A Hierarchical Training and Identification Method using Gaussian Process Models for Face Recognition in Videos

Negar Hassanpour and Liang Chen
Computer Science Department, University of Northern British Columbia, Prince George, Canada

***Abstract*— In a video based face identification task, a sequence of frames can be utilized to identify the subject in the video. The information extracted from frames can provide samples of the subject in different head poses and facial expressions and under various lighting conditions which enriches the training process. However, some of these frames may not be useful for identification due to noise from various sources (such as occlusion, low resolution, and face tracking errors). It is important to reduce the effect of noisy samples by designing a representation structure that is capable of alleviating the noise in each sequence, complemented by developing a recognition procedure that rejects the wrong decisions affected by noise.**

In this paper we propose a sequence representation called Ensemble of Abstract Sequence Representatives (EASR) that is aimed at reducing the effect of noisy frames in a sequence. EASRs are used to guide the sampling process in a learning scheme called specialization – generalization which is used to train an ensemble of binary Gaussian Process (GP) models. Identification is done using: (i) the similarity between the EASRs of the gallery and probe images, and (ii) the label provided by the ensemble of GP classifier models. Evaluation of our approach on three publicly available benchmark datasets demonstrates significantly better performance compared to the state-of-the-art.

I. Introduction

A classic face identification problem involves identifying a subject from a single image and with only a few training samples available. For such configuration, these sample images must be carefully recorded in a controlled environment. However, in real-world applications, such good quality samples are not easily attainable. Fortunately, with the extensive availability of digital imaging devices, sufficient data is available to allow the recognition process to be based on image-set to image-set matching. In this sense, image-set based face identification in general and video based face identification in particular is potentially more promising than using single images. This type of face identification tends to be more robust since the recognizer gets to see more possible variations in appearance of the subject (e.g., different illumination, pose, facial expressions, etc.).

However, in video based face identification, we face a new challenge caused by uncertainty about the level of relevance of each frame which is due to the presence of noise in the frames of each video. Noisy frames in a sequence, similar to any other type of outlier, affect the accuracy of image-set based recognition in general, and methods that rely on individual samples in particular [23]. In this work we focus on dealing with the noise caused by low resolution, occlusion of the face, or failure of the face tracking algorithm to properly detect the face area in video.

We propose a vector set structure called Ensemble of Abstract Sequence Representatives (EASR) for representing a sequence. Each EASR is built by sampling and superposition to reduce noise, followed by a filtering mechanism to deal with outliers. Similar to the majority of image-set based approaches (e.g., [22], [24], [13], [10], [29]) that use a single structure to model each image-set, each EASR tries to model variations of the subject in an image-set and similarity of EASRs can be used for identification purposes. However, our method does not solely rely on EASRs for identification.

On top of the EASR representation method, we use an ensemble of binary GP models in a one-versus-rest setting for capturing the underlying non-linear structure of the data. To reduce the amount of noise presented to the GP models during the training, we use a sampling process called specialization – generalization. In the specialization step, the EASR similarity measure is used to find the nearest sequences of other subjects (the most challenging cases in the training set) and limit the sampling process to these sequences. In the generalization step, we attempt to reduce the effect of possibly noisy training samples by retraining the models on failed samples of the rest of the sequences in the training set. Finally, a fast identification process combines predictions of both methods to identify the subject in the probe sequence.

The main contribution of this paper is two-fold: First, we propose a representation structure for image-sets that minimizes the effect of noisy frames. This structure especially targets those frames that are not useful for the identification task, due to occlusion, low resolution, or failure of the face tracker algorithm. Second, we propose a learning scheme for tainting an ensemble of binary GP models for identification task in image-sets. This learning scheme selectively samples the training data to build the models efficiently and with minimum introduction of noise. Assessment of the proposed method on three publicly available benchmark datasets demonstrates better results compared to the previous methods [28], [7], [27], [3], [25], [22], [24], [13], [10], [4], [29] including state-of-the-art, especially on the more challenging YouTube Celebrities dataset.

The rest of this paper is organized as follows: First, we review the related work in the area of image-set based face identification. Then, we describe the components of our proposed approach. Next, the datasets and experimental setup used for evaluation of the work is explained. Later, the experimental results are discussed. Finally, the paper is concluded with a summary of the contributions and highlights on the future work.

