is reasonable to consider the whole shot S_i as typical ($id(S_i) = ID_1$) if most of the n_i frames in shot S_i are predicted as typical—otherwise, as atypical ($id(S_i) = ID_2$), increasing the number of interveners K. Thus, we have decided to use the Winner-Takes-All (WTA) principle to this purpose.

3.3. ILRA Stage

Once the system has registered at least two individuals ($K \geq 2$), it is necessary to determine whether the individual of the next shot S_i is registered or not. For this purpose, this stage comprises three main processes: modelling, novelty detection and, if the current shot is typical, classification. This stage has similarities with a *One Vs. All (OVA) strategy*. The available data at this stage are, on the one hand, the extended matrices $X_1^e, \dots, X_{i-1}^e$, which are the descriptors of each previous shot frames plus the label of their respective associated identities, and, on the other hand, the descriptors of the frames in shot S_i, i.e., X_i.

The aim of the modelling stage is to obtain the a posteriori probability, $p_{jk}^i = \text{Prob}(ID_k|\mathbf{x}_j^i)$, of each frame j in shot S_i belonging to each registered identity k. Thus, for shot i-th, a matrix P_i is computed:

$$P_i = \begin{pmatrix} p_{11}^i & \cdots & p_{1K}^i \\ \vdots & \ddots & \vdots \\ p_{n_i1}^i & \cdots & p_{n_iK}^i \end{pmatrix}, \quad (3)$$

where $\sum_{k=1}^K p_{jk}^i = 1$. On the one hand, for shots $S_1, \dots, S_{i-1}$, where an identity has been assigned, the estimation of the a posteriori probability is done using a leave-one-out strategy. Therefore, for each frame $fr_j \in \{S_1, \dots, S_{i-1}\}$, the a posteriori probabilities are computed using a Naïve Bayes classifier trained with all the frames minus frame fr_j, $\{S_1, \dots, S_{i-1}\} \setminus fr_j$. On the other hand, for each frame $fr_j \in S_i$, the a posteriori probabilities are computed using a Naïve Bayes classifier trained with all the frames of previous shots, $\{S_1, \dots, S_{i-1}\}$.

Once the a posteriori probabilities are computed, the second step of the modelling process is carried out. The ILR transformation is applied to $P_1, \dots, P_i$ matrices. This is a well-known transformation in the field of Compositional Data, which obtains a real coordinate representation, preserving the Aitchison metric in the original space of the a posteriori probabilities [38]. Formally defined as:

$$Z_i = ilr_v = clr(P_i)V, \quad (4)$$

where *clr* is the Centered Log Ratio (CLR) transformation and V is a matrix whose columns form an orthonormal basis of the CLR plane [38]. As a summary, each jth frame is normalized as follows:

$$\mathbf{x}_j^i \in \mathbb{R}^D \quad \Rightarrow \quad \mathbf{p}_j^i \in \mathbb{R}^K \quad \Rightarrow \quad \mathbf{z}_j^i \in \mathbb{R}^{K-1}. \tag{5}$$

Then, all transformed vectors are organized by rows in a matrix Z_i and this is the matrix that characterizes the shot S_i to determine the identity of the intervener. A similar transformation procedure is followed for all frames in shots $S_1, \ldots, S_{i-1}$, obtaining matrices $Z_1, \ldots, Z_{i-1}$. To determine the novelty in shot S_i, a one-class SVM classifier is trained with the extended matrices $Z_1^e, \ldots, Z_{i-1}^e$, and, similarly to the novelty detection approach of the initialization stage, predictions are obtained for input matrix Z_i. Again, WTA is used to determine if S_i is atypical or typical. In the first situation, $id(S_i) = ID_{K+1}$ is assigned and the number of identities known by the system increases. In the other case, when S_i is considered typical, a classifier is used to identify which of the known ones it belongs to. The classification module could be performed by any classifier, which could be trained with the extended matrices $Z_1^e, \ldots, Z_{i-1}^e$ to determine $id(S_i)$. Moreover, a WTA strategy is chosen to determine the identity that characterizes shot S_i.

3.4. ILRA Time Complexity

The time complexity for computing ILRA comprises the a posteriori probability computation, the ILR transformation, the novelty detection stage, and, in some cases, a classification. The time complexity of a posteriori probability computation (Naïve Bayes classifier) is $O(N \times D)$, where N is the number of frames and D is the number of attributes. The time complexity of ILR transformation is $O(N \times K)$, where K is the number of interveners. Finally, the time complexity of novelty detection and classification approach (SVM) is $O((K-1) \times N^2)$. Thus, the overall time complexity for each shot is $O(N(D + K(N+1) - N))$ in case of atypical detection; otherwise, $O(N(D + K(2N+1) - 2N))$.

4. Experimental Evaluation and Results

In order to evaluate our proposal, recordings from the Canary Islands Parliament (Santa Cruz de Tenerife, Canary Islands, Spain), which are publicly available in the Parliament web site [39], were processed on a workstation with an Intel Core i7-2600 at 3.40 GHz and 16 GB of RAM. The source code is available in github [40]. For the experiments, we chose six videos with different characteristics which are summarized in Table 1. The selected videos cover a wide range of interveners (5—21) and shots, so the influence of the number of interveners could be evaluated. Shots shorter than 30 s were skipped as they were considered not relevant for the diarization. In addition, frames without a detected face are avoided. For this aim, a face detector based on Histogram of Oriented Gradients features and SVM classifier is applied [41], where the face is normalized, establishing as a vertical symmetry axis through the center of the eyes position in the image, which are estimated by the face detector. Attending to the number of interveners, the videos could be classified as short with less than ten interverners (video identifiers 2771, 2918, 3015) and large with more than ten (video identifiers 2792, 2907, 3011).

Table 1. Description of the videos analyzed. The columns, "Shots" and "Frames" indicate the number of shots and frames.

