electronics

MDPI

Article

Zero-Shot Deep Learning for Media Mining: Person Spotting and Face Clustering in Video Big Data

Mohamed S. Abdallah [1,2,*], HyungWon Kim [1,*], Mohammad E. Ragab [2] and Elsayed E. Hemayed [3,4]

1 Department of Electronics, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea
2 Informatics Department, Electronics Research Institute (ERI), Giza 12622, Egypt; mehab@eri.sci.eg
3 Computer Engineering Department, Faculty of Engineering, Cairo University, Giza 12613, Egypt; hemayed@ieee.org
4 University of Science and Technology, Zewail City of Science and Technology, October Gardens, Giza 12578, Egypt
* Correspondence: sameer@cbnu.ac.kr or sameer@eri.sci.eg (M.S.A.); hwkim@cbnu.ac.kr (H.K.)

Received: 23 October 2019; Accepted: 19 November 2019 ; Published: 22 November 2019

Abstract: The analysis of frame sequences in talk show videos, which is necessary for media mining and television production, requires significant manual efforts and is a very time-consuming process. Given the vast amount of unlabeled face frames from talk show videos, we address and propose a solution to the problem of recognizing and clustering faces. In this paper, we propose a TV media mining system that is based on a deep convolutional neural network approach, which has been trained with a triplet loss minimization method. The main function of the proposed system is the indexing and clustering of video data for achieving an effective media production analysis of individuals in talk show videos and rapidly identifying a specific individual in video data in real-time processing. Our system uses several face datasets from Labeled Faces in the Wild (LFW), which is a collection of unlabeled web face images, as well as YouTube Faces and talk show faces datasets. In the recognition (person spotting) task, our system achieves an F-measure of 0.996 for the collection of unlabeled web face images dataset and an F-measure of 0.972 for the talk show faces dataset. In the clustering task, our system achieves an F-measure of 0.764 and 0.935 for the YouTube Faces database and the LFW dataset, respectively, while achieving an F-measure of 0.832 for the talk show faces dataset, an improvement of 5.4%, 6.5%, and 8.2% over the previous methods.

Keywords: face clustering; face recognition; face detection; CNN; KL divergence; triplet loss

1. Introduction

Many methods have been studied to achieve the target of producing, processing, and recording of talk show videos in an effective way. However, a meaningful analysis of media content requires substantial manual efforts. This problem is encountered in TV production analysis and media mining applications, where the number of faces of individuals can be on the order of millions. Many talk show hours are broadcasted daily. The majority of these talk shows contain millions of frames. We consider clustering these large amounts of face images into a few hundred discrete identities to properly organize these vast amounts of data. In our research, a frame-based analysis is needed to make talk show videos searchable for identities (public figures) and useful for media mining and TV production analysis.

Video face clustering is a challenging problem. In talk show videos, the faces of individuals have high variations in appearance due to the diversity in head pose, facial expression, scale, illumination

and occlusions caused by other matters, such as glasses, hair, and hands. Other uncontrolled circumstances add challenges such as complex background, low resolution, image blurring, and motion speed. In our case of face clustering, we encounter other challenges, including the scalability of an algorithm; thus, it can be linearly extended to a larger number of frames. Some people may appear very often, whereas other people may appear less frequently; thus, the number of frames per individual is unbalanced.

In this study, we address the problem of partitioning numerous unlabeled face frames into the identities present in talk show videos. This paper presents an efficient method of indexing and clustering video data for achieving a media production analysis of identities in talk show videos. We present experimental results of the face clustering method and analyze them in terms of clustering accuracy and run time.

In addition to the indexing and clustering of all faces that appear in a video, the proposed system includes a face recognition function for public figures with training data.

In this paper, we present a TV media mining system that is based on deep convolutional neural networks (DCNNs) algorithms for face detection, face recognition, and face clustering. In recent years, DCNN approaches have made significant contributions in advancements in computer vision domains. Using these new technologies in media production analysis and mining in videos enables media production companies and TV channels to index their vast multimedia data with high accuracy and speed.

The remainder of the paper is organized as follows: Section 2 presents related studies. Section 3 describes the proposed system. Section 4 presents and discusses the experimental results and datasets in our research. Section 5 concludes the paper and outlines future research.

2. Related Research

The aim of object recognition is to recognize important objects in video frames. The aim of person spotting (face recognition) is to identify the appearance of a specific identity in TV talk shows. The appearance data includes the duration of the appearance and the title used to identify this appearance. The aim of the face clustering task is to partition individuals faces in TV talk show videos into a group of similar faces (cluster) that have the same appearance.

Numerous techniques have been suggested in the literature to address face recognition and face clustering problems [1–5]. Most of these techniques rely on the creativity of researchers to extract better features. These techniques remain inferior to their human vision counterpart. Thus, a paradigm shift is needed to develop face recognition techniques that can approach the performance of the human visual system. Inspired by the powerful structure of the human brain and the recent development in the neuroscience field, researchers paid more attention to the visual cortex. The researchers were able to partially mimic the visual cortex and achieved substantially higher accuracy than classical techniques [6–8]. With the recent developments in deep neural networks, the DCNN was proposed as an object or face recognition technique that can close the gap between computational models and the human vision system. Various DCNN approaches have been proposed in the literature.

In the following subsections, we provide a summary of face recognition and face clustering surveys. We first briefly introduce the classical methods followed by remarkable convolutional neural network (CNN) models.

2.1. Classical Techniques

In recent decades, classical techniques of object recognition, face recognition and clustering have played a very important role in advancing machine learning fields. In classical techniques, hand-crafted feature extractors are used to gather related information from an input image and eliminate the related variability. This process reduces the required memory and computation power, prevents overfitting, and generates a feature vector. Subsequently, a trained classifier categorizes the feature vectors into the corresponding classes. Classifiers include standard fully-connected feed-forward neural networks,

multi-layer neural networks and other classifiers such as support vector machine (SVMs). Therefore, the power of a classic technique depends on its feature extractor.

Table 1 shows a comparison of three examples of classical face recognition systems.

