applied sciences

MDPI

Article

Sample Reduction-Based Pairwise Linear Regression Classification for IoT Monitoring Systems

Xizhan Gao, Wei Hu, Yu Chu and Sijie Niu *

School of Information Science and Engineering, University of Jinan, Jinan 250022, China
* Correspondence: ise_niusj@ujn.edu.cn

Abstract: At present, the development of the Internet of Things (IoT) has become a significant symbol of the information age. As an important research branch of it, IoT-based video monitoring systems have achieved rapid developments in recent years. However, the mode of front-end data collection, back-end data storage and analysis adopted by traditional monitoring systems cannot meet the requirements of real-time security. The currently widely used edge computing-based monitoring system can effectively solve the above problems, but it has high requirements for the intelligent algorithms that will be deployed at the edge end (front-end). To meet the requirements, that is, to obtain a lightweight, fast and accurate video face-recognition method, this paper proposes a novel, set-based, video face-recognition framework, called sample reduction-based pairwise linear regression classification (SRbPLRC), which contains divide SRbPLRC (DSRbPLRC), anchor point SRbPLRC (APSRbPLRC), and attention anchor point SRbPLRC (AAPSRbPLRC) methods. Extensive experiments on some popular video face-recognition databases demonstrate that the performance of proposed algorithms is better than that of several state-of-the-art classifiers. Therefore, our proposed methods can effectively meet the real-time and security requirements of IoT monitoring systems.

Keywords: IoT monitoring system; video face recognition; recognition performance optimization; attention mechanism; anchor point; large-size video

check for updates

Citation: Gao, X.; Hu, W.; Chu, Y.; Niu, S. Sample Reduction-Based Pairwise Linear Regression Classification for IoT Monitoring Systems. *Appl. Sci.* **2023**, *13*, 4209. https://doi.org/10.3390/app13074209

Academic Editors: Hejun Wu and Shigeng Zhang

Received: 22 February 2023
Revised: 12 March 2023
Accepted: 24 March 2023
Published: 26 March 2023

1. Introduction

At present, the development of the Internet of Things [1,2] has become a significant symbol of the information age, and the video monitoring systems are an important basic aspect within the IoT field, which can be used in many real word scenarios, such as intelligent elderly care monitoring systems, intelligent access control systems, and intelligent anti-theft systems.

Traditional monitoring systems usually consist of two parts, the front-end cameras, and the back-end server, and their basic processes are usually as follows: the front-end cameras are first used to collect the monitoring data, then the collected data will be uploaded to the back-end server for storage and processing; finally, the useful information is obtained through human inspection or some intelligent algorithms. However, such manner has the following disadvantages: (1) uploading the collected data to the server will cost a lot of time, resulting in a failure to meet real-time requirement; (2) compared with PCs and other devices, IoT devices are generally weak in performance and more vulnerable to attacks, resulting in their inability to ensure the security of data during transmission; (3) the collected video data usually need to be analyzed and recognized using intelligent algorithms. However, the existing video face-recognition methods still have some defects. For the above problems (1) and (2), the currently mature solution is to introduce edge computing [3] and migrate the data analysis process to the edge end. Specifically, some intelligent algorithms to each edge end (for example, deploy them to the surveillance cameras) are firstly deployed, in order to process the collected data in real time; secondly, the recognition or analysis results processed by the algorithm are returned to the back-end

(server) for storage, or even alarming. In such a manner, we can not only effectively reduce the amount of data transmission and meet the real-time requirements, but also gain higher security, because the transmission data no longer comprise the original data but the analysis results. However, this scheme has high requirements for the intelligent algorithms to be deployed, that is, the intelligent algorithms need to have fewer parameters (because the memory of the edge device is limited), faster computing speed and higher recognition accuracy. In view of this practical demand, this paper will focus on the research of video face-recognition methods applicable to IoT monitoring systems.

So far, face recognition is already widely used in everyday life, including sign-in systems, fugitive tracking systems, etc. Many related algorithms have been developed [4,5], such as the principal component analysis (PCA) algorithm [6], linear regression classification (LRC) algorithm [7], support vector machine (SVM) algorithm [8], linear discriminant analysis (LDA) algorithm [9], k-nearest neighbor (K-NN) approach [10], canonical correlation analysis (CCA) method [11], and the sparse representation-based classifications (SRC) algorithm [12]. All of these single image-based image recognition algorithms have achieved satisfactory performances; however, with the development of imaging technology, a great number of videos is being produced everyday, and how to recognize video [13] data, i.e., how to measure the distance between videos, is still a challenging problem. Many researchers have pointed out that temporal information is not important for face recognition; hence, the term video face recognition in this study refers to set-based video face recognition, i.e., where temporal information is not considered.

When compared to a single image, videos and image sets contain useful information, such as varying expressions, varying poses, and varying illumination conditions, that describe the objects within in the images. Hence, it is very important to study the classification problem from the perspective of image sets, i.e., the image set classification task. The purpose of image set classification (or set-based video recognition) is to assign labels to the probe set (or video) by measuring the similarities between the gallery videos (or sets) and the probe videos (or sets). When compared with the single image-based recognition task, the set-based video recognition task can directly calculate the labels for whole videos, without classifying each frame of the videos separately, which can effectively accelerate the calculation speed. Since every video includes a wide range of appearance variations, the key to set-based video face recognition lies in two strategies: the representation of the videos and the accurate measurement of the distance between different videos. From the perspective of model representation and measurement learning, many set-based video face-recognition methods have been presented in recent decades, in order to identify the optimal recognition performance, including the discriminant canonical correlations (DCC) algorithm [14], covariance discriminative learning (CDL) algorithm [15], dual-linear regression classification (DLRC) algorithm [16], manifold discriminant analysis (MDA) algorithm [17], pairwise linear regression classification (PLRC) algorithm [18], image set-based collaborative representation and classification (ISCRC) algorithm [19], and set-level joint sparse representation classification (SJSRC) algorithm [20]. Among these algorithms, LRC-based methods have gradually attracted the attention of a greater number of researchers because of their superior performance.