Video Identifier	Interveners	Shots	Frames	Duration
2771	5	13	2440	0:33:23
2918	7	33	7142	1:21:23
3015	8	52	22,088	3:02:44
2792	11	55	13,956	1:48:00
2907	12	57	9542	2:20:20
3011	21	73	6525	2:01:42

First, a set of offline experiments were carried out to focus and to evaluate different situations involved in the proposed approach. The evaluation comprised three main experiments: (1) novelty detection in the initialization; (2) novelty detection and (3) classification in the ILRA stage. In this way, the performance of the different stages of our approach can be evaluated. With this objective, the shots of the same ID were reorganized to carry out the experiments properly, as shown in Figure 4.

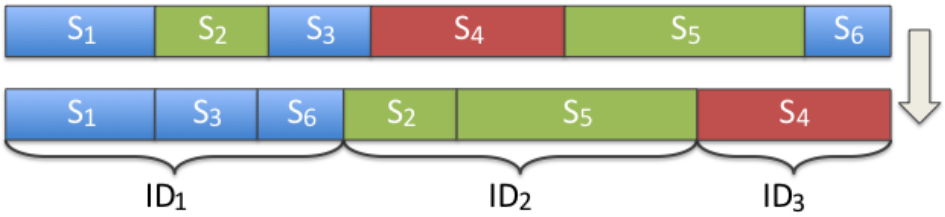

Figure 4. The original shots are reorganized with the purpose of grouping by ID for the novelty detection (initialization and ILRA stages) and classification (ILRA stage) experiments.

As a result of the rearrangement of the samples, the training sets are unbalanced because there are IDs more present than others; to avoid that, 500 frames were randomly chosen per identity. When the number of frames for an identity was lower, all shot frames were used. To validate the process, we carried out 100 repetitions.

The dimensionality of the individuals was reduced, $\mathbb{R}^{w \times h} \rightarrow \mathbb{R}^{D}$, as mentioned in Section 3. This reduction is based on applying a descriptor to the intervener face area ($w \times h$) where w and h represent the width and height, respectively. Two descriptor types have been evaluated, local descriptor and deep descriptor. The former type used a grid of 3×3 cells over an aligned image of 59×65. The following local descriptors were evaluated: Histogram of Oriented Gradients (HOG) [42], Local Binary Patterns (LBP) [43], LBP Uniform (LBPu2) [44], Neighborhood Intensity based LBP (NILBP) [45], and Weber Local Descriptor (WLD) [46] with a dimensionality of 81, 2304, 531, 531, and 2304, respectively. The latter type corresponds to a feature vector extracted from a deep network. In this case, a triplet network based on Inception Resnet backbone (Resnet$_T$) [47,48] is used. Mainly, a triplet network embedded the samples in a new feature space, where the samples that belong to the same identity are close and samples from different identities are far. Thus, three instances of Inception Resnet are used that share the same weight matrix. The embedded space is represented by the last fully connected layer, with a dimensionality of 128 in our experiments. Resnet$_T$ is used due to its excellent scores in different kinds of problems in recent years. The network was trained on Ms-celeb-1m [49] because the dataset consists of 1 million identities and we obtained a generalized model to extract the feature vectors from the faces. The network was initialized with the following parameters: mini-batches of size 90 along 500 epochs; the initial learning rate was 0.1, and this was decreased with a factor of 10 after every 100 epochs. Thus, the margin between positive and negative pairs (α) is set to 0.2. We set multiple descriptors due to the importance to evaluate the influence of different feature vectors for both stages of the algorithm.

Once the experimental setup is defined, it is necessary to adopt a metric. The accuracy (Acc.) is used with the purpose of evaluating the offline experiments, being formally defined as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{6}$$

where TP and FP are the number of true and false positives, respectively; TN and FN are the number of true and false negatives, respectively. Accuracy is used to measure typical and atypical detections. Instead of calculating the mean of typical and atypical values, the F measure is adopted to obtain only a measure providing a trade-off between both accuracies. Its formal definition is presented in the following equation:

$$F = 2\frac{precision \times recall}{precision + recall}, \tag{7}$$

where

$$precision = \frac{TP}{TP + FP} \tag{8}$$

and

$$recall = \frac{TP}{TP + FN}, \tag{9}$$

where *precision* is the fraction of relevant samples among the retrieved samples; moreover, *recall* is the fraction of relevant samples that have been retrieved over the total amount of relevant samples. Below, we present and discuss the results obtained in the experiments.

4.1. Evaluation of Novelty Detection in the Initialization Stage

The purpose of this first experiment is to evaluate the ability of the system to detect a novel identity when a single identity is known, i.e., $K = 1$. The typical or atypical detection was performed as follows: for each identity ID_k, we considered its corresponding samples as a test set, and, to conform the training set, we considered two different situations.

In the first case, the training set was composed of those samples with identity $ID_j \neq ID_k$. In such situation, the tested identity should be labelled as atypical (Figure 5a) and the number of different comparisons is $K^2 - K$. Note that, for each comparison, the detection of the individuals has to be atypical to be a success.

In the other case, the training set was composed by those samples with the same identity ID_k. To avoid having identical training and test sets, one third of the original samples of identity ID_k is used as a test set and the remaining two thirds as a training set. In this situation, the detection of the individuals has to be typical to be a success (Figure 5b). We performed this experiment for all K identities in the video.

Novelty detection in initialization stage columns of Table 2 summarize the results of the initialization stage experiments. It can be observed that, in all videos, the best F measure is obtained using Resnet$_T$, with an average value of 97.66%. In general, the atypical detection results are greater than or equal to 90% in 30 of 36 settings.

