International Journal of Advanced Multidisciplinary Research
ISSN: 2393-8870
www.ijarm.com
DOI: 10.22192/ijamr Volume 6, Issue 2 -2019

Research Article

DOI: http://dx.doi.org/10.22192/ijamr.2019.06.02.011

Video face recognition system for large scale person Re-identification using grassman algorithm

Dr. S. Saravanan, Dr. A.V. Prathapkumar, Dr. G. Ramprabhu

Dhanalakshmi Srinivasan Engineering College, Perambalur

Abstract

For face recognition in surveillance scenarios, identifying a person captured on image or video is one of the key tasks. This implies matching faces on both still images and video sequences. Automatic face recognition for still images with high quality can achieve satisfactory performance, but for video-based face recognition it is hard to attain similar levels of performance. Compared to still images face recognition, there are several disadvantages of video sequences. First, images captured by CCTV cameras are generally of poor quality. The noise level is higher, and images may be blurred due to movement or the subject being out of focus. Second, image resolution is normally lower for video sequences. If the subject is very far from the camera, the actual face image resolution can be as low as 64 by 64 pixels. Last, face image variations, such as illumination, expression, pose, occlusion, and motion, are more serious in video sequences. The approach can address the unbalanced distributions between still images and videos in a robust way by generating multiple "bridges" to connect the still images and video frames. So in this paper, we can implement still to video matching approach to match the images with videos using Grassmann manifold learning approach to know unknown matches. Finally provide voice alert at the time unknown matching in real time environments. And implement neural network classification algorithms to classify the face images in real time captured videos.

Keywords

Image Resolution,
Grassmann learning,
Real time environments,
Face detection

I. Introduction

Video signal is basically any sequence of time varying images. A still image is a spatial distribution of intensities that remain constant with time, whereas a time varying image has a spatial intensity distribution that varies with time. Video signal is treated as a series of images called frames. An illusion of continuous video is obtained by changing the frames in a faster manner which is generally termed as frame rate. The demand for digital video is increasing in areas such as video teleconferencing, multimedia authoring systems, education, and video-on-demand systems. Video indexing is necessary to facilitate efficient content-based retrieval and browsing of visual information stored in large multimedia databases. To create an efficient index, a set of representative key frames are selected which capture and encapsulate the entire video content.

In recent years, increasing attention has been paid to the video-based face recognition. Many approaches were proposed to use temporal information to enhance face recognition for videos. One direct approach is temporal voting. A still image-matching mechanism is proposed by Satoh for matching two video sequences. The distance between two videos is the minimum distance between two frames across two videos. And

presented a sequential importance sampling (SIS) method to incorporate temporal information in a video sequence for face recognition. A state space model with tracking state vector and recognizing identity variable was used to characterize the identity by integrating motion and identity information over time. However, this approach only considers identity consistency in temporal domain, and thus it may not work well when the face is partially occluded. Zhang and Martinez applied a weighted probabilistic approach on appearance face models to solve the occlusion problem. Their experiment shows that this approach can improve the performance for PCA, LDA, and ICA. The approach proposed uses the condensation algorithm to model the temporal structures. The video face recognition framework is shown in fig 1.