Recently, the PLRC algorithm [18] has been proven to be a valid image set-classification method. By analyzing PLRC, the distance between images has been identified as the key point in the construction of unrelated sets. However, previously developed unrelated set construction strategies (i.e., S1 and S2 in [18]) are not optimal distance metric methods. Fortunately, based on collaborative and sparse representation, Gao et al. proposed another two unrelated set construction strategies: S3 and S4 [21]. In addition, when the size of the videos or image sets is small, PLRC can work well (see Figure 4). Therefore, how to make PLRC suitable for large-size videos has become an interesting problem. Gao et al. proposed the kernel PLRC (KPLRC) algorithm [21], with the aim of increasing the dimensionality of image samples in an image set, in order to increase the linear separability and overcome

this issue. Unfortunately, KPLRC not only increases the classification accuracy, but also the additional computational overhead.

Motivated by this, and starting from the direction of decreasing the number of images in each video, this paper proposes a novel video face recognition or image set classification framework, named sample reduction-based PLRC (SRbPLRC), which consists of divide SRbPLRC (DSRbPLRC), anchor point SRbPLRC (APSRbPLRC), and attention anchor point SRbPLRC (AAPSRbPLRC) methods. Specifically, the DSRbPLRC algorithm was evaluated first, which simply divides large-size videos into several small-size videos/subsets randomly. However, DSRbPLRC increases the number of videos, which may reduce the computing speed, and it does not consider the influence of any noise or outliers that may exist in the videos. Then, we looked at the APSRbPLRC algorithm, which first uses the clustering method to divide each video into several sub-videos, and then the centroid and mean values of each sub-video were used as anchor points. APSRPLCR can greatly reduce the number of samples in videos and does not increase the number of videos. Thus, it can realize set-based video face recognition efficiently and accurately. Additionally, APSRPLCR can overcome the influence of noise and outliers to some extent. Nevertheless, APSRPLCR considers all images within each sub-video to be of equal importance while computing the anchor points; however, we believe that a good anchor point construction strategy should adaptively weigh and combine the images in each sub-video/subset. Thus, the AAPSRbPLRC algorithm was developed, which utilizes an attention mechanism to learn the optimal anchor points. An overall flow chart of this algorithm is shown in Figure 1. Finally, several experiments were conducted on some popular databases (i.e., Honda, Mobo, and YTC) to demonstrate the effectiveness of our proposed method for large-size video face recognition. The main contributions of this paper are given as follows:

- A simple, yet effective, sample reduction-based pairwise linear regression classification framework is proposed, which consists of three methods, and these methods are augmented one by one.
- By introducing the attention mechanism, our attention anchor point SRbPLRC method can adaptively weigh and combine the images in each cluster, so as to achieve faster and effective image set classification.
- Experimental results can demonstrate the effectiveness of our proposed three methods for large size video face recognition. Therefore, our proposed methods can effectively meet the real-time and security requirements of IoT monitoring system.

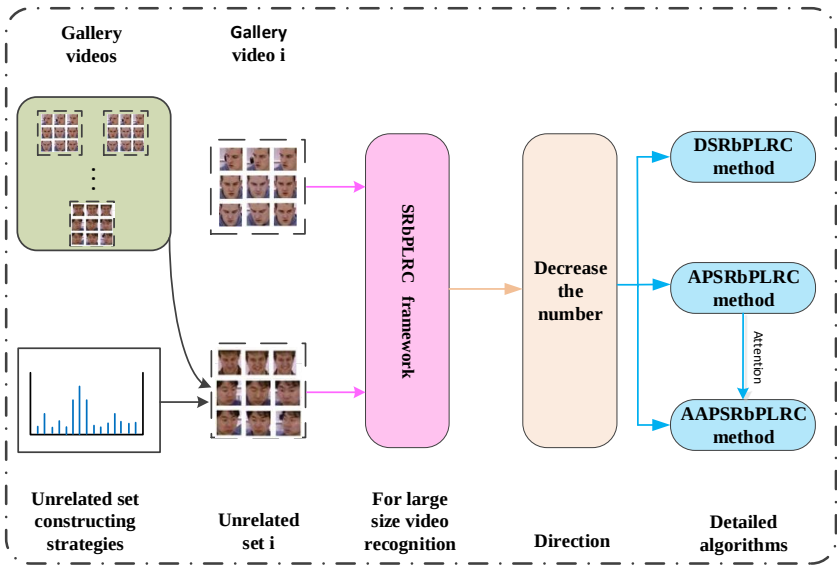


Figure 1. An illustration of the proposed framework. Aiming at obtaining an effective and efficient video face recognition method, a new SRbPLRC framework is proposed, by decreasing the number of images in each video. This framework contains DSRbPLRC, APSRbPLRC, and AAPSRbPLRC methods.

The rest of this paper is organized as follows. In Section 2, we briefly review the existing set-based video face-recognition algorithms. In Section 3, we present the SRbPLRC framework in detail. The experiments and the results analysis are discussed in Section 4. In Section 5, we present our conclusions.

2. Related Works

In this section, we briefly review some related works, which cover Internet of Things and video face-recognition methods for IoT monitoring systems.

2.1. Internet of Things

The definition of IoT is as follows: "through various information sensors, collect any object or process that needs to be monitored, connected and interacted in real time, collect all kinds of required information such as sound, light, heat, electricity, mechanics, chemistry, biology and location, and realize the universal connection between things and people, and realize the intelligent perception, recognition and management of things and processes through various possible network access". According to this definition, there are many research branches in IoT, such as IoT basic theory and technology, IoT data management and middleware technology, IoT electrical automation and remote monitoring, etc.

In the IoT basic theory and technology branch, researches mainly study the IoT communication protocol and node technology, sensor network architecture deployment and performance evaluation, information security and privacy, multi-sensor information fusion, etc. For example, regarding the IoT scenario, Cauteruccio et al. had developed many works on social networks [22–24]. For instance, in the literature [24], they proposed a new method to investigate anomalies in a multiple IoT scenario. In the literature [22], the authors proposed computing the scope of a social object in a multi-IoT scenario; and in another study [23], they proposed three new measures of betweenness centrality, specifically conceived for a multiple IoT scenario. Besides, Nicolazzo et al. [25] proposed a privacy-preserving approach to prevent feature disclosure in a multiple IoT scenario, i.e., a scenario where objects can be organized into (partially overlapped) networks that interact with each other.

In the IoT data management and middleware technology branch, researches mainly study massive data processing and analysis technology, as well a sdata storage, query, mining, analysis and fusion technology, in other words, applying artificial intelligence (AI) algorithms to IoT systems, and thus producing the AIoT concept. The AI technology is usually used to interpret and respond to some human-to-machine and machine-to-machine data flows in real time, and it can reduce latency, as well as increase the privacy and real-time intelligence of the system at the edge. This also means that fewer data need to be sent and stored on the server. Note that the research in this paper just belongs to this research field.