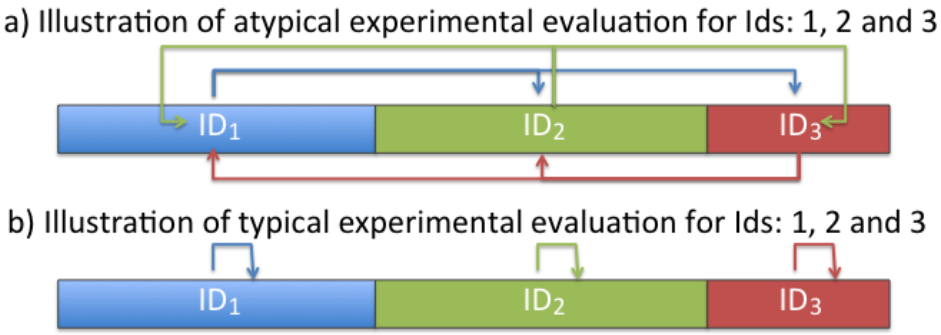

Figure 5. Initialization stage. (**a**) atypical experimental evaluation where each ID is matched individually with the remaining IDs (colored arrows); (**b**) typical experimental evaluation where each ID is matched with itself (colored arrows).

Table 2. Results of the offline experiments in terms of accuracy measure for novelty detection in initialization stage, novelty detection in the ILRA stage, and intervener classification in the ILRA stage. The results comprise the evaluation for different descriptors. The highest score is in bold.

Video Features		Descriptor	Novelty Detection in Initialization Stage			Novelty Detection in ILRA Stage			Intervener Classification in ILRA Stage	
Id	*K*		Typical	Atypical	*F*	Typical	Atypical	*F*	MAP Acc.	SVM Acc.
2771	5	HOG	100.0	90.00	94.74	80.00	60.00	**68.57**	96.52	96.92
		LBP	80.00	90.00	84.71	40.00	60.00	48.00	62.15	64.07
		LBPu2	80.00	90.00	84.71	80.00	60.00	**68.57**	96.72	96.89
		NILBP	100.0	90.00	94.74	20.00	80.00	32.00	72.45	81.13
		$Resnet_T$	100.0	100.0	**100.0**	60.00	80.00	**68.57**	**98.51**	**98.05**
		WLD	80.00	100.0	88.89	40.00	80.00	53.33	94.17	94.17
2918	7	HOG	85.71	52.38	65.02	100.0	85.71	**92.31**	92.15	91.12
		LBP	85.71	85.71	85.71	85.71	57.14	68.57	44.34	47.57
		LBPu2	85.71	90.48	88.03	28.57	100.0	44.44	**98.63**	**98.03**
		NILBP	100.0	85.71	92.31	100.0	0.00	0.00	41.05	50.20
		$Resnet_T$	100.0	100.0	**100.0**	29.57	86.71	42.86	97.49	97.16
		WLD	85.71	80.95	83.27	42.85	85.71	57.14	92.89	92.99
3015	8	HOG	100.0	85.71	92.31	100.0	100.0	**100.0**	95.30	94.24
		LBP	87.50	100.0	93.33	25.00	37.50	30.00	54.92	56.47
		LBPu2	87.50	100.0	93.33	50.00	100.0	66.67	**97.93**	**98.01**
		NILBP	100.0	96.43	98.18	87.50	12.50	21.88	63.00	67.58
		$Resnet_T$	100.0	100.0	**100.0**	27.27	100.0	42.85	97.78	97.52
		WLD	100.0	96.43	98.18	87.50	0.00	0.00	63.57	68.68
2792	11	HOG	81.82	89.09	85.30	90.91	100.0	**95.24**	92.60	91.26
		LBP	90.91	98.18	94.41	18.18	72.72	29.09	51.84	53.42
		LBPu2	81.82	96.36	88.50	45.45	90.91	60.61	97.12	96.84
		NILBP	81.82	96.36	88.50	100.0	0.00	0.00	45.16	55.76
		$Resnet_T$	100.0	100.0	**100.0**	36.00	100.0	52.94	**97.94**	**97.83**
		WLD	81.82	98.18	89.26	9.09	90.91	16.53	85.13	85.31
2907	12	HOG	75.00	90.91	82.19	41.66	83.33	55.56	96.42	96.02
		LBP	75.00	96.97	84.58	66.67	75.00	70.59	64.30	64.47
		LBPu2	66.67	98.48	79.51	25.00	83.33	38.46	98.11	98.19
		NILBP	83.33	100.0	90.91	50.00	25.00	33.33	76.19	79.07
		$Resnet_T$	100.0	100.0	**100.0**	41.67	100.0	58.83	**98.90**	**98.98**
		WLD	58.33	100.0	73.68	75.00	91.67	**82.50**	92.25	91.87
3011	21	HOG	52.38	94.76	67.47	47.62	71.43	57.14	41.29	**96.55**
		LBP	42.86	97.14	59.48	61.90	61.90	61.90	20.18	49.65
		LBPu2	42.86	96.19	59.30	42.86	76.19	54.86	40.85	94.92
		NILBP	57.14	96.19	71.69	61.90	52.38	56.75	36.98	84.26
		$Resnet_T$	76.19	98.57	**85.95**	23.81	90.48	37.70	**41.49**	94.64
		WLD	28.57	99.05	44.35	85.71	80.95	**83.27**	36.70	86.09
Mean		HOG	82.49	83.81	81.17	76.70	83.41	**78.14**	85.71	94.35
		LBP	77.00	94.67	83.70	49.58	60.71	51.36	49.62	55.94
		LBPu2	74.09	95.25	82.23	45.31	85.07	55.60	88.23	97.15
		NILBP	87.05	94.12	89.39	69.90	28.31	23.99	55.81	69.67
		$Resnet_T$	96.03	99.76	**97.66**	36.22	92.70	50.62	**88.69**	**97.36**
		WLD	63.51	95.77	79.60	56.69	71.54	48.80	77.45	86.52

4.2. Evaluation of Novelty Detection in the ILRA Stage

The experiments related to the ILRA stage for offline scope are motivated by the need to evaluate the capacity of the approach to detect the novel identity of a new shot when several identities are known. Therefore, two evaluations are considered for each identity: atypical and typical. The former comprises all ID_k identity samples in the test set, while the rest of identity samples, $ID_{j \neq k}$, are used for training (Figure 6a). This experiment is carried out to evidence the approach behaviour for atypical identity detection, as the tested identity ID_k should be labelled as atypical. The latter comprises all identities in both training and test set, splitting randomly and balanced their respective samples, using one third for testing and the rest for training (Figure 6b). This experiment is carried out to evidence the approach behaviour for typical identity detection, as the tested identity ID_k should be labelled as typical.