Table 1. Examples of classical face recognition systems.

Publication	Dataset	Features	Method
Mozhde Elahi et al. [1]	ORL face dataset	Wavelet transform, PCA, DCT	SVM
Sanjeev Kumar et al. [2]	Yale face	PCA, LDA, BPN	SVM
Bo Dai et al. [3]	AT&T face, AR face	SIFT, PCA, 2DPCA	SVM

Elahi et al. [1] compared the performance of principal component analysis (PCA) and discrete cosine transform (DCT) methods for feature reduction in a face recognition system. The researchers also applied wavelet transform for feature extraction and an SVM classifier for training and recognition.

Kumar et al. [2] investigated various methods for face recognition, such as PCA, linear discriminant analysis (LDA), neural networks, backpropagation networks (BPNs), and radial basis function networks (RBFNs) and discuss their advantages and disadvantages.

Dai et al. [3] reviewed well-known face recognition methods such as the scale-invariant feature transform (SIFT), PCA, and 2DPCA. Their results demonstrated that SIFT has significant advantages over both PCA and 2DPCA.

Belalia et al. [9] proposed a region-based image retrieval (RBIR) approach using Shape adaptive discrete cosine transform (SA-DCT). The features can be extracted directly from compressed images by using a discrete cosine transform (DCT). First, low-level features for each region are constructed from the coefficients of quantized block transforms, then histograms of local image features are used as descriptors of statistical information, and finally the combination of histograms of image regions (objects) is defined to integrate high-level semantic information.

Belhallouche et al. [10] presented an RBIR using shape adaptive discrete wavelet transform (SA-DWT). The features can be extracted using multi-features color, texture, and edge descriptors. SA-DWT represented the best way to exploit the coefficients characteristics and properties such as the correlation.

Table 2 shows a comparison of four examples of face clustering systems.

Table 2. Examples of face clustering systems.

Publication	Dataset	Features	Method
Zhao et al. [4]	Personal photo album	2DHMM, Contextual	Hierarchical clustering
Cui et al. [5]	Family photo album	LBP, clothing color, Texture	Spectral clustering
Ho et al. [11]	CMU PIE, Yale Face	Gradient, pixel intensity	Spectral clustering
Tian et al. [12]	Four disjoint datasets	Image + contextual	Partial clustering

Zhao et al. [4] clustered a personal photo album dataset by combining contextual information (time of clustering and the probability that faces simultaneously appear in images) with identities obtained by a two-dimensional hidden Markov model (2DHMM) and hierarchical clustering results.

Cui et al. [5] developed a tool that employs spectral clustering as an initial method for organizing photographs. Local binary pattern (LBP) features and color and texture features are extracted from detected faces and detected bodies, respectively.

Ho et al. [11] developed a spectral clustering algorithm that is based on variations in the affinity matrix by computing the probability that two face images depict the same object.

Tian et al. [12] developed a probabilistic clustering model that enables an algorithm to reject clusters that do not have distributed samples.

2.2. Deep Neural Network Techniques

From the feature extractor perspective, generating handcrafted features is a difficult task that should be performed by an expert for better performance. In addition, the feature vector becomes too long to obtain advanced techniques. CNNs eliminate the need for feature vectors and automatically extract features [13,14].

From the classifier perspective, techniques such as fully connected neural networks comprise a large number of parameters and, consequently, a large memory for storing weights, compared with convolutional neural networks of the same size. In addition, they require large training instances to make them invariant to shift, scale, and distortion, while CNNs are automatically invariant to these changes via parameter sharing across space. They disregard the topology of the input images using fixed representations that are not affected by local features, while CNNs extract local features and then combine them using the concept of local receptive fields [15].

CNN is a popular deep learning technique that is employed for object recognition. CNN is a more popular solution than other deep learning techniques because CNN is the most relevant technique for the human visual system among other deep learning techniques. The CNN hierarchical structure is inspired by the feed-forward hierarchal system of the human visual system. CNN is biologically inspired by the cat's visual cortex system based on Hubel and Wiesel research and the Fukushima model, which is named Neocognitron [6–8].

LeCun et al. provided LeNet-5 [16], which is the first CNN for handwriting recognition and applied backpropagation for learning. In object recognition and detection, Krizhevsky et al. provided the first CNN, which is named AlexNet [17] and achieved a better performance than state-of-the-art methods. Additional research provided various CNNs with different techniques. Zeiler and Fergus developed ZFNet [18], which was trained using only 1.3 million images compared with AlexNet, which was trained using the entire ImageNet dataset with approximately 15 million images. ZFNet outperforms AlexNet with an 11.74% in the top-5 test error. GoogLeNet [19], which was developed by Szegedy et al., suggested a new CNN architecture named Inception, which won in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) for both the object recognition field and detection field. Their main contribution is the usage of very deep convolutional networks and the advanced utilization of resources.

2.3. Recent Related Research

Recently, a vast collection of face recognition and clustering studies has been proposed. In this subsection, we only discuss the most recent research that is related to DCNN.

The research of [20] proposes a multi-stage complex system that merges the output of a DCNN with PCA for dimensional reduction and an SVM for classification.

Zhenyao et al. propose a DCNN that warps faces into a canonical frontal view and then train it to classify each known identity. The authors also combine PCA of the network output with a collection of SVMs for face verification [21].

Sun et al. [20,22] present 25 networks, each operating on a different face patch. The authors combine 50 responses (regular and flipped) for their performance in the Labeled Faces in the Wild (LFW) dataset. The researchers also employ both PCA and a joint Bayesian model [23] in the embedding space that effectively corresponds to a linear transform.

Taigman et al. employ a multi-stage approach that aligns faces to a 3D shape model. A multi-class network is trained on approximately four thousand identities to perform the face recognition task. The authors also used a Siamese network, where they optimize the L1-distance between two face features [24].