2.2. Video Face-Recognition Methods for IoT Monitoring Systems

Video- or image set-based face-recognition has been actively studied for decades. In this paper, we only considered inputs of orderless sets of face images. Since many researchers have pointed out that temporal information is not important for face-recognition tasks, existing methods that exploit temporal dynamics are not considered here.

Up until now, existing video recognition or image set classification approaches have been roughly grouped into three classes: model representation-based methods, metric learning-based methods, and simultaneous model representation- and metric learning-based methods.

Model representation-based methods mainly focus on how to build more discriminative model representations. For example, the discriminant canonical correlations (DCC) algorithm [14] used subspaces to model videos or image sets, and employed the canonical correlation metric to measure the similarities between two videos. The log-Euclidean metric learning (LEML) algorithm [26] and covariance discriminative learning (CDL) al-

gorithm [15] both used covariance matrices to model videos and logarithm kernels, in order to learn the distance between videos or image sets. Single Gaussian or Gaussian mixture models (GMM) used probability distributions to model videos, and utilized existing K-L divergence to measure the similarities between the distributions. The disadvantage of model representation-based methods is that they ignore the importance of distance metric learning.

Different from model representation-based approaches, metric learning-based methods mainly use existing strategies to represent videos, and focus on how to learn accurate distance metrics. Cevikalp and Triggs [27] modeled image sets and videos as affine hulls and convex hulls, and then proposed two metric learning methods: the convex hull-based image set distance (CHISD) algorithm and the affine hull-based image set distance (AHISD) algorithm. They then developed two optimization functions to learn distance metrics, based on collaborative representation. The SANP [28] method also used affine hulls to represent videos, and employs a new distance learning method, which tries to identify the SNAP between two videos. The ISCRC algorithm [19] represents probe videos as regularized/convex hulls and uses all gallery videos to collaboratively reconstruct these hulls in order to learn the discriminative distances. Dual-linear regression classification (DLRC) [16] took LRC as its baseline and extended it to video recognition. Specifically, DLRC used spanning subspaces to represent image sets and videos, and tried to learn more suitable set-based distances. Inspired by DLRC, the concept of unrelated sets was first defined in PLRC, which learned related distance metrics, unrelated distance metrics, and combination metrics to improve classification performance. Unfortunately, experiments have shown that even though DLRC and PLRC can achieve satisfactory performances in small-size video recognition or image set classification tasks, they do not perform well on large-size videos. Additionally, the existing unrelated set construction strategies (i.e., S1 and S2 in PLRC) are not optimal. So, how to construct more discriminative unrelated sets and extend the methods to large-size video recognition tasks remain challenging problems. Gao et al. proposed the KPLRC algorithm, which uses two new, unrelated set-constructing strategies, and tries to increase the dimensionality of the image samples in each image set to overcome the limitations of PLRC. Unfortunately, KPLRC not only increases the classification accuracy but also the additional computational overhead. Inspired by this, a novel SRbPLRC framework is proposed in this paper, which combines the two new unrelated set construction strategies from KPLRC with three extensions of SRbPLRC, for large-size video recognition or image set classification.

Simultaneous model representation- and metric learning-based methods consider model representation and metric learning at the same time, in order to achieve more accurate video recognition. For instance, the joint metric learning-based class-specific representation (JMLC) [29] framework modeled image sets as affine hulls, and simultaneously learned the related and unrelated distance metrics. The projection metric learning (PML) algorithm [30] used subspaces to represent videos, and learned the projection metric directly using Grassmann manifolds, in order to reduce the computation costs. The MDA algorithm [17] and manifold–manifold distance (MMD) algorithm [31] both modeled videos as local linear models. Thus, the similarities between two videos are then characterized by the distance between their corresponding local linear models. The covariate-relation graph (CRG) algorithm [32] used graphs to represent videos, and tried to learn the graph guide distance metric. The multi-model fusion metric learning (MMFML) algorithm [33] was developed because it was thought that using one model to represent each image set was not enough; therefore, this algorithm used multi-models to jointly model each video or image set and adopted a distance fusion method. Other methods in the literature [34–36] also used subspace points, which lie on Grassmannian manifolds, to represent videos, and different discriminant analysis approaches have been developed to learn more effective distance metrics.

3. SRbPLRC Framework for Large-Size Video Face Recognition

In one study [18], the PLRC algorithm has been developed and a definition of unrelated sets has been introduced, with the aim of improving discriminative information. According to that definition of unrelated sets, N_i samples need to be selected from all classes (except the *ith* class) to construct an unrelated set. Additionally, two unrelated set construction strategies have been developed in PLRC, which are denoted as S1 and S2. In S1, response vectors are reconstructed using a complete dataset $X = [X_1, X_2, \cdots, X_C] \in \mathcal{R}^{q \times L}$ and a probe set $Y \in \mathcal{R}^{q \times n}$, where L is the sum of all N_i and N_i is the number of samples in X_i. Obviously, q, L, n satisfy the inequality $q < L + n$, which indicates that S1 does not satisfy the basic conditions of the DLRC algorithm. Hence, the learned reconstruction coefficient of S1 is unbelievable (see Figure 4). Inspired by S1, two novel unrelated set construction strategies have further been proposed in the literature [21]: S3 and S4. The details of these strategies are shown in Figure 2. However, although KPLRC overcomes the disadvantages of PLRC with the help of S3 and S4, its computing time is very high. Aiming at obtaining an effective and efficient video face recognition method, we developed the new SRbPLRC framework by decreasing the number of images in each video. Note that our proposed framework will borrow the above mentioned unrelated set-construction strategies.

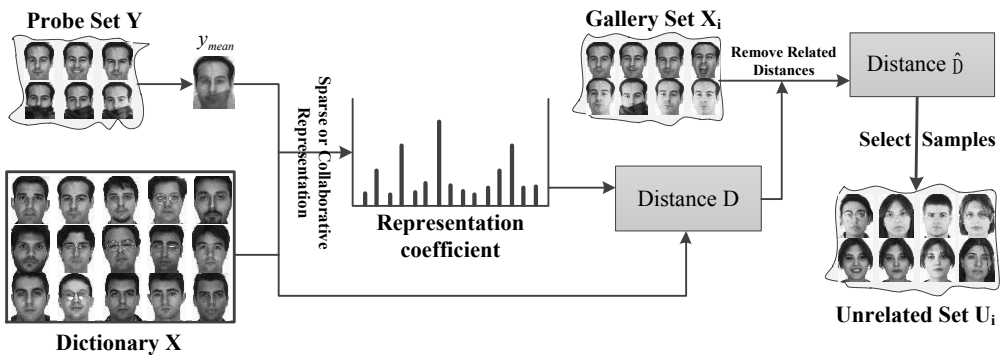


Figure 2. A diagram of the unrelated S3 and S4 set-constructing strategies. In S3, the sparse representation is used to learn the distances, while in S4, the collaborative representation is used.