Novelty detection in the ILRA stage columns of Table 2 allude that the descriptor with the highest *F* accuracy is HOG, reporting 78.14%. It is also observed that, when the number of interveners is low, the best descriptor is HOG and, over a large number of interveners, WLD behaves apparently better than the remaining descriptors.

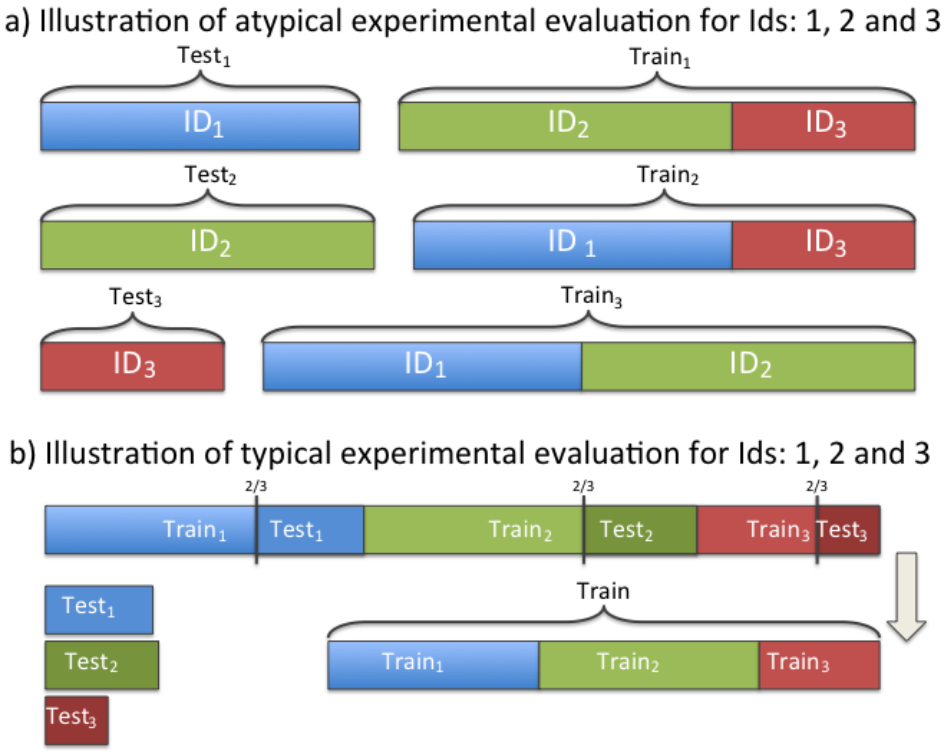

Figure 6. ILRA stage. (**a**) atypical experimental evaluation where each ID is matched with the remaining IDs; (**b**) typical experimental evaluation where each set is split into one third to test and the rest for training.

4.3. Evaluation of Intervener Classification in the ILRA Stage

The purpose of this experiment is to evaluate the capacity of the approach to correctly assign the identity of a new intervener shot when multiple identities are known. That means, when the identity of the new shot ($id(S_i)$) is present among the known identities, this intervener has been considered as typical in the ILRA stage, and the approach should match it to whom ID belongs to. Two classifiers are considered: the Maximum A Posteriori (MAP) probability extracted from the samples (see Figure 7a); and an SVM classifier to continue using the same typology of classifiers that we used throughout this proposal (see Figure 7b). In the case of the SVM, a Radial Basis Function (RBF) kernel is selected with main parameters $\nu = 0.1$, $\gamma = 0.1$ and $C = 1$. A repeated holdout validation is carried out using 100 repetitions with re-sampling of the individuals, one third of the samples to test and the remaining to train.

The results are summarized in intervener classification in the ILRA stage columns of Table 2. Among the six descriptors, Resnet$_T$ yields the best accuracy in seven of the twelve experiments, giving an average value for the MAP and SVM classifiers of 88.69% and 97.36%, respectively.

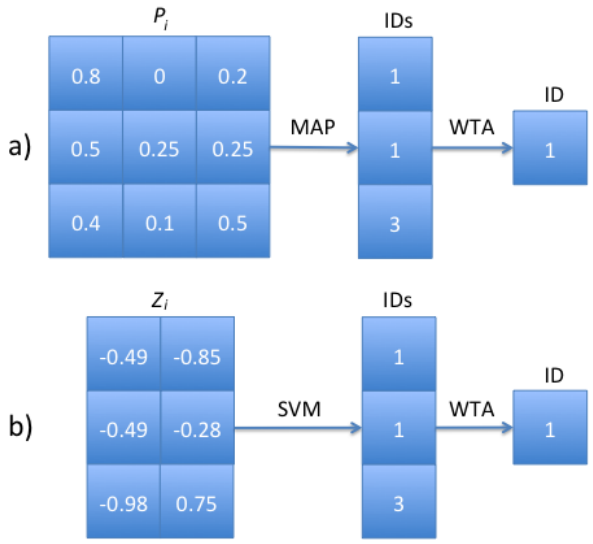

Figure 7. Procedure to determine $id(S_i)$. (**a**) represents the process to extract the ID of the intervener, using the Maximum A Posteriori (MAP) probability to each sample; (**b**) shows the use of an SVM to obtain the intervener ID.