Schroff et al. [25] propose a system named FaceNet, which is based on learning a Euclidean embedding per image using a DCNN. The CNN is trained in a way that makes the embedding space (squared L2 distances) corresponds to the face similarity. Once this embedding has been produced, face verification, face recognition, and clustering can be achieved. Schroff et al. explore two different DCNN architectures that have been recently utilized to achieve success in the computer vision community [24]. The first architecture is based on the Zeiler and Fergus [18] model and the second architecture is based on the Inception model of Szegedy et al. [19].

The authors in [26] show how Guillaumin et al. employ two methods for learning distance measures: a logistic discriminant method and the nearest neighbor method. The logistic discriminant method learns the metric from a set of labeled face pairs, while the nearest neighbor method computes the probability that a pair of faces are related to the same class.

Shi et al. [27] introduced a representation based on ResNet, which performs very well in classification problems, and design a Conditional Pairwise Clustering (ConPaC) algorithm, which estimates the adjacency matrix based on the similarities between faces. This algorithm expresses the clustering problem as a Conditional Random Field model and uses Loopy Belief Propagation to find an approximate solution for maximizing the posterior probability of the adjacency matrix.

Zhu et al. [28] introduced a nonlinear subspace clustering (NSC) algorithm for image clustering. This NSC algorithm exhibits the multi-cluster nonlinear structure of samples via a nonlinear neural network and achieves improved scalability and clustering accuracy more than kernel-based clustering methods.

Baoyuan et al. employ face tracking and a frame index to partition all faces from videos into multiple clusters. Once pairwise correlations between faces can be obtained from the temporal and spatial data, a clustering model based on hidden Markov random fields can be achieved [29,30].

Gokberk et al. [31] explore the identification problem for unconstrained TV data for face tracks. They use pairs of faces within a track as positive examples and pairs of face tracks of different people in the same frame as negative examples.

With advances in deep learning, numerous efforts have been made for face clustering, especially in videos. Zhang et al. [32] address the clustering problem by training a nonlinear metric function with a DCNN from the input image to a low-dimensional feature embedding. This network is trained to optimize the embedding space such that the Euclidean distances correspond to a measure of face similarity using an improved triplet loss function, which maximizes the distance between the negative pairs and minimizes the distance of the positive pairs.

Face clustering is a challenging problem that has been addressed only by a few studies given the vast amount of unlabeled face frames and the number of individuals in talk show videos. A globally accepted methodology or metric for face clustering does not exist. In talk show videos, the unknown number of public figures is large, which is difficult in terms of scalability (run-time). Additionally, the number of images per individual is unbalanced, which is challenging for a lot of previous clustering algorithms. We developed a DCNN-based system to address the problem of face clustering in talk show videos to achieve improved scalability and clustering accuracy.

3. Proposed System

The proposed system uses a DCNN model that we design and train from scratch. Figure 1 shows the architecture of our model, which follows the style of convolutions layers of Zeiler and Fergus [18] model. We additionally add $1 \times 1 \times$ d convolutional layers and max-pooling layers between the standard convolutional layers and, consequently, produce a model with 19 layers. In the training stage, our model extracts the face feature vector (embedding) and optimizes the training parameters based on a triplet loss objective function [33,34].

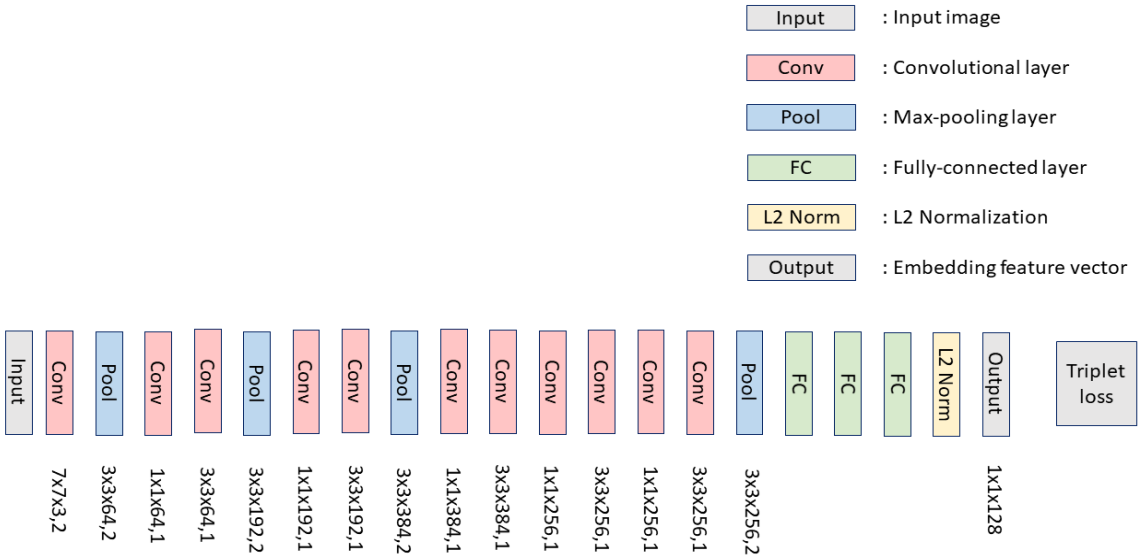

Figure 1. Architecture of our model.

The triplet method considers three inputs: f_a ,f_p, and f_n, which are the anchor, positive, and negative, respectively. The anchor image and positive image belong to the same person, while the negative image is obtained from a different person. The anchor and positive pair use different images within the same class, from which the triplet method trains the network to extract similar feature information. The anchor and negative pair have different classes, from which the method trains the network to extract different feature information.

The triplet loss function forces the network to produce a smaller distance between the anchor and the positive pair than the distance between the anchor and the negative pair with the margin α. Thus, we must define a proper distance measurement function to calculate the similarity between the anchor and the positive pair and the dissimilarity between the anchor and the negative pair.