3.1. DSRbPLRC Algorithm

Suppose that there are C large-size videos/image sets, denoted as $\{X_1, \cdots, X_C\}$, with each video containing N_i images and each image being $x_m^i \in \mathcal{R}^{q \times 1}$. Additionally, suppose that the image vector x_m^i is normalized to [0, 1]. Using these definitions, the ith gallery video or image set can be formulated as $X_i = \{x_j^i | j = 1, \cdots, N_i\}$ and the probe video or image set can be represented as $Y = \{y_j | j = 1, \cdots, n\}$, where n is the number of images in the probe video Y.

The DSRbPLRC algorithm aims to divide large-size videos into several small-size sub-videos to improve its ability to deal with the large-size video recognition task. Assuming that t is the division threshold, i.e., the number of samples in the divided video is no larger than t, the related (or unrelated) gallery video X_i can be randomly divided into $X_{i1}, X_{i2}, \cdots, X_{ia}$, where $a = [N_i/t + 1]$ is the number of the divided sub-videos, and

$$\begin{aligned} X_{ij} &= \{x_1^{ij}, \cdots, x_t^{ij}\}, \ \ j = 1, \cdots, a-1, \\ X_{ia} &= \{x_1^{ia}, \cdots, x_{N_i - t(a-1)}^{ia}\}, \end{aligned} \tag{1}$$

where i means the sub-video X_{ij} or X_{ia} comes from the original i^{th} gallery video, and j means that it is the jth sub-video of X_i.

Similarly, the probe video Y can be divided into $Y_1, Y_2, \cdots, Y_b$, where $b = [n/t + 1]$ is the number of the divided sub-videos of Y, and

$$\begin{aligned} Y_e &= \{y_1^e, \cdots, y_t^e\}, \ \ e = 1, \cdots, b-1, \\ Y_b &= \{y_1^b, \cdots, y_{n-t(b-1)}^b\}, \end{aligned} \tag{2}$$

where e means the eth sub-video of Y.

The DSRbPLRC algorithm uses spanning subspaces to represent each sub-video; thus, the gallery set X_{ij} can be represented as $X_{ij}\alpha = [x_1^{ij}, \cdots, x_t^{ij}]\alpha \in \mathcal{R}^{q\times 1}$, $i = 1, \cdots, C$, $j = 1, \cdots, a$, which can represent all images in X_{ij}. The probe set Y_e can be represented as $Y_e\eta = [y_1^e, \cdots, y_t^e]\eta$, $e = 1, \cdots, b$. Finally, the unrelated set U_{ij} that corresponds to the gallery set X_{ij} can be represented as $U_{ij}\gamma = [u_1^{ij}, u_2^{ij}, \cdots, u_{N_i}^{ij}]\gamma \in \mathcal{R}^{q\times 1}$.

By minimizing the Euclidean distance between $Y_e\eta$ and $X_{ij}\alpha$ (or $U_{ij}\gamma$), we obtain the following optimization problem:

$$\begin{aligned} &\min \ \ ||X_{ij}\alpha - Y_e\eta||_2^2 \\ &s.t. \ \ \sum_i \alpha_i = 1, \ \sum_i \eta_i = 1, \end{aligned} \tag{3}$$

and

$$\begin{aligned} &\min \ \ ||U_{ij}\gamma - Y_e\eta||_2^2 \\ &s.t. \ \ \sum_i \gamma_i = 1, \ \sum_i \eta_i = 1. \end{aligned} \tag{4}$$

Then, using the least squares algorithm, the related and unrelated representation coefficients $\beta_{e,r}^{ij}$ and $\beta_{e,u}^{ij}$ between (X_{ij}, Y_e) and (U_{ij}, Y_e), respectively, can be computed as follows:

$$\begin{aligned} \beta_{e,r}^{ij} &= ({S_{e,r}^{ij}}^T S_{e,r}^{ij})^{-1} {S_{e,r}^{ij}}^T s_{e,r}^{ij}, \\ \beta_{e,u}^{ij} &= ({S_{e,u}^{ij}}^T S_{e,u}^{ij})^{-1} {S_{e,u}^{ij}}^T s_{e,u}^{ij}, \end{aligned} \tag{5}$$

where $s_{e,r}^{ij} = y_{mean}^e - x_{mean}^{ij}$, $s_{e,u}^{ij} = y_{mean}^e - u_{mean}^{ij}$, $S_{e,r}^{ij} = [\bar{X}_{ij}, -\bar{Y}_e]$, $S_{e,u}^{ij} = [\bar{U}_{ij}, -\bar{Y}_e]$, $\bar{X}_{ij} = [x_1^{ij} - x_{mean}^{ij}, \cdots, x_{N_i}^{ij} - x_{mean}^{ij}]$, $\beta_{e,r}^{ij} = [\alpha; \eta]$, $\beta_{e,u}^{ij} = [\gamma; \eta]$, and the definitions of $\bar{Y}_e$, $\bar{U}_{ij}$ are the same as in $\bar{X}_{ij}$.

Thus, the related metric distance $d_{e,r}^{ij}$ and the unrelated metric distance $d_{e,u}^{ij}$ between X_{ij} and Y_e can be calculated as follows:

$$\begin{aligned} d_{e,r}^{ij} &= ||s_{e,r}^{ij} - S_{e,r}^{ij}\beta_{e,r}^{ij}||_2, \\ d_{e,u}^{ij} &= ||s_{e,u}^{ij} - S_{e,u}^{ij}\beta_{e,u}^{ij}||_2. \end{aligned} \tag{6}$$

Finally, the combination metric $d_e^{ij} = d_{e,r}^{ij}/d_{e,u}^{ij}$ is used to obtain the label of sub-video Y_e (the detail can be found in [29]). Thus, the label of probe video Y can be obtained by using the majority voting strategies.