4.4. Evaluation of the Proposed Online System

After evaluating the different offline stages, we carried out an online experiment. The number of frames per shot has been modified compared to the offline configuration. In addition, 200 frames per shot were used because the experiment comprises a larger number of shots, some of them containing a reduced number of frames. This situation brought about unbalanced shots that affect the performance of the algorithm. Given the best performance provided by the SVM classifiers in previous offline experiments, SVM is adopted to identify the interveners in the case of typical individuals.

To evaluate the online system, we adopted, from [50], True Re-identification Rate (TRR) and True Distinction Rate (TDR) measures. TRR evaluates how good the method is to re-identify interveners, while TDR evaluates how good the method is to distinguish among the interveners. Both measures are formulated as follows:

$$TRR = \frac{\mathrm{tr}(score)}{N}, \tag{10}$$

$$TDR = 1 - \frac{(score\,\mathbf{1}_N)^T\,\mathbf{1}_N - \mathrm{tr}(score)}{N(N-1)}, \tag{11}$$

where $\mathbf{1}_N$ is a vector of dimension N with all the elements to one; and $\mathrm{tr}(score)$ is the trace of $score$ that is a $N \times N$ matrix that has the result of comparing each proposed intervener shot identity with respect to all proposed intervener shot identities, 1 is assigned to equal identities and 0 to different ones. Thus, 1 in the diagonal elements and 0 in off-diagonal elements compose a perfect score. To obtain a single measure, the F measure is adopted, relating TRR (considered as recall) and TDR (considered as precision).

The last evaluated experiment is the online process where real online video processing is comprised, evaluating the same descriptor for each stage of the algorithm. The results of the experiments are summarized in Table 3. In most of the processed videos, a descriptor beats the others, but there is no common behaviour across the entire video collection. In this case, the use of the descriptor depends on the video, not on the number of interveners. On the one hand, we would like to highlight the F measure obtained in video 3011, 88.25%, covering a population of 21 interveners and two hours of recording in an open world problem that means a real complex problem. On the other hand, the result achieved for recording 2907 is interesting because it brings forward a deficiency in traditional feature vectors, aroused by an occlusion issue due to most of the interveners putting

the glasses on or taking them off during the intervention. In this situation, Resnet$_T$ improves at least 44.69% compared to the other descriptors, reaching 76.51% in recording 2907.

Table 3. Results of the online experiments in terms of TDR, TDR, and *F*. The highest *F* is in bold.

Video ID	Descriptor	TRR	TDR	*F*
2771	HOG	83.33	74.07	**78.43**
	LBP	25.00	94.44	39.53
	LBPu2	16.67	96.30	28.42
	NILBP	16.67	88.89	28.07
	Resnet$_T$	58.33	90.74	71.01
	WLD	8.33	96.30	15.34
2918	HOG	38.50	99.08	55.45
	LBP	95.72	11.63	20.75
	LBPu2	40.11	83.28	54.14
	NILBP	56.68	76.89	65.26
	Resnet$_T$	59.15	97.65	**73.68**
	WLD	31.55	93.76	47.21
3015	HOG	71.05	95.96	**81.65**
	LBP	29.47	75.76	42.44
	LBPu2	34.21	96.80	50.55
	NILBP	40.53	88.89	55.67
	Resnet$_T$	56.83	99.58	72.37
	WLD	36.84	96.46	53.32
2792	HOG	71.83	69.18	70.48
	LBP	28.17	85.18	42.34
	LBPu2	70.42	94.12	**80.56**
	NILBP	59.15	83.76	69.34
	Resnet$_T$	47.59	97.69	64.00
	WLD	54.93	82.82	66.05
2907	HOG	52.27	48.49	50.31
	LBP	15.91	87.09	26.90
	LBPu2	31.82	95.12	47.69
	NILBP	27.27	92.54	42.13
	Resnet$_T$	65.79	91.41	**76.51**
	WLD	40.91	74.75	52.88
3011	HOG	82.08	95.43	**88.25**
	LBP	55.66	91.34	69.17
	LBPu2	49.06	99.69	65.75
	NILBP	73.58	65.98	69.58
	Resnet$_T$	54.68	97.85	70.15
	WLD	71.70	89.29	79.53
Mean	HOG	66.51	80.37	70.76
	LBP	41.66	74.24	40.19
	LBPu2	40.38	94.22	54.52
	NILBP	45.65	82.83	55.01
	Resnet$_T$	57.06	95.82	**71.29**
	WLD	40.71	88.90	52.39

Furthermore, our system is compared with our previous work [51]—as far as we know, the only existing approach in this scenario, i.e., face-based intervener re-identification in open-world parliamentary debates sessions. Additionally, face recognition approaches focusing on the closed world are used to extend the comparative of the proposed ILRA approach. In particular, HOG, LBP LBPu2, NILBP, WLD and Resnet$_T$ are used as feature vectors. In order to detect atypical samples, we use a threshold with a value of 0.5, an atypical sample being the corresponding one with a value larger than the threshold. In the case that the sample is typical, a distance vector is calculated from the samples previously analyzed with respect to the current sample. The identity with the minimum distance will represent the current sample.

Our method obtains in most of the experiments the best *F* measure for the different videos, compared with the above methods. These results are summarized in Table 4. On the one hand, the highest increase in performance is video 3015 where there is an improvement of 63.80% with

respect to our previous work, ILRA being widely superior to traditional methods of face recognition. On the other hand, the recent technique, Resnet$_T$, achieves a significant increase in results compared to the techniques mentioned above. However, it does beat the proposed method, reaching an average difference of 1.12% for the analyzed videos.