II. Related Work

In the literature, the task of image-set based face identification is addressed in two steps: (i) representation of the image-sets, and (ii) finding a suitable similarity measure between them. Representation of the image-sets is either parametric or non-parametric. Parametric methods attempt to represent each image-set with a data-driven distribution function (e.g. Gaussian mixture model), and measure the similarity between them by calculating the between-set distribution distance (e.g. Kullback-Leibler divergence) [1]. Parametric methods, however, suffer from the assumption that all image-sets representing the same identity are drawn from the same distribution – which is most likely not the case. Therefore, the majority of current works use a non-parametric approach.

Representation method in non-parametric approaches can be either linear or non-linear. Most notable linear methods include: Mutual Subspace Method *MSM* [28] which constructs a linear subspace for each image-set and calculates the similarity from the Euclidean angle between the two subspaces; and Discriminant Canonical Correlations *DCC* [13] which finds an optimal discriminant function that transforms image-sets to another space in which the within-class canonical correlations are maximized while between-class canonical correlations are minimized.

Non-linear methods include: Kernel Grassmannian Distance *KGD* [25] which is a kernel generalization of the Grassmannian distance in order to capture the non-linear structures in the image-sets; Manifold Discriminant Analysis *MDA* [22] that forms the subspaces for each set with locally linear models (manifolds) and attempts to learn an embedding space, where each manifold is compact but manifolds of different classes are as separated as possible; and Manifold-Manifold Distance *MMD* [24] which formulates the recognition task as computation of distance between two locally linear subspaces of data, i.e., manifolds.

Measurement of similarity in non-parametric representations can be based on calculating the distance between representatives of the two image-sets, as in *MMD* [24]; and Affine/Convex Hull based image-set Distance *AHISD* and *CHISD* [3] where each image-set is represented by an affine/convex hull derived by spanning the subspace using the images in the set and similarity is the distance between closest exemplars. Similarity can also be measured based on the representation structure as a whole, as in *MSM* [28]; and Covariance Discriminant Learning method *CDL* [23] that represents the image-set by its covariance matrix second-order statistic, and formulated as a function which converts covariance matrix from Riemannian manifold to Euclidean space where similarity measurement is straightforward.

Sparse Approximated Nearest Point *SANP* [10] propose a similarity method that utilizes both the structural information of the image-sets, as well as their representatives. The kernel extension of this method *KSANP* allows for modelling the complex non-linear structures that are embedded in the data. As an improvement over the *SANP* method in terms of complexity reduction, Regularized Nearest Points *RNP* [29] models each image-set as a regularized affine hull and measures the similarity between two sets by calculating the distance between the nearest points between the two hulls.

Some recent methods have a holistic approach regarding the representation of each video sequence. For instance, Mean Sequence Sparse Representation-based Classification *MSSRC* [18] method performs a joint optimization to determine a linear relationship between all available training images.Joint Sparse Representation *JSR* [5] represents all the frames in a probe video sequence as an ensemble to suppress the effect of noise for a more stable recovery. Image-Set based Collaborative Representation and Classification *ISCRC* [30] models the probe video sequence as a convex or regularized hull and calculates the distance to the gallery considering the correlation between these two.

Gaussian Process (GP) models have been previously used for probabilistic object categorization. In [11], Kapoor et al. use GP confidence estimates at unlabelled data points in an active learning paradigm for interactive labelling. This active learning approach is of interest for datasets in which abundant unlabelled data is available, given that manual labelling is often expensive and/or time consuming in large datasets. To the best of our knowledge, GP has not been used for the task of video based face recognition before.

In the next section we describe our method which in part uses GP regression models for face identification in videos.

III. Proposed Method

Our proposed method consists of two main modules: EASR module and GP module. EASR module is based on a vector-based representation of the image-sets called Ensemble of Abstract Sequence Representatives (EASR), that offers better resistance to noisy data; along with a method for similarity measurement between EASRs. The GP module incorporates a learning scheme called specialization – generalization for effective training of an ensemble of binary GP classifiers (enabling further noise reduction). The identification process combines both modules using a hierarchical structure to maximize identification rate. The rest of this section describes each module in more details.

A. Ensemble of Abstract Sequence Representatives

In image-set based face identification, each image may not fully characterize the individual's face. This may be due to (i) poor quality of the image (e.g. low resolution, illumination, etc.), (ii) partial existence of the face in the image's field of view (e.g. occlusion, pose, etc.), and (iii) failure of the face detector algorithm to accurately spot the face. Such issues would cast uncertainty on the degree that a face identification method should rely on each individual image in an image-set (for images in both gallery and probe sets).