3.2. APSRbPLRC Algorithm

From the above subsection, we can see that the DSRbPLRC algorithm increases the number of videos, which may reduce the computing speed, and it does not consider the influence of any noise or outliers that may exist in the videos. Thus, the APSRbPLRC algorithm is introduced in this subsection, which first uses the clustering method to divide each video into several sub-videos, and then uses the centroid and mean values of each sub-video as anchor points.

Anchor points are the most representative points in videos or image sets, and the subsets formed by anchor points can approximately replace the original videos or sets. In order to decrease the size of image sets or videos and improve efficiency, subsets of anchor points are constructed by extracting the common characters from each image and singling out the most representative points. An illustration of this process is shown in Figure 3.

In this study, the hierarchical divisive clustering (HDC) algorithm [31] is used as the clustering method to extract the clusters, i.e., to divide each video into several sub-videos. Specifically, the HDC algorithm regards each image set or video as a manifold and measures the non-linearity degrees of different clusters according to the deviations between the Euclidean distances and geodesic distances. This ensures that there is a similar number of images in the extracted clusters, and that the samples in each cluster have similar visual characteristics. Then, the mean image of each cluster is computed as the anchor point, which can in turn represent the whole subset.

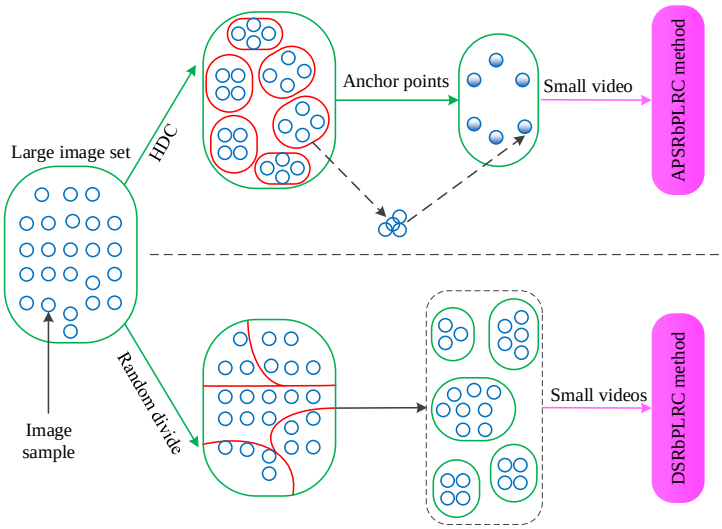


Figure 3. An illustration of the proposed DSRbPLRC and APSRbPLRC algorithms. DSRbPLRC directly divides a large-size video into several small-size sub-videos, and each sub-video is used for classification. Different from DSRbPLRC, APSRbPLRC uses the HDC algorithm to extract anchor points, which will not increase the number of videos.

After obtaining the anchor points, the size of the videos is greatly reduced, and due to the necessary information being preserved by the anchor points, recognition rates can also be retained. We should point out that other anchor points extraction strategies can also be used, such as the K-means clustering approach, spectral clustering algorithms, and one-class SVMs.

In this study, for C gallery videos $\{X_1, \cdots, X_C\}$ and a probe video Y, the HDC algorithm is first used to extract the anchor points of each video, so we could obtain the following anchor point gallery and probe videos: $\{X_1^{ap}, \cdots, X_C^{ap}\}$ and Y^{ap} (where X_i^{ap} contains N_i^{ap} representative images, Y^{ap} contains n^{ap} representative images, and $N_i^{ap} \ll N_i$, $n^{ap} \ll n$ is satisfied). As with the DSRbPLRC algorithm, we could then directly obtain the label of Y using Equations (5) and (6).

3.3. AAPSRbPLRC Algorithm

The APSRbPLRC algorithm uses the mean image of each subset as the anchor point; however, we believe that a good anchor point construction strategy should adaptively weigh and combine the images in each subset. Hence, we adopted an attention mechanism to obtain more discriminative anchor points.

Suppose that we obtained subset $\{x_1, \cdots, x_m\}$ using the HDC algorithm, and then initialized the query vector $q^0 \in \mathcal{R}^{q \times 1}$. We can then obtain the score of each image x_i by computing the dot product as follows:

$$e_k = (q^0)^T x_k, \tag{7}$$

where T is the transpose operation. The attention distribution α_k can then also be computed, as follows:

$$\alpha_k = \frac{\exp(e_k)}{\sum_j \exp(e_j)}. \tag{8}$$

Finally, the weighted average of subset can be obtained as follows:

$$r^0 = \sum_{k=1}^{m} \alpha_k x_k. \tag{9}$$

The obtained r^0 can be input into the following equation to obtain the new query vector q^1, q^2:

$$\begin{aligned} q^1 &= Tanh(W_1 r^0 + b_1), \\ q^2 &= Tanh(W_2 q^1 + b_2). \end{aligned} \tag{10}$$

Finally, the anchor point r^2 of the subset generated by q^2 can be obtained using Equations (7)–(9).

In order to solve the parameters in Equation (10), the subspace-based contrastive loss is used in this paper. Suppose that $\{X_1^{ap}, \cdots, X_C^{ap}\}$ are anchor point gallery videos computed after the above attention mechanism, then the distance d^{ij} between X_i^{ap} and X_j^{ap} can be computed by Equations (5) and (6). Thus, the subspace based contrastive loss can be defined as:

$$L = \sum_{ij} y_{ij} d^{ij} + (1 - y_{ij}) \max(0, m - d^{ij}) \tag{11}$$

where y_{ij} equals 1 or 0 ($y_{ij} = 1$ means that X_i^{ap} and X_j^{ap} belong to the same class, whereas $y_{ij} = 0$ means that X_i^{ap} and X_j^{ap} belong to different classes).

After obtaining more discriminative anchor points for each video, the AAPSRbPLRC method can effectively produce the labels for the probe video.

In summary, when we suppose that each sub-video or subset has a similar number of images, then the PLRC algorithm needs to compute C large-size distances to classify one probe video, while the DSRbPLRC algorithm needs to compute $C \times a \times b$ small-size distances, and the APSRbPLRC and AAPSRbPLRC algorithms both need to compute C small-size distances.

4. Experiments

4.1. Experimental Settings

Next, we verified the effectiveness of the proposed framework for large-size video recognition tasks. In this study, the division threshold t in DSRbPLRC was set as $t = 100$. There are two parameters for the APSRbPLRC algorithm: the threshold parameter θ and the nearest neighbor parameter k. According to [31], it is known that larger θ values imply fewer local models (thus, higher efficiency) but larger linearity deviations, and vice versa. Thus, the parameter θ was selected by searching from $1, 2, \cdots, 5$, while parameter k was selected by searching from $1, 2, \cdots, 12$. In the AAPSRbPLRC algorithm, the parameter m was set as 1.5.