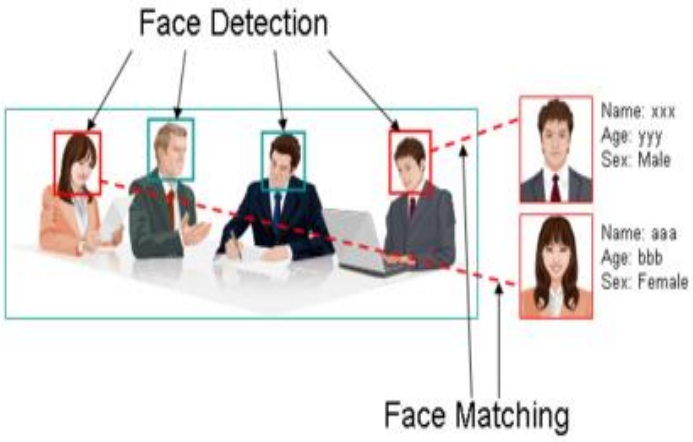

Fig 1. Face matching framework

II. Related Work

Y. Yan, et.al,…[1] proposed a novel active sample selection approach (a.k.a. active learning) for image classification by using web images. Previous research has shown that cross-media modeling of various media types is beneficial for multimedia content analysis. The web images are often associated with rich textual descriptions (e.g., surrounding texts, captions, etc). While such text information is not available in testing images, we show that text features are useful for learning robust classifiers, enabling better active learning performance of image classification. Typical active sampling methods only deal with one media type which cannot simultaneously utilize different media types. The new supervised learning paradigm, namely learning using privileged information (LUPI), can be used to solve this problem. In a LUPI scenario, in addition to main features, there is also privileged information available in the training procedure. Privileged information can only be used in training, and is not available in testing. Uncertainty sampling is the most frequently used strategy in the active learning. In this work, we propose to exploit both visual and text features for active sample selection by taking text as privileged information. By LUPI, we train SVMs on visual features and slacking function on text features. We present five strategies to combine the uncertainty measure of these two classifiers. To ensure the selected samples to be representative, we exploit the diversity measurement, such that the selected samples are less similar to each other. We formulate a ratio objective function to maximize cross-media uncertainty and minimize the similarity of selected data. Then we propose to measure uncertainty and diversity for training sample selection. A new optimization method is proposed to solve the proposed model, which automatically learns the optimal ratio of uncertainty to similarity. In this way, we avoid introducing the trade-off parameter between the two types of measurements.

Y. Yang, et.al,…[2] implemented a new feature selection algorithm, which leverages the knowledge from related multiple tasks to improve the performance of feature selection. In our study, the following lessons have been learned: Sharing information among related tasks is beneficial for supervised learning. However, if the multiple tasks are not correlated, the performance is not necessarily improved. Compared to single task learning, the advantages of multitask learning are usually more visible when we only have few training examples per task. As we increase the number of positive training data, the intra-task knowledge is sufficient for training, and thus adapting inter-task knowledge does not

necessarily help. It is not always the case that feature selection improves the performance. However it is still beneficial because it improves the efficiency. Also, feature selection would provide us with better interpretability of the features. The improvement of feature selection varies when different classifiers are used. For example, since linear SVM actually has the ability to assign different weights to different features, the performance improvement of SVM is less than KNN, after feature selection. The rest of this paper is organized as follows. We give the objective function. The optimization approach is proposed, followed by the proof of its convergence. We then show experimental results and conclude the paper. The performance improvement of feature selection varies when different classifiers are used. Taking KTH dataset as an example, we compare the performance of different feature selection algorithms when different classifiers are used. In particular, we use KNN as an alternative classifier. The average recognition accuracy of the six action types. Comparing Table and Figure, we can see that when using all features for action recognition, the accuracy of KNN is lower than that of SVM. However, after feature selection, KNN gains higher accuracy than SVM. One possible explanation is that SVM has the ability to weigh different features, and thus the benefit from feature selection is less.

X. Chang, et.al,…[3] aimed to solve the limitations of the existing discriminant analysis algorithms for high-order data and propose a compound rank-k projection algorithm for discriminant bilinear analysis. Different from, the convergence of our optimization approach is explicitly guaranteed. We adopt multiple orthogonal projection models to obtain more discriminant projection directions. In particular, we use h sets of projection matrices to find a low dimensional representation of the original data. The h projection matrices are orthogonal to each other. By doing so, we can project the original data into different orthogonal basis and information from various perspectives can be obtained. The key novelty of our method is that it adopts multiple projection models, which are integrated and work collaboratively. In this way, a larger search space is provided to find the optimal solution, which will yield better classification performance. We name the proposed algorithm as Compound Rank-k Projection for Bilinear Analysis (CRP). It is worthwhile noting that the algorithm can be readily extended to high-order tensor discriminant analysis. The main contributions of our work can be summarized as CRP can deal with matrix representations directly without converting them into vectors. Hence, spatial correlations within the original data can be preserved. Compared with the conventional algorithms, the computation complexity is reduced. Compared to the classical 2-dimensional linear discriminant analysis methods, CRP benefits from the trade-off between the degree of freedom and the avoidance of the over-fitting problem. Although the classical 2DLDA gains good performance, its iterative optimization algorithm may not converge due to the singularity of the between-class scatter matrix. Differently, the convergence of our algorithm is explicitly guaranteed. The rest of this paper is organized as summarizes an overview of the classical LDA as well as 2DLDA. A novel compound rank-k projection for bilinear analysis is proposed. We present our experimental results on five different datasets. The conclusion of our work is discussed.