Table 4. Results of the online experiments compared with other approaches in terms of TDR, TDR, and *F*. The highest *F* is in bold.

Video ID	Descriptor	TRR	TDR	*F*
2771	[42]	58.33	61.11	59.69
	[43]	41.67	70.37	52.34
	[44]	41.67	70.37	52.34
	[45]	33.33	79.63	46.99
	[48]	79.33	64.52	**71.16**
	[46]	41.67	70.37	52.34
	[51]	53.91	75.36	62.86
	Ours (Resnet$_T$)	58.33	90.74	71.01
2918	[42]	49.41	79.85	61.05
	[43]	42.35	95.82	58.74
	[44]	48.82	97.18	64.99
	[45]	57.65	85.07	68.72
	[48]	96.00	44.71	61.01
	[46]	43.53	94.78	59.66
	[51]	50.59	75.16	60.47
	Ours (Resnet$_T$)	59.15	97.65	**73.68**
3015	[42]	85.81	12.56	21.91
	[43]	43.02	58.85	49.71
	[44]	45.49	57.84	50.93
	[45]	48.18	51.49	49.78
	[48]	80.17	47.98	60.03
	[46]	69.52	38.58	49.62
	[51]	85.93	9.96	17.85
	Ours (Resnet$_T$)	56.83	99.58	**72.37**
2792	[42]	20.33	96.28	33.57
	[43]	31.17	94.87	46.92
	[44]	31.05	95.64	46.88
	[45]	48.18	51.49	49.78
	[48]	89.05	58.17	**70.37**
	[46]	31.17	91.56	57.27
	[51]	23.85	93.58	38.01
	Ours (Resnet$_T$)	47.59	97.69	64.00
2907	[42]	23.49	88.23	37.10
	[43]	33.73	87.79	48.74
	[44]	28.31	89.67	43.03
	[45]	26.20	88.58	40.44
	[48]	91.26	88.68	**89.95**
	[46]	34.94	82.92	49.16
	[51]	21.99	84.91	34.93
	Ours (Resnet$_T$)	65.79	91.41	76.51
3011	[42]	57.61	77.38	66.05
	[43]	51.78	70.62	59.75
	[44]	53.41	70.55	60.80
	[45]	58.38	73.59	65.11
	[48]	66.45	70.61	68.47
	[46]	50.76	78.68	61.71
	[51]	58.12	79.36	67.10
	Ours (Resnet$_T$)	54.68	97.85	**70.15**
Mean	[42]	49.16	69.24	46.56
	[43]	40.62	79.72	52.70
	[44]	41.46	80.21	53.16
	[45]	45.32	71.64	53.47
	[48]	83.71	62.44	70.17
	[46]	45.27	76.15	54.96
	[51]	49.07	69.72	46.87
	Ours (Resnet$_T$)	57.06	95.82	**71.29**

5. Discussion

In this paper, we analyzed the ILRA approach in offline and online contexts. On the one hand, offline experiments were carried out to evaluate the method in a controlled scenario. In this way, we could analyze each stage of the approach. On the other hand, online experiments allowed us to test the method in real conditions, where the system starts without any registered person.

A feature of the proposed method is the need of an initialization stage because it is not possible to calculate the ILRA with less than three registered identities (Section 3). The performance in detecting the second identity to start the ILRA process will affect the rest of the system. For this reason, we evaluate the initialization stage in an offline context (Section 4.1), where we have obtained that the Resnet$_T$ descriptor achieves a better score than local descriptors.

The modelling process is evaluated in the offline ILRA stage, which is split into two processes, novelty detection (Section 4.2) and classification of identities (Section 4.3). Firstly, the Resnet$_T$ descriptor is not the best descriptor for novelty detection in the ILRA stage. In this instance, a local descriptor, HOG, obtains the best average performance. Secondly, the Resnet$_T$ descriptor is better than local descriptors for classifying, as much as using a MAP as an SVM classifier.

This disaggregated analysis of the offline experiments shows that there is not a common best descriptor for each stage. This issue is translated into the online experiments, where a decreasing of the average score for each descriptor is obtained. This is due to having failures in the recognition at the first stage, which generates more false positive identifications. A way to alleviate this issue is to choose a specific descriptor for each stage; as shown in Section 4, there is no single descriptor that stands out in all stages. The selection of a single descriptor for any stage affects the system performance, making it less robust. Certainly, the system is simpler, but the use of a single descriptor in all system stages seems not to be the ideal approach. A further observation suggests that Resnet$_T$ is well suited to detect outliers in a one-class problem, HOG fits to novel detection in the ILRA stage and Resnet$_T$ performs better to classify in the ILRA stage.

The existence of short videos with very few detected faces favors SVM over MAP as can be observed in video 3011 (intervener classification in the ILRA stage of Table 2). This is due to the estimation of the Naïve Bayes parameters used in the MAP, as the average that is more affected by unbalanced classes. Some authors have verified that SVM performs better than Naïve Bayes dealing with unbalanced classes [52–54]. Moreover, the feature vector transformation using ILR alleviates the unbalanced problem as it is suggested in [55].

6. Conclusions

A feasible face-based intervener re-identification to open world solutions has been presented in order to be applied to diarization problems. We have evaluated the approach in parliamentary debate sessions, a challenging scenario, where people vary their pose and appearance, and do not necessarily appear while speaking.

In this scenario, the novelty intervener detection is relevant, as those identities must be properly registered. If novelty detection fails to detect a new intervener, s/he will be incorrectly assigned to a previously detected intervener. On the contrary, if a previously detected intervener is considered as a new one, the number of interveners will be erroneously increased. We have used and evaluated descriptors for identity registration. In the one-class problem, Resnet$_T$ has shown a good performance in novelty detection. The use of HOG yields the highest accuracy for a low number of interveners. However, when the number of interveners is larger, WLD achieves the best results. The best configuration is the Resnet$_T$ with an SVM classifier in the classification stage.