4.2. Methodology

In this subsection, we describe the comparison experiments that we conducted, in order to compare our proposed method with some state-of-the-art set-based video recognition methods, including DCC [14], AHISD, CHISD [27], SANP [28], MMD [31], COV [15], ISCRC [19], LEML [26], PML [30], the covariate-relation graph (CRG) algorithm [32], DLRC [16], and PLRC (PLRC-I and PLRC-II) [18]. The implementation codes of these algorithms were provided by the original authors, and their parameters were empirically tuned, according to the recommendations in the original references. The attention mechanism in our AAPSRbPLRC algorithm was trained using a standard backpropagation algorithm. On the Honda and Mobo databases, we randomly divided the gallery videos into two small videos, as each class only has one gallery video.

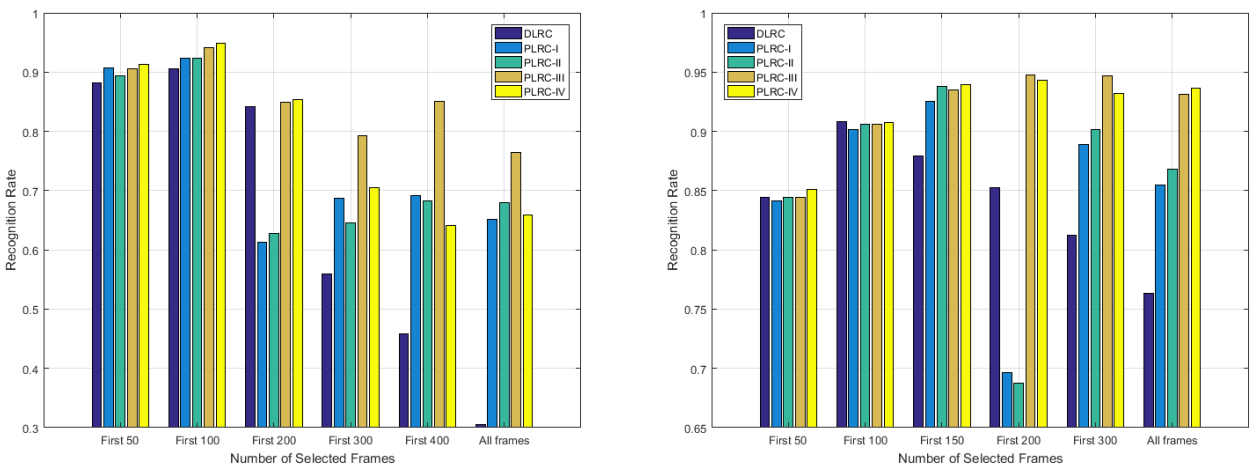
4.3. Databases

In this study, three popular large-size databases were used in our experiments: the Honda database, Mobo database, and YouTube Celebrities (YTC) database. The Honda/UCSD database contains 59 video sequences from 20 different people and there are about 300∼500 frames in each sequence. For fair comparison, 20 sequences were randomly selected for training (one sequence was selected from each class), and the other 39 video sequences were selected for testing. So, the video frames from each video sequence became a face image set. The Mobo database contains 96 video sequences from 24 different people, with each person having four video sequences in each class. In our experiments, we randomly selected one sequence from each person for training, and used the remaining three video sequences for testing. The YTC database is a large-size video recognition database, which consists of 1910 video clips from 47 celebrities on the YouTube website. In each clip, there are hundreds of frames. In our experiments, we randomly selected three video clips or image sets from each celebrity for training and six video clips or image sets for testing. All of the face images in these databases were resized to 20×20.

4.4. SRbPLRC for Large-Size Video Recognition

On the Honda database, the biggest video sequence contains 618 images. On the Mobo database, the biggest video sequence includes 340 images, while on the YouTube Celebrities database, the biggest video sequence includes 349 frames. Hence, we will perform large-size video recognition experiments on these databases.

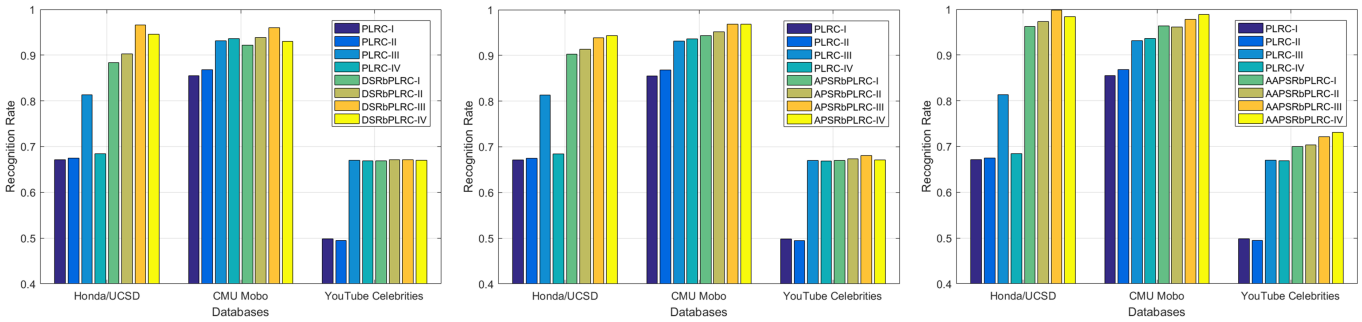
Figure 4 illustrates the recognition rates of 10 random experiments using the DLRC and PLRC methods. The videos were from the Honda and Mobo databases and contained varying numbers of frames. In the experiments, PLRC-III and PLRC-IV are the combinations of PLRC and the S3 and S4 unrelated set construction strategies. From the figure, it can clearly be observed that the DLRC, PLRC-I, and PLRC-II methods only work well on small videos (when there are fewer than 150 frames in each video). Specifically, on the Honda database, the performance of the DLRC, PLRC-I, and PLRC-II methods is significantly reduced when the number of frames is larger than 150. For the Mobo database, the demarcation point for the DLRC, PLRC-I, and PLRC-II methods is also 150 frames. We also noticed that the PLRC-III and PLRC-IV methods perform better than the DLRC, PLRC-I, and PLRC-II methods.