We propose a vector-based representation for each image-set called Ensemble of Abstract Sequence Representatives (EASR) that addresses the uncertainties mentioned above as follows: it relaxes the noise in the raw data points by transferring them into a higher level representation structure using stratified sampling and superposition. Furthermore, the

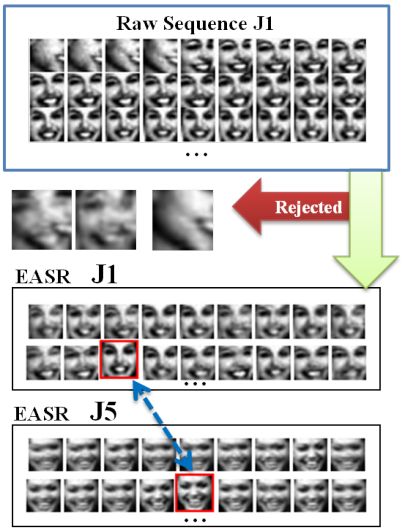

Fig. 1. A sample sequence from the YouTube Celebrities dataset [12] (top), its EASR based on intensity features (middle), and a matched EASR from another clip (bottom)

construction process of EASRs is supplemented with an outlier filtering scheme to counter face tracking errors. More interestingly, this representation structure can help construct GPs that are specialized in discriminating the correct class versus the most similar ones to it. This will be further discussed in section III-C.

In the rest of this section, we first explain the representation structure of EASR, and then discuss the similarity measurement between each EASR that is necessary for either performing the identification task, or ranking the classes.

1) Representation: A video sequence (top of Fig. 1) can be represented as a set of normalized n dimensional feature vectors (each referred to as α) extracted from every frame. However, because such primary representation is prone to noise, it is beneficial to transform these vectors into a noise-relaxed secondary representation structure. Using stratified sampling, we draw (with replacement) a set of α vectors from each stratum (i.e., video sequence), which are then grouped into several non-overlapping subsets of size m (i.e., m vectors per subset). Then, for each subset, a new feature vector (represented by β) is constructed using (1).

$$\beta = \frac{\gamma}{\|\gamma\|} \quad , \quad \gamma = \sum_{i=1}^{m} \alpha_i \tag{1}$$

We refer to these new n dimensional unit feature vectors as Abstract Sequence Representatives (ASRs). Superposition leads to constructing more robust samples by minimizing the effect of undesired variations in single noisy images and therefore actively improving the recognition rate.

For each sequence we construct a set of ASRs of size M, and refer to it as Ensemble of Abstract Sequence Representatives (EASR). Top of Fig. 1 shows the first 27 frames of a raw sequence labelled J1. In the middle of Fig. 1 a subset of ASRs forming the J1's EASR is presented. The idea of utilizing ensembles is close to the concept of exploiting the knowledge of the crowd in decision trees and random forests [2] that are known for their robust performance in noisy data.

The pair-wise similarity between two ASRs β_p and β_q (i.e., ψ_{pq}) is calculated as the inner product of the two ASRs, as in (2),

$$\psi_{pq} = \langle \beta_p, \beta_q \rangle = |\beta_p||\beta_q| \cos(\theta_{pq}) = \cos(\theta_{pq}) \tag{2}$$

where θ_{pq} is the angle between the two unit vectors β_p and β_q. All the sequence-wise similarity values ψ_{pq} derived by (2) are collected in matrix Ψ that is an $M \times M$ matrix.

It is a good practice to monitor the quality of ASRs that are generated by random sub-sampling. For each ASR β_p, we calculate its mean pair-wise similarity $\bar{\Psi}_p$ by averaging over row p of Ψ. We then calculate average $(\bar{\Psi})$ and standard deviation $(\sigma_{\bar{\Psi}})$ of these $\bar{\Psi}_p$s for $p \in [1..M]$. Finally, we filter out the possible outliers, i.e., any ASR β_o with an average pair-wise similarity $(\bar{\Psi}_o)$ that is two standard deviations $(\sigma_{\bar{\Psi}})$ less than the average within-ensemble similarity $(\bar{\Psi})$.

$$\text{Reject}(\beta_o | \bar{\Psi}_o < \bar{\Psi} - 2\sigma_{\bar{\Psi}}) \tag{3}$$

We identified two sources for generating outlier ASRs: (i) superposition of frames that present the subject in highly different conditions; and (ii) presence of noisy frames (e.g., where the face tracking failed) in the ASR. Three sample ASRs that were rejected in the process of constructing the EASR for the J1 sequence are shown in Fig. 1. The two rejected ASRs on the left are generated due to source (i), while source (ii) is behind rejection of the third ASR (face tracker failure on a number of frames).