Unlike [25,33,34], we employ Kullback–Leibler divergence (KLD) [35] in our DCNN model, which is also referred to as relative entropy, as a distance measurement between two probability distributions. The KLD can be expressed as follows:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} p(x) \log \left(\frac{P(x)}{Q(x)} \right). \quad (1)$$

KLD is usually employed as the loss function in neural network optimization. Thus, the triplet loss function that is being minimized is expressed as follows:

$$Loss = \sum_{i}^{N} max(\alpha + D_{KL}(f_{a_i} \| f_{p_i}) - D_{KL}(f_{a_i} \| f_{n_i}), 0). \quad (2)$$

Although the use of other distance measurements is feasible, most related studies applied the squared Euclidean distance as a metric. While we chose KLD in this paper, our model accepts other kinds of distance metrics, such as the Euclidean distance, squared Euclidean distance, total variance, and Wasserstein distance. While the comparison of the metric functions is not the focus of our primary research, anyone can explore the advantages of these choices.

In all experiments, we train the model using the stochastic gradient descent (SGD) and start the training with a learning rate of 0.05. Our proposed system consists of two main tasks: (1) the person spotting task and (2) the face Indexing and clustering task. Figure 2 illustrates the flow diagram of the

first task of the proposed system, while the second task is presented in Figure 3. The details of each task are discussed in the following sections.

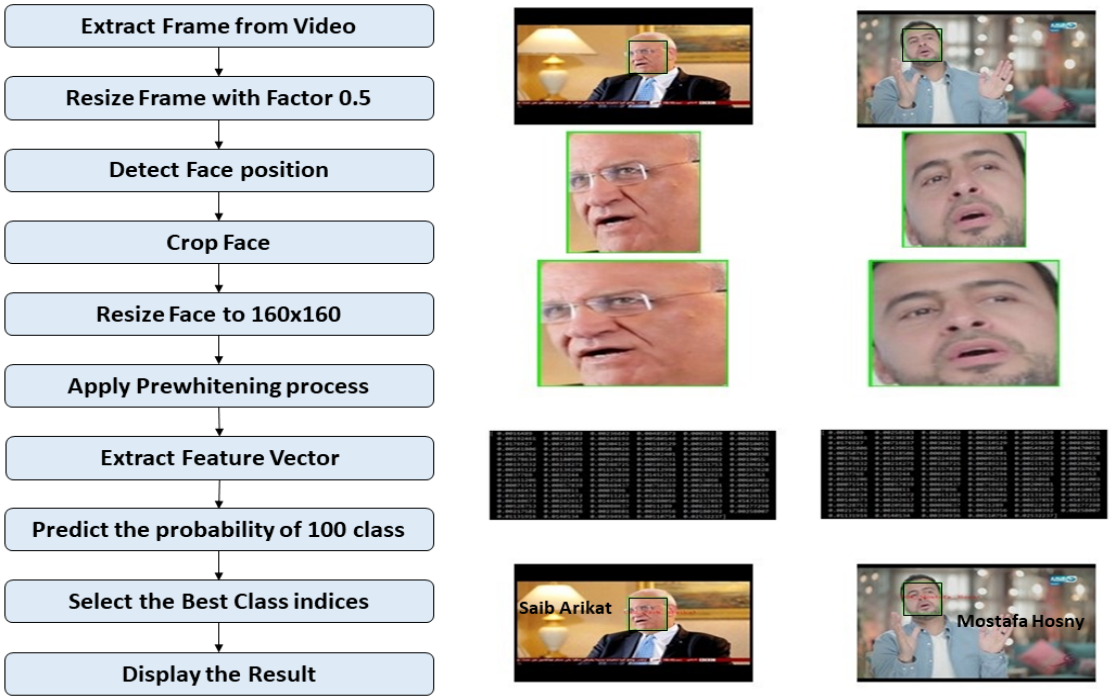

Figure 2. Flow diagram of our person spotting system.

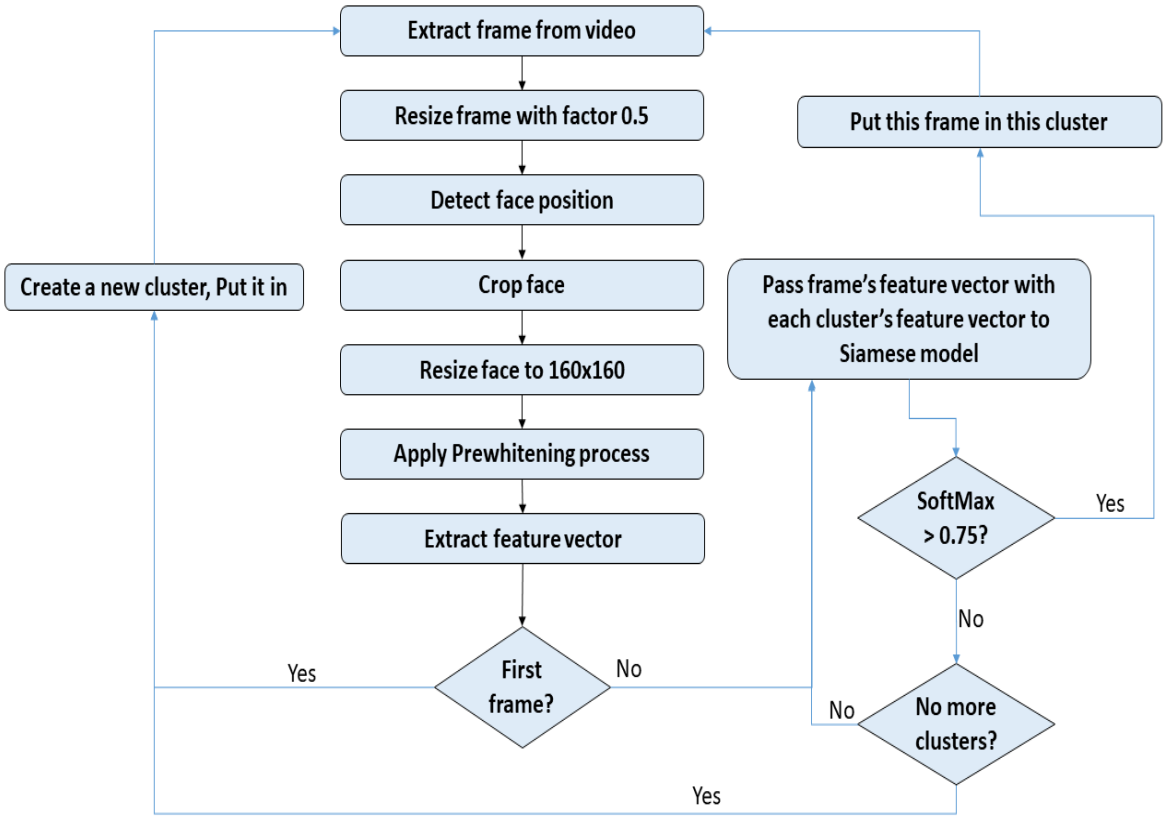

Figure 3. Flow diagram of our face clustering system.