(a) Honda database (b) Mobo database

Figure 4. The experimental results of DLRC and PLRC methods from two popular databases, with varying numbers of frames.

In subsequent experiments, the Honda, Mobo and YouTube Celebrities databases, with all their frames, were used as the large-sized videos. The average classification accuracies of the DSRbPLRC, APSRbPLRC, and AAPSRbPLRC algorithms are shown in Figure 5.



(a) DSRbPLRC method (b) APSRbPLRC method (c) AAPSRbPLRC method

Figure 5. The average recognition rates of the DSRbPLRC, APSRbPLRC, and AAPSRbPLRC algorithms, in terms of large-sized videos.

Figure 5 lists a quantitative evaluation (i.e., the average classification accuracy) of the DSRbPLRC, APSRbPLRC, and AAPSRbPLRC algorithms for large-size videos. It can be observed that our proposed method outperforms PLRC in all cases, which indicates that our SRbPLRC framework is effective, i.e., the DSRbPLRC, APSRbPLRC, and AAPSRbPLRC algorithms can overcome the limitations of PLRC to some extent. We also found that the DSRbPLRC and APSRbPLRC algorithms achieve similar performances, while the AAPSRbPLRC algorithm achieves a better performance, which indicates that the attention mechanism could find more discriminative anchor points. From Figure 5, it can also be seen that the unrelated S3 and S4 set -construction strategies outperform the other unrelated subspace-construction strategies. These quantitative and qualitative comparisons confirm the applicability of the proposed method for large-size video recognition.

Finally, Table 1 enumerates the performances of the different methods on the different databases when using all frames. The results of DSRbPLRC, APSRbPLRC and AAPSRbPLRC are the best classification accuracies selected from Figures 4 and 5. As can be seen from the table, we observed the following: (1) the proposed method performs better than PLRC, which indicates that our proposed approach is effective for the large-sized video face-recognition task; (2) the AAPSRbPLRC algorithm obtains the best classification accuracy on all databases, while the APSRbPLRC algorithm achieves a sub-optimal classification performance on the Mobo database, which means that decreasing the number of images in the videos used can overcome the shortcomings of the DLRC and PLRC algorithms, and that the attention mechanism is helpful for the large-sized video face-recognition task.

Table 1. The average classification accuracies of the different methods on the three popular databases (%), where the bold denotes the best results.

Method	Honda	Mobo	YTC
DCC	96.67	88.54	61.73
MMD	97.18	94.10	67.23
LEML	97.18	88.69	50.60
PML	96.67	89.69	66.13
CRG	88.46	70.48	49.38
COV	**100.0**	90.35	69.18
SANP	93.60	93.87	66.70
AHISD	96.34	93.13	65.13
CHISD	95.44	94.21	63.40
DLRC	30.54	77.30	37.31
PLRC-I	65.13	84.67	49.89
PLRC-II	67.95	86.03	49.43
DSRbPLRC	96.59	95.96	67.18
APSRbPLRC	96.36	96.83	68.11
AAPSRbPLRC	**100.0**	**97.98**	**72.23**

4.5. Computation Times

In this subsection, we present the testing times for identifying a query video on the Honda and Mobo databases, using different methods. Here, only the unrelated S1 and S2 set-construction strategies were used for all methods. The average computing times are shown in Table 2. From the table, we can see the following: (a) compared to PLRC, our APSRbPLRC and AAPSRbPLRC algorithms require less time for testing, possibly because our anchor point method can effectively decrease the set size; (b) compared to the other methods, the DSRbPLRC algorithm requires the most testing time, which is due to the fact that the DSRbPLRC algorithm needs to solve more optimization problems.

Table 2. The testing times for identifying a query video, on the Honda and Mobo databases.

Method	Honda	Mobo
PLRC-I	2.03	3.38
PLRC-II	6.33	9.38
DSRbPLRC-I	20.38	30.89
DSRbPLRC-II	80.80	100.8
APSRbPLRC-I	1.28	2.89
APSRbPLRC-II	4.98	8.83
AAPSRbPLRC-I	1.79	3.21
AAPSRbPLRC-II	5.67	8.87

5. Conclusions

In this study, a new set-based video face recognition framework for IoT monitoring systems, called sample reduction-based pairwise linear regression classification, was proposed, which can effectively classify large-size videos. Subsequently, three classifiers (DSRbPLRC, APSRbPLRC, and AAPSRbPLRC) were also developed by decreasing the number of images in each video. Our experimental results on some popular databases demonstrate the superiority of our DSRbPLRC, APSRbPLRC, and AAPSRbPLRC algorithms. Specifically, qualitative and quantitative comparisons between our proposed method, DLRC, and PLRC for the large-size video face-recognition task demonstrate that the DSRbPLRC, APSRbPLRC, and AAPSRbPLRC algorithms can overcome the limitations of PLRC. We also find that the AAPSRbPLRC method achieves the best classification results on all databases.

The main limitations of the proposed framework include the fact that we did not consider the importance of feature learning; only one model feature is used in our framework, while multi-model features are very common in IoT monitoring systems. Therefore, the unrelated set construction strategy can be further improved. Hence, our future studies will combine our framework with deep neural networks, classify multi-feature videos, and design new unrelated set-construction strategies. Besides this, we will also research the accelerating deploy problem in IoT systems using these algorithms.

Author Contributions: Conceptualization, X.G. and W.H.; Data curation, W.H.; Formal analysis, W.H.; Funding acquisition, X.G.; Investigation, X.G.; Methodology, X.G. and S.N.; Project administration, X.G.; Resources, S.N.; Software, W.H.; Supervision, S.N.; Validation, Y.C.; Visualization, Y.C.; Writing—original draft, X.G.; Writing—review and editing, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China (grant number: 62101213), the Shandong Provincial Natural Science Foundation (grant number: ZR2020QF107), the Development Program Project of Youth Innovation Team of Institutions of Higher Learning in Shandong Province, and the Big Data Project of University of Jinan (grant number: XKY1926).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