J. Luo, et.al,…[4] proposed a framework consisting of two algorithms for multimedia content analysis and retrieval. First, a new transductive ranking algorithm, namely, ranking with Local Regression and Global Alignment (LRGA), is proposed. Differently from distance-based ranking methods, the distribution of the samples in the whole data set is exploited in LRGA. Compared with the inductive methods, only the query example is required. In contrast to the MR algorithm that directly adopts the Gaussian kernel to compute the Laplacian matrix , LRGA learns a Laplacian matrix for data ranking. For each data point, we employ a local linear regression model to predict the ranking scores of its neighboring points. In order to assign an optimal ranking score to each data point, we propose a unified objective function to globally align local linear regression models from all the data points. In retrieval applications, there is no ground truth to tune the parameters of ranking algorithms like MR. Therefore, it is meaningful to develop a new method that learns an optimal Laplacian matrix for data ranking. Second, we propose a semi-supervised learning algorithm for long-term RF. A system log is constructed to record the history RF information marked by all of the users. We refine the vector representation of multimedia data according to the log information via a statistical approach. To that end, we convert the RF information into pairwise constraints, which are classified into two groups. The data pairs in the first group are semantically similar to each other, while the data pairs in the second group are dissimilar to each other. While LDA can be used to exploit these two types of information, the valuable information in the unlabeled data is not utilized. In this paper, we propose a semi-supervised learning algorithm to refine the vector

representation by considering the history RF information as well as the multimedia data distribution of both labeled and unlabeled samples. In this paper, we evaluate the performance of our algorithms in content-based cross-media retrieval where the query example and retrieval results can be of different media types. For example, the user can search images either by an example image or an example audio record. We also apply the proposed algorithms to content based image retrieval and 3D motion/pose data retrieval. Extensive experiments demonstrate that our algorithms achieve better retrieval performance when compared with the existing related works.

W. Li, et.al,…[5] proposed a filter pairing neural network (FPNN) for person re-identification. This deep learning approach has several important strengths and novelties compared with existing works. It jointly handles misalignment, photometric and geometric transforms, occlusions and background clutter under a unified deep neural network. During training, all the key components are jointly optimized. Each component maximizes its strength when cooperating with others. Instead of using handcrafted features, it automatically learns optimal features for the task of person re-identification from data, together with the learning of photometric and geometric transforms. Two paired filters are applied to different camera views for feature extraction. The filter pairs encode photometric transforms. While existing works assume cross-view transforms to be uni-modal, the deep architecture and its maxout grouping layer allow modeling a mixture of complex transforms. Secondly, we train the proposed neural network with carefully designed training strategies including dropout, data augmentation, data balancing, and bootstrapping. These strategies address the problems of misdetection of patch correspondence, overfitting, and extreme unbalance of positive and negative training samples in this task. Thirdly, we re-examine the person re-identification problem and build a large scale dataset that can evaluate the effect introduced by automatic pedestrian detection. All the existing datasets are small in size, which makes it difficult for them to train a deep neural network. Our dataset has 13,164 images of 1,360 pedestrians; see a comparison .Existing datasets only provide manually cropped pedestrian images and assume perfect detection in evaluation protocols. As automatic detection in practice introduces large misalignment and may seriously affect the performance of existing methods. Our dataset provides both manually cropped images and automatically detected bounding boxes with a state-of the-art detector for comprehensive evaluation.

III. Existing methodologies

The term multi-view face recognition, in a strict sense, only refers to situations where multiple cameras acquire the subject (or scene) simultaneously and an algorithm collaboratively utilizes the acquired images/videos. But the term has frequently been used to recognize faces across pose variations. This ambiguity does not cause any problem for recognition with (still) images; a group of images simultaneously taken with multiple cameras and those taken with a single camera but at different view angles are equivalent as far as pose variations are concerned. However, in the case of video data, the two cases diverge. While a multi-camera system guarantees the acquisition of multi-view data at any moment, the chance of obtaining the equivalent data by using a single camera is unpredictable. Such differences become vital in non-cooperative recognition applications such as surveillance. For clarity, we shall call the multiple video sequences captured by synchronized cameras a multi-view video and the monocular video sequence captured when the subject changes pose, a single-view video. With the prevalence of camera networks, multi-view surveillance videos have become more and more common. Nonetheless, most existing multi-view video face recognition algorithms exploit single-view videos. Given a pair of face images to verify, they look up in the collection to “align” the face part’s appearance in one image to the same pose and illumination of the other image. This method will also require the poses and illumination conditions to be estimated for both face images. This “generic reference set” idea has also been used to develop the holistic matching algorithm, where the ranking of look-up results forms the basis of matching measure. There are also works which handles pose variations implicitly without estimating the pose explicitly.