3.1. Person Spotting

The aim of the person spotting task is to identify the appearance information of a specific individual in TV talk shows. The appearance information includes the duration of the appearance. Our person spotting (or face recognition) system consists of three main tasks: identify all faces in the image (face detection), analyze facial features (face encoding), and perform face recognition that is compared with known faces and make a prediction. The details of each task are discussed in the following sections.

3.1.1. Face Detection and Alignment

The first step in our system is face detection, in which the faces in a frame are located before they can be recognized. Face detection and alignment in an uncontrolled circumstance are challenging due to variations in illuminations, poses, and occlusions. A well-known face detection method [36] invented by Paul Viola and Michael Jones achieves good performance with real-time efficiency. A more reliable solution, which was invented by Dalal and Triggs, is the histogram of oriented gradients (HOG) [37]. However, these detectors may deteriorate the performance in real applications due to larger variations of human faces. Inspired by the excellent performance of CNNs in computer vision tasks, various CNN models have been proposed in the last decade [38], which can be used for face detection approaches.

Our face detection task utilizes a framework developed by Zhang et al. [39], which introduces a deep cascaded multitask approach that employs the inherent correlation to increase performance. The CNN framework consists of three stages: the first stage rapidly produces candidate windows via a shallow CNN. The second stage refines the windows to reject all non-faces windows via a more complex CNN. The final stage uses a more powerful CNN to refine the results and output facial landmark positions.

3.1.2. Face Encoding

The approach that we propose to address the face recognition problem is directly based on our proposed model. Our approach trains the proposed DCNN architecture to generate 128 measurements (feature vectors) for each face. Our model is trained by employing the triplet loss function that compares three face images at a time: two face images of the same person and one image of a completely different person. The algorithm analyzes the measurements that are generated from these three images and tweaks the CNN to ensure that the measurements for similar faces are closer and the measurements for different faces are farther away. After repeating this step millions of times on a vast amount of images of different people, the CNN learns to reliably generate 128 measurements for each person. This system employs the KLD distance metric, which measures the face similarity as distance.

3.1.3. Face Recognition (Identifying a Person's Name from Encoding)

In this last step, we identify the individual in the dataset of public figures who have the shortest similarity distance to the test image. We accomplish this task by training a simple linear SVM based classifier. The computation time for this classifier is in the range of milliseconds.

3.1.4. Face of Unknown Person

We also modified the developed recognition system to report unknown faces for persons who are not part of our dataset. The recognition system makes a decision for unknown faces based on a predefined threshold. However, we perform another experiment by adding a new class for the unknown faces and train the network with a mix of faces that differ from our known faces to increase the system accuracy.

The flow diagram of the person spotting task is shown in Figure 2. It starts by extracting a frame from a talk show video, resizes it to half the original size, detects the face rectangle and crops, and

resizes the rectangle to 160×160. We apply a pre-whitening technique to process the sequence of images to help effective training regardless of the illumination conditions of the images. Pre-whitening is also referred to as zero component analysis (ZCA). The pre-whitening technique subtracts the average from the pixels to normalize the range of the pixel values of input images, which tends to simplify the training. The model extracts the face embedding and passes it to a SVM based classifier to select the best class indices and display the results.

3.2. Face Clustering

In the previous subsection, we focused on face recognition as a methodology to spot known faces in the media. Although our classifier accurately recognized known faces, we cannot extend these classifiers to thousands or millions of faces. To build such a classifier, several hundreds of images of each unknown face should be available to train the classifier, which would not be possible in real time TV media. Some faces that appear in the media are not immediately recognized but are subsequently recognized. For these reasons, we addressed the face indexing problem differently from the person spotting problem.

Instead of building a classifier for known faces, we build a face verifier to differentiate between two faces and decide if they are the same or different. The face verifier is used to cluster all faces in a video into similar faces regardless of the face. This clustering process can be referred to as face indexing since it converts the long video into a set of folders that contain similar faces with their timestamp. When we are interested in a specific person, the face verifier can be used to determine when this person appears in the video by comparing its face image against different folders.

Our face verification system consists of three main tasks: face detection with alignment, face feature extraction, and face clustering (identify common people among these faces). The first two tasks (face detection and face feature extraction) are equivalent to the tasks for the person spotting system. Details of the final task (face clustering) are discussed in this section.

The final step in our face verification system is face clustering, which partitions the appearance of persons in TV talk show videos (identify common people among these faces). Face clustering groups a large number of faces into clusters in a way that minimizes the distance among faces within the same cluster while maximizing the distance among faces in different clusters.

Recent studies show that deep learning techniques have achieved notable performance improvement using a clustering task. Thus, our face clustering task utilizes a deep learning technique called one-shot learning with Siamese networks [40]. As humans, we can recognize the face of any person after meeting the person once, which is also the target goal of computer vision systems. A Siamese network consists of two identical neural networks, each of which takes one of two input images. The last layers of the two networks are then fed to a loss function, which calculates the similarity between the two input images (refer to Figure 4).

Siamese networks plays a central role for the tasks of identifying the similarity between two comparable images. In our study, we employ a deep convolutional neural network architecture to construct a Siamese network. We adopt the LFW dataset to train the Siamese network model, so it can differentiate between faces. We then test the network model with our talk show videos dataset. Since the model was not trained with any frame of our talk show video dataset, we refer to it as zero-shot learning. We employ a SoftMax loss function to calculate the similarity between two input feature vectors.