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

References

1. Pedditi, R.B.; Debasis, K. Energy Efficient Routing Protocol for an IoT-Based WSN System to Detect Forest Fires. *Appl. Sci.* **2023**, *13*, 3026. [CrossRef]
2. Azfar, S.; Nadeem, A.; Ahsan, K.; Mehmood, A.; Siddiqui, M.S.; Saeed, M.; Ashraf, M. An IoT-Based System for Efficient Detection of Cotton Pest. *Appl. Sci.* **2023**, *13*, 2921. [CrossRef]
3. Chang, D.-M.; Hsu, T.-C.; Yang, C.-T.; Yang, J. A Data Factor Study for Machine Learning on Heterogenous Edge Computing. *Appl. Sci.* **2023**, *13*, 3405. [CrossRef]
4. Ullah, W.; Hussain, T.; Baik, S.W. Vision transformer attention with multi-reservoir echo state network for anomaly recognition. *Inf. Process. Manag.* **2023**, *60*, 103289. [CrossRef]
5. Ullah, W.; Hussain, T.; Khan, Z.A.; Haroon, U.; Baik, S.W. Intelligent dual stream CNN and echo state network for anomaly detection. *Knowl.-Based Syst.* **2022**, *253*, 109456. [CrossRef]
6. Turk, M.; Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **1991**, *3*, 71-86. [CrossRef] [PubMed]
7. Naseem, I.; Togneri, R.; Bennamoun, M. Linear regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *31*, 2106–2112. [CrossRef]
8. Gao, X.; Fan, L.; Xu, H. A novel method for classification of matrix data using twin multiple rank smms. *Appl. Soft Comput.* **2016**, *48*, 546–562. [CrossRef]
9. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J. Eigenfaces vs. fisher-faces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 711–720. [CrossRef]
10. Hastie, T.; Tibshirant, R. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 607–616. [CrossRef]
11. Gao, X.; Sun, Q.; Yang, J. MRCCA: A novel CCA based method and its application in feature extraction and fusion for matrix data. *Appl. Soft Comput.* **2018**, *62*, 45–56. [CrossRef]
12. Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 210–227. [CrossRef] [PubMed]
13. ur Rehman, A.; Belhaouari, S.B.; Kabir, M.A.; Khan, A. On the Use of Deep Learning for Video Classification. *Appl. Sci.* **2023**, *13*, 2007. [CrossRef]
14. Kim, T.K.; Kittler, J.; Cipolla, R. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1005–1018. [CrossRef] [PubMed]
15. Wang, R.; Guo, H.; Davis, L.S.; Dai, Q. Covariance discriminative learning: A natural and efficient approach to image set classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2496–2503.
16. Chen, L. Dual linear regression based classification for face cluster recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 2673–2680.
17. Wang, R.; Chen, X. Manifold discriminant analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 429–436.
18. Feng, Q.; Zhou, Y.; Lan, R. Pairwise linear regression classification for image set retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4865–4872.
19. Zhu, P.; Zuo, W.; Zhang, L.; Shiu, S.C.; Zhang, D. Image set-based collaborative representation for face recognition. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 1120–1132. [CrossRef]
20. Zheng, P.; Zhao, Z.Q.; Gao, J.; Wu, X. A set-level joint sparse representation for image set classification. *Inf. Sci.* **2018**, *448*, 75–90. [CrossRef]
21. Gao, X.; Sun, Q.; Xu, H.; Gao, J. Sparse and collaborative representation based kernel pairwise linear regression for image set classification. *Expert Syst. Appl.* **2020**, *140*, 112886. [CrossRef]
22. Cauteruccio, F.; Cinelli, L.; Fortino, G.; Savaglio, C.; Terracina, G.; Ursino, D.; Virgili, L. An approach to compute the scope of a social object in a Multi-IoT scenario. *Pervasive Mob. Comput.* **2020**, *67*, 101223. [CrossRef]
23. Cauteruccio, F.; Terracina, G.; Ursino, D.; Virgili, L. Redefining Betweenness Centrality in a Multiple IoT Scenario. In Proceedings of the 1st Workshop on Artificial Intelligence and Internet of Things Co-Located with the 18th International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, 19–20 November 2019; pp. 16–27.
24. Cauteruccio, F.; Cinelli, L.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L.; Savaglio, C.; Liotta, A.; Fortino, G. A framework for anomaly detection and classification in Multiple IoT scenarios. *Future Gener. Comput. Syst.* **2021**, *114*, 322–335. [CrossRef]
25. Nicolazzo, S.; Nocera, A.; Ursino, D.; Virgili, L. A privacy-preserving approach to prevent feature disclosure in an IoT scenario. *Future Gener. Comput. Syst.* **2020**, *105*, 502–519. [CrossRef]

26. Huang, Z.; Wang, R.; Shan, S.; Li, X.; Chen, X. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 720–729.
27. Cevikalp, C.; Triggs, B. Face recognition based on image sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2567–2573.
28. Zhao, Z.Q.; Xu, S.T.; Liu, D.; Tian, W.D.; Jiang, Z.D. A review of image set classification. *Neurocomputing* **2019**, *335*, 251–260. [CrossRef]
29. Gao, X.; Niu, S.; Wei, D.; Liu, X.; Wang, T.; Zhu, F.; Dong, J.; Sun, Q. Joint Metric Learning-Based Class-Specific Representation for Image Set Classification. *IEEE Trans. Neural Netw. Learn. Syst.* 2022, *Early Access*. [CrossRef]
30. Huang, Z.; Wang, R.; Shan, S.; Chen, X. Projection metric learning on grassmann manifold with application to video based face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 140–149.
31. Wang, R.; Shan, S.; Chen, X.; Gao, W. Manifold-manifold distance with application to face recognition based on image set. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 23–28 June 2008; pp. 1–8.
32. Chen, Z.; Jiang, B.; Tang, J.; Luo, B. Image set representation and classification with attributed covariate-relation graph model and graph sparse representation classification. *Neurocomputing* **2017**, *226*, 262–268. [CrossRef]
33. Gao, X.; Sun, Q.; Xu, H.; Wei, D.; Gao, J. Multi-model fusion metric learning for image set classification. *Knowl.-Based Syst.* **2019**, *164*, 253–264. [CrossRef]
34. Wei, D.; Shen, X.; Sun, Q.; Gao, X. Discrete Metric Learning for Fast Image Set Classification. *IEEE Trans. Image Process.* **2022**, *31*, 6471–6486. [CrossRef] [PubMed]
35. Wei, D.; Shen, X.; Sun, Q.; Gao, X.; Ren, Z. Sparse Representation Classifier Guided Grassmann Reconstruction Metric Learning with Applications to Image Set Analysis. *IEEE Trans. Multimed.* 2022, *Early Access*. [CrossRef]
36. Wang, T.; Shi, P. Kernel Grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognit. Lett.* **2009**, *30*, 1161–1165. [CrossRef]