IV. $proposed methodologies

Face detection is the first stage of a face recognition system. A lot of research has been done in this area, most of which is efficient and effective for still images only & could not be applied to video sequences directly. Face recognition in videos is an active topic in the field of image processing, computer vision and biometrics over many years. Compared with still face recognition videos contain more abundant information

than a single image so video contain spatio-temporal information. To improve the accuracy of face recognition in videos to get more robust and stable recognition can be achieved by fusing information of multi frames and temporal information and multi poses of faces in videos make it possible to explore shape information of face and combined into the framework of face recognition. The video-based recognition has more advantages over the image-based recognition. First, the temporal information of faces can be utilized to facilitate the recognition task. Secondly, more effective representations, such as face model or super-resolution images, can be obtained from the video sequence and used to improve recognition results. Finally, video- based recognition allows learning or updating the subject model over time to improve recognition results for future frames. So video based face recognition is also a very challenging problem, which suffers from following nuisance factors such as low quality facial images, scale variations, illumination changes, pose variations, Motion blur, and occlusions and so on. In the video scenes, human faces can have unlimited orientations and positions, so its detection is of a variety of challenges to researchers. In recent years, multi-camera networks have become increasingly common for biometric and surveillance systems. Multi view face recognition has become an active research area in recent years. In this paper, an approach for video-based face recognition in camera networks is proposed. Traditional approaches estimate the pose of the face explicitly. A robust feature for multi-view recognition that is insensitive to pose variations is proposed in this project. The proposed feature is developed using the spherical harmonic representation of the face, texture mapped onto a sphere. The texture map for the whole face is constructed by back-projecting the image intensity values from each of the views onto the surface of the spherical model. A particle filter is used to track the 3D location of the head using multi-view information. Videos provide an automatic and efficient way for feature extraction. In particular, self-occlusion of facial features, as the pose varies, raises fundamental challenges to designing robust face recognition algorithms. A promising approach to handle pose variations and its inherent challenges is the use of multi-view data. In video based face recognition, great success has been made by representing videos as linear subspaces, which typically lie in a special type of non-Euclidean space known as Grassmann manifold. To leverage the kernel-based methods developed for Euclidean space, several recent methods have been proposed to embed the Grassmann manifold into a high dimensional Hilbert space by exploiting the well-established Project Metric, which can approximate the Riemannian geometry of Grassmann manifold. Nevertheless, they inevitably introduce the drawbacks from traditional kernel-based methods such as implicit map and high computational cost to the Grassmann manifold. To overcome such limitations, we propose a novel method to learn the Projection Metric directly on Grassmann manifold rather than in Hilbert space. From the perspective of manifold learning, our method can be regarded as performing a geometry-aware dimensionality reduction from the original Grassmann manifold to a lower-dimensional, more discriminative Grassmann manifold where more favorable classification can be achieved. And also provide neural network classification algorithm to classify faces with improved accuracy. Finally provide voice based alert system with real time implementation. The proposed framework is shown in fig 2.

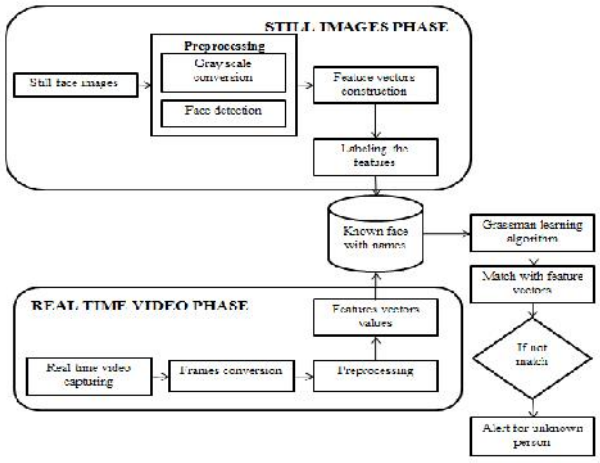

Fig 2. Proposed work

The algorithm as follows:

Grassmann algorithm:

Representing the data on Grassmann manifolds is popular in quite a few image and video recognition tasks. In order to enable deep learning on Grassmann manifolds, this paper proposes a deep network architecture which generalizes the Euclidean network paradigm to Grassmann manifolds. In particular, we design full rank mapping layers to transform input Grassmannian data into more desirable ones, exploit orthogonal re-normalization layers to normalize the resulting matrices, study projection pooling layers to reduce the model complexity in the Grassmannian context, and devise projection mapping layers to turn the resulting Grassmannian data into Euclidean forms for regular output layers. To train the deep network, we exploit a stochastic gradient descent setting on manifolds where the connection weights reside on, and study a matrix generalization of backpropagation to update the structured data. The popular applications of Grassmannian data motivate us to build a deep neural network architecture for Grassmannian representation learning. For this purpose, the new network architecture is designed to take Grassmannian data directly as input, and learns new favorable Grassmannian data that are able to improve the final visual tasks. In other words, the new network aims to deeply learn Grassmannian data on their underlying Riemannian manifolds in an endto-end learning architecture. To perform discriminant learning on Grassmann manifolds, many works embed the Grassmannian into a Euclidean space. This can be achieved either by tangent space approximation of the underlying manifold, or by exploiting a positive definite kernel function to embed the manifold into a reproducing kernel Hilbert space. In both of such two cases, any existing Euclidean technique can then be applied to the embedded data, since Hilbert spaces respect Euclidean geometry. For example, first embeds the Grassmannian into a high dimensional Hilbert space, and then applies traditional Fisher analysis method. Obviously, most of these methods are limited to the Mercer kernels and hence restricted to use only kernel based classifiers. Moreover, their computational complexity increases steeply with the number of training samples.

The Grassmann manifold G(m, D) is the set of m-dimensional linear subspaces of the R D. The G(m, D) is a m(D−m)-dimensional compact Riemannian manifold.

An element of G(m, D) can be represented by an orthonormal matrix Y of size D by m such that Y = Im, where Im is the m by m identity matrix. For example, Y can be the m basis vectors of a set of pictures in R D.

However, the matrix representation of a point in G(m, D) is not unique: two matrices Y1 and Y2 are considered the same if and only if span(Y1) = span(Y2), where span(Y) denotes the subspace spanned by the column vectors of Y . Equivalently, span(Y1) = span(Y2) if and only if Y1R1 = Y2R2 for some R1, R2 ∈ O(m). With this understanding, we will often use the notation Y when we actually mean its equivalence class span(Y), and use Y1 = Y2 when we mean span(Y1) = span(Y2), for simplicity.

Formally, the Riemannian distance between two subspaces is the length of the shortest geodesic connecting the two points on the Grassmann manifold. However, there is a more intuitive and computationally efficient way of defining the distances using the
principal angles

V. Conclusion

In this paper, we reviewed face recognition technique for still images and video sequences. Most of these existing approaches need well-aligned face images and only perform either still image face recognition or video-to video match. They are not suitable for face recognition under surveillance scenarios because of the following reasons: limitation in the number (around ten) of face images extracted from each video due to the large variation in pose and lighting change; no guarantee of the face image alignment resulted from the poor video quality, constraints in the resource for calculation influenced by the real time processing. We then proposed a local facial feature-based framework for still image and video-based face recognition under surveillance conditions. This framework is generic to be capable of still-to-still, still-to-video and video-to video matching in real-time. While the training process uses static images, the recognition task is performed over video sequences. Our results show that higher recognition rates are obtained when we use video sequences rather than statics – even when the algorithm using static images and that using video sequences address the same problems with exactly the same techniques. Evaluation of this approach is done for still image and video based face recognition on real time image datasets.

References

[1] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," IEEE Transactions on Multimedia, vol. 18, no. 12, pp. 2494–2502, 2016

[2] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," IEEE Transactions on Multimedia, vol. 15, no. 3, pp. 661–669, 2013.

Access this Article in Online

Website: www.ijarm.com

Subject: Information Technology

Quick Response Code

DOI:10.22192/ijamr.2019.06.02.011

How to cite this article:
S. Saravanan, A.V. Prathapkumar, G. Ramprabhu . (2019). Video face recognition system for large scale person Re-identification using grassman algorithm. Int. J. Adv. Multidiscip. Res. 6(2): 101-107.
DOI: http://dx.doi.org/10.22192/ijamr.2019.06.02.011