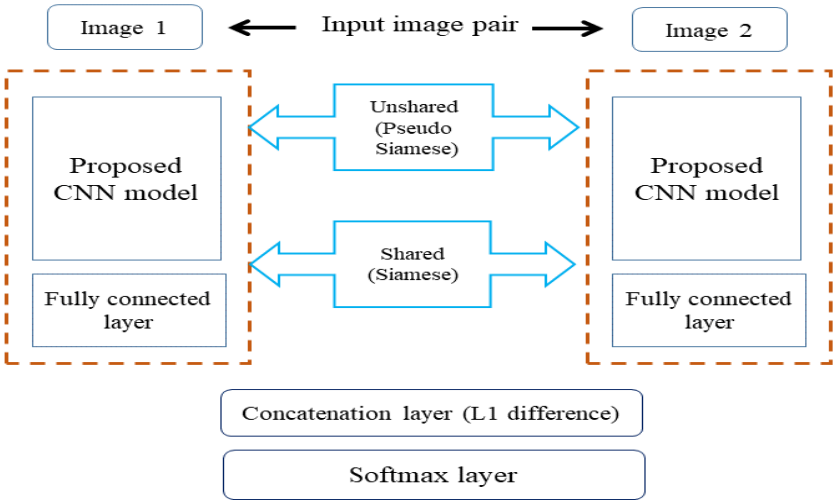

Figure 4. A Siamese network structure used by the proposed DCNN model.

The SoftMax loss function is expressed as follows:

$$S(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{K} \exp(y_j)}. \quad (3)$$

Here, $i = 1, \ldots, K$, which enables us to normalize the K-dimensional vector y of arbitrary real values into the K-dimensional probability vector S. As demonstrated in the next section, our developed face clustering task achieves superior accuracy over state-of-the-art techniques.

Figure 3 illustrates a flow diagram of the proposed face clustering system, where the primary functions of clustering start after extracting a face feature vector from our DCNN model. If the input frame is the first frame, its face feature vector is allocated to the first cluster. Otherwise, its feature vector (embedding) is compared with the existing clusters using the Siamese network. The face feature vector is allocated to the correlated cluster if they exhibit high similarity; otherwise, the face feature vector is allocated to a new cluster. The face clustering task involves thresholding the distance between the two feature vectors.

4. Experimental Results

In this section, we present the experimental results of the proposed person spotting and face clustering system. We have implemented, trained, and fine-tuned the proposed neural network models and algorithms using the TensorFlow framework. The person spotting and face clustering results have been verified with several datasets that were prepared in different environments. The datasets in our study are presented in this section.

4.1. Datasets

Datasets have a major role in advancing face recognition, clustering research, and deep learning techniques. Until recently, computer vision algorithms could not approach the performance of the human visual system because the available labeled image datasets were relatively not enough and so they did not represent the diversity of the real world. Face recognition and clustering approaches that are based on DCNN require a large volume of data and large face dataset for training. The greater the size of the dataset gets, the more efficient the training process can get and consequently the higher the recognition performance that can be achieved. Recently, these techniques employing very large-scale datasets achieved significant improvement in many recognition tasks.

In our study, therefore, we utilize an extensive collection of face datasets—Labeled Faces in the Wild (LFW) [41], which is a collection of unlabeled web face images for training, the YouTube

Faces Database [42], and our talk show faces dataset—to evaluate the developed face recognition and clustering framework.

Labeled Faces in the Wild (LFW): is a dataset of face images created to solve the problem of unconstrained face recognition. LFW contains 13,233 images of labeled faces of 5749 celebrities and public figures collected from the web.

YouTube Faces Database (YTF): is a dataset of face videos downloaded from YouTube. The YTF database contains 3425 videos of 1595 different persons and 621,126 frames that include the persons of interest.

Our dataset of unlabeled web face images: This dataset contains 2000 face images of 100 public figures, including 200 face images for each public figure. We constructed this dataset by collecting images of famous public figures from various TV channels.

Our Talk Show Video Dataset: We constructed a face dataset of talk show videos, which is a comprehensive dataset of labeled videos of faces in challenging unconstrained environments. This dataset contains recorded videos of 2160 h captured from TV channels including Sky-News, Al-Arabia, BBC Arabic, MBC-Masr, and Al-Nahar.

4.2. Evaluation of the Person Spotting System

In the person spotting task, we need to recognize the appearance of a specific person in TV talk shows and the duration of the appearance. Our person spotting (or face recognition) system consists of three steps: identify all faces in the image (face detection), analyze and encode facial features, perform face recognition, which is compared against known faces, and make a prediction.

We trained our proposed DCNN model using the LFW dataset for the face feature extraction process. We also trained our SVM classifier system for the face recognition process using 180 images from our dataset of unlabeled web face images and then tested it using 20 images for each person. For this experiment, we used a system with a 64-bit core i7 processor, 16 GB RAM and Nvidia GeForce 920M GPU (Shantou, China) which operates Linux Ubuntu 16.04. The result of our classifier is the name of the predicted person.

In the face detection and alignment step, we locate the faces in each frame and then attempt to recognize them. We observed that the developed face detection and alignment task produces a superior accuracy to the state-of-the-art methods while maintaining real-time performance. Figure 5 shows a sample of the original images and their detected faces.

Figure 5. Face detection process.

After extracting the face area in the detection process, we generate 128 features for this face. We pass this feature vector to our classifier to recognize the face. After confirming the accuracy of the

detection process, we computed the average accuracy of our face recognition process. The average person spotting accuracy was 99.6%, while the F1 measure was 0.996. Samples of the recognized faces are shown in Figure 6.

Figure 6. Face recognition process.

We also modified the developed system to report unknown faces for persons who are not part of our dataset of unlabeled web face images. We used a predefined threshold to determine unknown faces. After adding the unknown faces, the average accuracy decreased to 98.2% with an F1 measure of 0.98. To increase the accuracy, we conducted another experiment by adding a new class for the unknown faces and trained the network with the addition of faces that are differed from the 100 persons in our original database. The average recognition accuracy increased to 99.2%, with the F1 measure increased to 0.99.