2) Similarity Measurement: In order to find the similarity between two sequences i and j (denoted as S_{ij}), first, we find the similarity ψ_{xy}^{ij} between all possible ASR pairs of the form (β_x^i, β_y^j) where β_x^i is the x^{th} ASR from the EASR of sequence i and β_y^j is the y^{th} ASR from sequence j's ensemble, following (2). This yields the Ψ^{ij} matrix, where $\Psi^{ij} = [\psi_{xy}^{ij}]$, for $x, y \in [1..M]$. The nearest ASR pair of the two sequences i and j (i.e., the β_x^i and β_y^j with the maximum ψ_{xy}^{ij} among all pairs) determines the similarity measure S_{ij}.

$$S_{ij} = \text{Max}\ \{\Psi_{ij}\} \tag{4}$$

Following our illustrated example, the bottom of Fig. 1 shows the EASR for a sequence labelled as J5, which represents the same subject as in sequence J1 but from another clip of hers. The similarity between these two EASRs is measured by similarity of their closest pair of ASRs as highlighted in Fig. 1, calculated via (4).

Finally, predicting the identity of a probe video sequence is posed as finding the most similar EASR l in the gallery to the probe EASR p, where $l \in [1..L]$ with L being the total number of subjects in gallery. We form the matrix $S^p = [S_{lp}]$ for $l \in [1..L]$, and report the identity of the subject with the highest similarity: $identity = \text{ArgMax}\ \{S^p\}$.

Although we designed EASRs to be resilient to noise, their inherent linear structure does not allow for capturing large and complex variations in data. Thus, we add a second non-linear component (i.e., binary Gaussian process models) to address this issue. In the next section, we provide a brief

review on GP regression and describe how it is employed to perform the task of video based face identification.

B. Gaussian Process Models

A Gaussian process is a generalization of the Gaussian probability distribution and is a Bayesian alternative to the kernel methods such as Support Vector Machines. Since models learned by GP are non-parametric [19], any hard assumptions on the structure of the model are safely avoided (e.g. assuming all data points are drawn from the same distribution, i.e., parametric models described in section II). In this section, we briefly discuss GPs for regression and classification following the notation used by Murphy [16].

Given a set of labelled samples $X = \{x_1, x_2, ..., x_N\}$, where each x_i represents a feature vector, and observed class labels $y = \{y_1, y_2, ..., y_N\}$, we are interested in classifying a set of unlabelled samples X_*. Gaussian process regression solution assumes a latent function $f(x)$ exists such that $y = f(x) + \epsilon$, where $\epsilon \sim (0, \sigma_y^2)$ links the observed label y to hidden label $f(x)$ via a Gaussian noise model.

GP assumes that $p(f|X) = p(f(x_1), ..., f(x_N))$ is jointly Gaussian, with mean $m(x) = \mathbb{E}[f(x)]$ and covariance $k(x_i, x_j) = \mathbb{E}[(f(x_i) - m(x_i))(f(x_j) - m(x_j))^T]$. In more abstract terms, $f \sim \mathcal{N}(\mu, K)$, where, $\mu = (m(x_1), ..., m(x_N))$, and $K_{ij} = k(x_i, x_j)$ that is a positive definite kernel function, defined based on our prior beliefs over the kinds of functions we expect to observe in data (e.g. level of smoothness [1]). In this work, we use a radial basis function (RBF) kernel that is in form of $k(x_i, x_j) = \sigma_f^2 \exp(-\frac{1}{2l^2}(x_i - x_j)^2)$. Parameters σ_f and l are optimized based on cross-validation over the training data.