Our proposed system experiments have exhibited good results with an average F measure of 71.29% for the best descriptor for each video. In addition, we have compared the ILRA with respect to different techniques used in face recognition in a closed world, exhibiting an increase of 1.6% with respect to the deep descriptor extracted from a triplet network based on an Inception Resnet backbone. In the offline experiments, the results for the novelty detection in the initialization stage reach an average 97.66% accuracy for the Resnet_T descriptor. In the ILRA stage, for the novelty detection, an average accuracy of 78.14% is obtained for the HOG descriptor. In the ILRA stage for the intervener classification experiments, the average accuracy is 97.36% using Resnet_T descriptor with our method.

As future work, we plan to apply this approach using only audio features and the fusion of audio and video features. In this way, we could determine the influence of the audio over the image representation and verify if we obtain a better feature vector. Moreover, we intend to use deep learning techniques in order to replace the one-class SVM in the novelty detection module.

Author Contributions: Data curation, P.A.M.-R., J.L.-N and M.C.-S.; formal analysis, P.A.M.-R., I.I., C.A.; funding acquisition, B.S., I.I., J.L.-N, M.C.-S. and C.A.; investigation, P.A.M.-R., B.S. and I.I.; methodology, P.A.M.-R., B.S., I.I. and C.A.; project administration, B.S. and I.I.; resources, J.L.-N and M.C.-S.; software, P.A.M.-R., J.L.-N. and I.I.; supervision, J.L.-N, M.C.-S., B.S., I.I. and C.A.; validation, P.A.M.-R., B.S., I.I. and C.A.; visualization, P.A.M.-R., B.S., I.I. and C.A.; writing - original draft, P.A.M.-R., J.L.-N and M.C.-S.; writing - review and editing, P.A.M.-R., J.L.-N, M.C.-S., B.S., I.I. and C.A.

Funding: This research was funded by the Spanish Ministry of Economy and Competitiveness, Spain RTI2018-093337-B-I00, by AGAUR 2014SGR464, by the Office of Economy, Industry, Commerce and Knowledge of the Canary Islands Government (CEI2018-4), and the Computer Science Department at the Universidad de Las Palmas de Gran Canaria.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gheissari, N.; Sebastian, T.B.; Hartley, R. Person Reidentification Using Spatiotemporal Appearance. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1528–1535.
2. Vezzani, R.; Baltieri, D.; Cucchiara, R. People reidentification in surveillance and forensics: A survey. *ACM Comput. Surv.* **2013**, *46*, 29:1–29:37. [CrossRef]
3. Prosser, B.; Zheng, W.S.; Gong, S.; Xiang, T. Person Re-Identification by Support Vector Ranking. In Proceedings of the British Machine Vision Conference (BMVC), Aberystwyth, UK, 31 August–3 September 2010; pp. 21.1–21.11. [CrossRef]
4. Roth, P.M.; Hirzer, M.; Köstinger, M.; Beleznai, C.; Bischof, H. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*; Gong, S., Cristani, M., Yan, S., Loy, C.C., Eds.; Springer: London, UK, 2014; pp. 247–267.
5. Bedagkar-Gala, A.; Shah, S.K. A survey of approaches and trends in person re-identification. *Image Vis. Comput.* **2014**, *32*, 270–286. [CrossRef]
6. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
7. Markou, M.; Singh, S. Novelty detection: a review-part 1: Statistical approaches. *Signal Process.* **2003**, *83*, 2481–2497. [CrossRef]
8. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58. [CrossRef]
9. Pimentel, M.A.; Clifton, D.A.; Clifton, L.; Tarassenko, L. A review of novelty detection. *Signal Process.* **2014**, *99*, 215–249. [CrossRef]
10. Anguera, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; Vinyals, O. Speaker diarization: A review of recent research. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 356–370. [CrossRef]