Another experiment was performed to test our person spotting system using our talk show video dataset. The average recognition accuracy was 99.1%, while the F1 measure was 0.972.

4.3. Evaluation of Face Clustering System

We developed the proposed Face Clustering system based on a Siamese network and trained it using our dataset of unlabeled web face images. We then evaluated the system using the LFW dataset, the YouTube Faces database, and our talk show face dataset.

The developed Siamese network algorithm is capable of clustering unknown faces for persons who are not part of the training dataset. The proposed algorithm demonstrated an F-measure of 0.764 and 0.935 with the YouTube Faces database and the LFW dataset, respectively. The algorithm also exhibited an F-measure of 0.832 with our talk show face dataset.

The results of our proposed face clustering system are compared with the results obtained by classical methods (k-means clustering, spectral clustering and hierarchical clustering) and the results obtained with a recent deep learning clustering system developed by Otto et al. [43].

We applied K-means clustering with the Euclidean distance metric, spectral clustering from a graph theory point of view, and hierarchical clustering based on a Euclidean distance to cluster the LFW dataset. Although K-means, spectral, and hierarchical clustering are the most well-known clustering algorithms, they are implemented using a prior fixed number of clusters. Thus, we evaluated all algorithms with several effective number of clusters and recorded the best results.

Table 3 shows the clustering performance on the LFW dataset of k-means and spectral clustering algorithms with the number of clusters close to the actual number of individuals is very poor. K-means and spectral clustering algorithms can not handle unbalanced data. Therefore, the optimal value of the number of clusters in terms of clustering accuracy (F-measure) is relatively low (Approximately 150 clusters).

Table 3. Clustering accuracy on the LFW dataset based on the number of clusters.

Clustering Method	No. of Clusters	F-Measure	Run Time
K-Means	90	0.36	16 s
K-Means	5749	0.072	More than 6 h
Spectral	100	0.25	12 min, 23 s
Spectral	150	0.2	20 min, 18 s
Proposed system	5749	0.935	7 min, 21 s

Table 4 shows the F-measure of each system. The clustering method of [43] was developed using the same datasets—YouTube Faces database and LFW—as the datasets used for our proposed face clustering system.

Table 4. Comparison between proposed system and previously reported method.

Clustering Method	LFW	YouTube Face	Talk Show
K-Means	0.36	–	–
Spectral	0.2	–	–
Otto et al. [43]	0.87	0.71	0.750
Proposed system	0.935	0.764	0.832

When the classical methods (k-means, spectral clustering and hierarchical clustering) are tested with a pre-determined number of clusters similar to the actual number of identities, they produce very poor clustering performance. This is attributed to the fact that the classical methods are unable to handle a vast amount of unbalanced data and thus tend to produce a large number of wrong clusters.

Table 4 demonstrates that our proposed face clustering system significantly outperforms all the previous methods compared in this experiment in a number of clusters close to the actual number of individuals.

In terms of runtime measured using our hardware system described above, Table 3 shows that the k-means and spectral clustering algorithms take a large compute time even for 13k images in the LFW dataset, while our proposed clustering model is much faster. Figure 7 shows samples of four different clusters.

Figure 7. Samples of four different clusters.

If we neglect to confirm the accuracy of the face detection process, our clustering system can cluster these fault detection images into separate clusters (each image in a specific cluster), as shown in Figure 8.

Figure 8. Face clustering process with bad detection.

5. Conclusions and Future Research

In this paper, we proposed a DCNN-based system to address the problem of person spotting and face clustering in talk show videos. The main contribution of the proposed system is training the DCNN model, indexing, and clustering of video data including unknown faces. The target application is an effective media production analysis of faces in talk show videos and rapidly searches and identifies a specific person in the total video data in real-time processing.

We show that our triplet loss minimization method and Siamese network significantly enhance the average spotting and clustering performance. The proposed system shows the effectiveness of different datasets in different conditions.

The experimental results demonstrate that our proposed system remarkably outperforms the previous state-of-the art techniques across several challenging datasets regarding performance and real-time processing. This improvement is in part attributed to our triplet loss minimization method, which has been proved to be highly effective in extracting facial appearance features. In addition, the Siamese network, a key functional element of the proposed DCNN model, exhibited substantial contribution in representing the similarity of face images under large variations and different conditions.

For example, the implemented clustering system exhibited performance improvement: our F-measure values for the LFW, YouTube Faces, and our talk show faces datasets are 0.935, 0.764, and 0.832, respectively, whereas the recent deep learning clustering system [43] reported 0.87, 0.71, and 0.75, respectively—5.4%, 6.5%, and 8.2% improvement. Therefore, the proposed person spotting and clustering system can be an effective approach to analyzing massive TV production videos and identifying either known or unknown faces.

Author Contributions: Conceptualization, M.S.A.; methodology, M.S.A. and E.E.H.; software, M.S.A.; formal analysis, M.S.A.; investigation, M.S.A.; writing—original draft preparation, M.S.A.; writing—review and editing, M.S.A., M.E.R., H.K. and E.E.H.; supervision, M.E.R., H.K. and E.E.H.; project administration, M.E.R., H.K. and E.E.H.; funding acquisition, H.K.

Funding: This research was supported by the KIAT (Korea Institute for Advancement of Technology) grant funded by the Korea Government (MSS: Ministry of SMEs and Startups). (No. S2755555, HRD program for 2019) and in part by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as Global Frontier Project, South Korea, (CISS-2019). It was also supported by The Egyptian National Telecom Regulatory Authority (NTRA), Project: A Framework for Big Arabic Media Data Mining and Analytics) and The Engineering Company for the Development of Digital Systems (RDI), Cairo, Egypt.

Acknowledgments: The authors would like to thank the RDI Corporation for their assistance in acquiring the talk show video dataset in this study. We are grateful to thank Mohsen Rashwan, Sherif Abdou, Omar Nasr, and Galal for their support throughout the process of this research work.