11. El Khoury, E.; Sénac, C.; Joly, P. Audiovisual diarization of people in video content. *Multimed. Tools Appl.* **2014**, *68*, 747–775. [CrossRef]
12. Liu, K.; Chen, J.H.; Chang, K.M. A Study of Facial Features of American and Japanese Cartoon Characters. *Symmetry* **2019**, *11*, 664. [CrossRef]
13. Kamachi, M.G.; Chiba, T.; Kurosumi, M.; Mizukoshi, K. Perception of Human Age from Faces: Symmetric Versus Asymmetric Movement. *Symmetry* **2019**, *11*, 650. [CrossRef]
14. Bredin, H.; Gelly, G. Improving Speaker Diarization of TV Series Using Talking-Face Detection and Clustering. In Proceedings of the ACM International Conference on Multimedia (ACMMM), Amsterdam, The Netherlands, 15–19 October 2016; pp. 157–161.
15. Gebru, I.; Ba, S.; Li, X.; Horaud, R. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, in press. [CrossRef] [PubMed]
16. Le, N.; Wu, D.; Meignier, S.; Odobez, J.M. EUMSSI Team at the Mediaeval Person Discovery Challenge. In Proceedings of the Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15 September 2015.
17. Friedland, G.; Hung, H.; Yeo, C. Multi-Modal Speaker Diarization of Real-World Meetings Using Compressed-Domain Video Features. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 19–24 April 2009; pp. 4069–4072.
18. Bazzani, L.; Cristani, M.; Murino, V. Symmetry driven accumulation of local features for human characterization and re-identification. *Comput. Vis. Image Underst.* **2013**, *117*, 130–144. [CrossRef]
19. Tao, D.; Guo, Y.; Song, M.; Li, Y.; Yu, Z.; Tang, Y.Y. Person re-identification by dual-regularized kiss metric learning. *IEEE Trans. Image Process.* **2016**, *25*, 2726–2738. [CrossRef]
20. Yu, H.X.; Wu, A.; Zheng, W.S. Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
21. Ustinova, E.; Ganin, Y.; Lempitsky, V. Multi-Region Bilinear Convolutional Neural Networks for Person Re-Identification. In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
22. Zheng, Z.; Zheng, L.; Yang, Y. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 13. [CrossRef]
23. Yong, S.P.; Deng, J.D.; Purvis, M.K. Novelty detection in wildlife scenes through semantic context modelling. *Pattern Recognit.* **2012**, *45*, 3439–3450. [CrossRef]
24. Clifton, D.A.; Clifton, L.; Hugueny, S.; Wong, D.; Tarassenko, L. An extreme function theory for novelty detection. *IEEE J. Sel. Top. Signal Process.* **2013**, *7*, 28–37. [CrossRef]
25. Irigoien, I.; Arenas, C. INCA: New statistic for estimating the number of clusters and identifying atypical units. *Stat. Med.* **2008**, *27*, 2948–2973. [CrossRef] [PubMed]
26. Boucenna, S.; Cohen, D.; Meltzoff, A.N.; Gaussier, P.; Chetouani, M. Robots learn to recognize individuals from imitative encounters with people and avatars. *Sci. Rep.* **2016**, *6*, in press. [CrossRef]
27. Markov, K.; Nakamura, S. Improved Novelty Detection for Online GMM Based Speaker Diarization. In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), Brisbane, Australia, 22–26 September 2008; pp. 363–366.
28. Zheng, W.S.; Gong, S.; Xiang, T. Transfer Re-Identification: From Person to Set-Based Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2650–2657.
29. Chan-Lang, S.; Pham, Q.C.; Achard, C. Closed and Open-World Person Re-Identification and Verification. In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 29 November–1 December 2017; pp. 1–8.
30. Zhu, X.; Wu, B.; Huang, D.; Zheng, W.S. Fast open-world person re-identification. *IEEE Trans. Image Process.* **2018**, *27*, 2286–2300. [CrossRef]
31. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-dissimilarity for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 994–1003.

32. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
33. Li, X.; Wu, A.; Zheng, W.S. Adversarial Open-World Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 280–296.
34. Khan, S.S.; Madden, M.G. One-class classification: Taxonomy of study and review of techniques. *Knowl. Eng. Rev.* **2014**, *29*, 345–374. [CrossRef]
35. Castrillón-Santana, M.; Lorenzo-Navarro, J.; Ramón-Balmaseda, E. Descriptors and regions of interest fusion for in- and cross-database gender classification in the wild. *Image Vis. Comput.* **2017**, *57*, 15–24. [CrossRef]
36. Castrillón-Santana, M.; Lorenzo-Navarro, J.; Travieso-González, C.M.; Freire-Obregón, D.; Alonso-Hernández, J.B. Evaluation of local descriptors and CNNs for non-adult detection in visual content. *Pattern Recognit. Lett.* **2017**, in press. [CrossRef]
37. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
38. Egozcue, J.J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **2003**, *35*, 279–300. [CrossRef]
39. de Canarias, P. Web Site of Canary Islands Parliament. 2018. Available online: http://www.parcan.es/ (accessed on 7 June 2018).
40. Marín-Reyes, P.A. ILRA Source Code. 2019. Available online: https://github.com/foumacray/ILRA (accessed on 12 August 2019).
41. Kazemi, V.; Sullivan, J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874.
42. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893. [CrossRef]
43. Ojala, T.; Pietikainen, M.; Harwood, D. Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions. In Proceedings of the International Conference on Pattern Recognition (ICPR), Jerusalem, Israel, 9–13 October 1994; Volume 1, pp. 582–585. [CrossRef]
44. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
45. Liu, L.; Zhao, L.; Long, Y.; Kuang, G.; Fieguth, P. Extended local binary patterns for texture classification. *Image Vis. Comput.* **2012**, *30*, 86–99. [CrossRef]
46. Chen, J.; Shan, S.; He, C.; Zhao, G.; Pietikainen, M.; Chen, X.; Gao, W. WLD: A robust local image descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1705–1720. [CrossRef]
47. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.[CrossRef]
48. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-resnet and the Impact of Residual Connections on Learning. In Proceedings of the Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017.
49. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. Ms-celeb-1m: A Dataset and Benchmark for Large-Scale Face Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 87–102.
50. Cong, D.N.T.; Khoudour, L.; Achard, C.; Meurie, C.; Lezoray, O. People re-identification by spectral classification of silhouettes. *Signal Process.* **2010**, *90*, 2362–2374. [CrossRef]
51. Sánchez-Nielsen, E.; Chávez-Gutiérrez, F.; Lorenzo-Navarro, J.; Castrillón-Santana, M. A multimedia system to produce and deliver video fragments on demand on parliamentary websites. *Multimed. Tools Appl.* **2016**, *76*, 6281–6307. [CrossRef]
52. Liu, Y.; Loh, H.T.; Sun, A. Imbalanced text classification: A term weighting approach. *Expert Syst. Appl.* **2009**, *36*, 690–701. [CrossRef]

53. Zhang, S.; Sadaoui, S.; Mouhoub, M. An empirical analysis of imbalanced data classification. *Comput. Inf. Sci.* **2015**, *8*, 151. [CrossRef]
54. Zhuang, L.; Dai, H. Parameter Estimation of One-Class SVM on Imbalance Text Classification. In Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence, Québec City, QC, Canada, 7–9 June 2006; pp. 538–549.
55. de Deus, J.L.; Neves, J.C.L.; Corrêa, M.C.d.M.; Parent, S.É.; Natale, W.; Parent, L.E. Balance design for robust foliar nutrient diagnosis of "Prata" banana (Musa spp.). *Sci. Rep.* **2018**, *8*, 15040. [CrossRef] [PubMed]