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

References

1. Elahi, M.; Gharaee, M. Performance comparison for face recognition using PCA and DCT. *J. Electr. Electron. Eng.* **2015**, *3*, 62–65. [CrossRef]
2. Kumar, S.; Kaur, H. Face recognition techniques: Classification and comparisons. *Int. J. Inf. Technol. Knowl. Manag.* **2012**, *5*, 361–363.
3. Dai, B.; Zhang, D. Evaluation of face recognition techniques. In Proceedings of the PIAGENG 2009: Image Processing and Photonics for Agricultural Engineering, Zhangjiajie, China, 11–12 July 2009; International Society for Optics and Photonics: Bellingham WA, USA, 2009; Volume 7489, p. 74890M.
4. Zhao, M.; Teo, Y.W.; Liu, S.; Chua, T.S.; Jain, R. Automatic person annotation of family photo album. In Proceedings of the International Conference on Image and Video Retrieval, Tempe, AZ, USA, 13–15 July 2006; pp. 163–172.
5. Cui, J.; Wen, F.; Xiao, R.; Tian, Y.; Tang, X. Easyalbum: An interactive photo annotation system based on face clustering and re-ranking. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 28 April–3 May 2007; ACM: New York, NY, USA, 2007; pp. 367–376.
6. Fukushima, K. Neocognitron. *Scholarpedia* **2007**, *2*, 1717. doi:10.4249/scholarpedia.1717. [CrossRef]
7. Hubel, D.H.; Wiesel, T.N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **1959**, *148*, 574–591. [CrossRef]
8. Matsugu, M.; Mori, K.; Mitari, Y.; Kaneda, Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw.* **2003**, *16*, 555–559. [CrossRef]
9. Belalia, A.; Belloulata, K.; Kpalma, K. Region-based image retrieval in the compressed domain using shape-adaptive DCT. *Multimed. Tools Appl.* **2016**, *75*, 10175–10199. [CrossRef]
10. Belhallouche, L.; Belloulata, K.; Kpalma, K. A new approach to region based image retrieval using shape adaptive discrete wavelet transform. *Int. J. Image Graph. Signal Process.* **2016**, *1*, 1–14. [CrossRef]
11. Ho, J.; Yang, M.H.; Lim, J.; Lee, K.C.; Kriegman, D. Clustering appearances of objects under varying illumination conditions. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; p. 11.
12. Tian, Y.; Liu, W.; Xiao, R.; Wen, F.; Tang, X. A face annotation framework with partial clustering and interactive labeling. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
13. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
14. Aghdam, H.H.; Heravi, E.J. *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*, 1st ed.; Springer Publishing Company, Incorporated: Berlin/Heidelberg, Germany, 2017.
15. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
16. LeCun, Y.; Boser, B.E.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.E.; Jackel, L.D. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*; NIPS: Denver, CO, USA, 1989; pp. 396–404.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; NIPS: Denver, CO, USA, 2012; pp. 1097–1105.
18. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 818–833.
19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

20. Sun, Y.; Wang, X.; Tang, X. Deeply learned face representations are sparse, selective, and robust. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2892–2900.
21. Zhu, Z.; Ping, L.; Tang, X. Recover canonical-view faces in the wild with deep neural networks. *arXiv* **2014**, arXiv:1404.3543.
22. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Comprehensive Study on Center Loss for Deep Face Recognition. *Int. J. Comput. Vis.* **2019**, *127*, 668–683. [CrossRef]
23. Zhao, M.; Song, B.; Zhang, Y.; Qin, H. Face verification based on deep Bayesian convolutional neural network in unconstrained environment. *Signal Image Video Process.* **2018**, *12*, 819–826. [CrossRef]
24. Wang, H.; Hu, J.; Deng, W. Face Feature Extraction: A Complete Review. *IEEE Access* **2018**, *6*, 6001–6039. [CrossRef]
25. 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, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
26. Ding, G.; Khan, S.; Tang, Z.; Porikli, F. Feature mask network for person re-identification. *Pattern Recognit. Lett.* **2019**. [CrossRef]
27. Shi, Y.; Otto, C.; Jain, A.K. Face clustering: Representation and pairwise constraints. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 1626–1640. [CrossRef]
28. Zhu, W.; Lu, J.; Zhou, J. Nonlinear subspace clustering for image clustering. *Pattern Recognit. Lett.* **2018**, *107*, 131–136. [CrossRef]
29. Wu, B.; Lyu, S.; Hu, B.G.; Ji, Q. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2856–2863.
30. Wu, B.; Hu, B.G.; Ji, Q. A coupled hidden markov random field model for simultaneous face clustering and tracking in videos. *Pattern Recognit.* **2017**, *64*, 361–373. [CrossRef]
31. Cinbis, R.G.; Verbeek, J.; Schmid, C. Unsupervised metric learning for face identification in TV video. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1559–1566.
32. Zhang, S.; Gong, Y.; Wang, J. Deep metric learning with improved triplet loss for face clustering in videos. In *Pacific Rim Conference on Multimedia*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 497–508.
33. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
34. Hu, T.Y.; Chang, X.; Hauptmann, A.G. Learning Distributional Representation and Set Distance for Multi-shot Person Re-identification. *arXiv* **2018**, arXiv:1808.01119.
35. Wang, S.; Qian, Y.; Yu, K. Focal Kl-Divergence Based Dilated Convolutional Neural Networks for Co-Channel Speaker Identification. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5339–5343. [CrossRef]
36. Wang, Y.Q. An analysis of the Viola-Jones face detection algorithm. *Image Process. Line* **2014**, *4*, 128–148. [CrossRef]
37. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
38. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
39. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
40. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.
41. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*; Tech. Rep. 07-49; University of Massachusetts: Amherst, MA, USA, October 2007.

42. Wolf, L.; Hassner, T.; Maoz, I. Face recognition in unconstrained videos with matched background similarity. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 529–534.
43. Otto, C.; Wang, D.; Jain, A.K. Clustering millions of faces by identity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 289–303. [CrossRef] [PubMed]