Now, in case of observing a new set of data samples X_* GP needs to predict f_*. For the sake of simplicity and without loss of generality, let us assume μ and μ_* are 0. By definition, the joint distribution is updated to:

$$\begin{pmatrix} y \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K_y & K_* \\ K_*^T & K_{**} \end{pmatrix}\right) \tag{5}$$

where $K_y = k(X, X) + \sigma_y^2 I_N$ is $N \times N$, $K = k(X, X_*)$ is $N \times N_*$, and $K = k(X_*, X_*)$ is $N_* \times N_*$.

The goal is to compute posterior $p(f_*|X_*, X, y)$ which has the following form:

$$\begin{aligned} p(f_*|X_*, X, y) &= \mathcal{N}(f_*|\mu_*, \Sigma_*) \\ \mu_* &= \mu(X_*) + K_*^T K_y^{-1} y \\ \Sigma_* &= K_{**} - K_*^T K_y^{-1} K_* \end{aligned} \tag{6}$$

derived by applying the rules for conditioning Gaussian distributions. We use Cholesky decomposition (suggested by [19]) to compute $K_y^{-1} = L^{-T} L^{-1}$ instead of direct inversion of the matrix to avoid numerical stability issues.

To summarize, for each new unlabelled sample x_*, GP regressor described above gives a mean μ_* that is the expected output of the function predicted by GP at this point, accompanied by a variance σ_*^2 which demonstrates the GP's confidence on its prediction.

[1]Kernel function $k(x_i, x_j)$ controls the relativeness of points x_i and x_j, i.e., if the kernel considers x_i and x_j as similar, then output of the function at those points is expected to be similar as well.

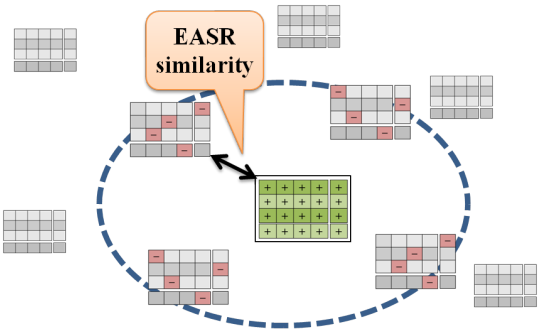

Fig. 2. The specialization step for $k = 4$ (best viewed in color)

In the current work, we use the GP regressor to construct a GP binary classifier[2] (i.e., $y_i \in \{-1, +1\}$). For each subject i, this classifier is capable of identifying the subject i versus the rest of the subjects. Let us assume subject i has m_i samples in total for training. We label these samples as $(+1)$. In order to collect the (-1) labelled samples, we sub-sample equal number of data points from the training set belonging to the rest of subjects such that the total number of samples for training is closest possible to $2m_i$ (i.e., balanced).

For each subject i in a gallery with L subjects, we construct one GP model GP^i. This model is trained to predict whether a sequence belongs to the subject i or not. To predict the identity of a probe video sequence p with m_p frames, the frames are presented to all L models. Each GP^i predicts the μ_* that is a vector of m_p length, where the j^{th} item shows the expected value of GP^i's underlying function (f_*) with the j^{th} frame as input. Classification of each frame is based on the sign of μ_*, if it is negative, it means GP^i rejects the possibility that this frame belongs to the subject i and vice versa. In order to aggregate the outputs of all m_p frames, we calculate the average of all f_*^j, $j \in [1..L]$ and record it as the overall output of GP^i (i.e., sum-fusion). After calculating the aggregated output for every model, identity of the subject with the highest aggregated output is reported as the predicted identity by the GP ensemble.

C. Specialization – Generalization Learning Scheme

GP binary classifiers are sensitive to the quality of training samples, thus a simple random sampling process without any provision for avoiding noisy samples reduces the identification power of the resulting model. In this section, we describe our learning scheme which relies on EASRs for finding the most relevant sequences for training each binary GP model (i.e., specialization step, schematically shown in Fig. 2), complemented by a generalization step which tries to alleviate the effect of potentially noisy frames in the training samples.

Starting with n subjects and m sequences for each subject in the training data, we have a $SQ_{n \times m}$ which contains sequences for each subject.

[2]Note that this satisfies the assumption of $\mu = 0$ as mentioned in the previous section, and therefore (6) holds true.

Fig. 3. Sample noisy frames detected in the generalization step when training a GP model for the sequence J1 in Fig. 1

Specialization step (Fig. 2):

1) Calculate EASRs for all training sequences.
2) Calculate the pair-wise similarity S_{ij} between each two subjects i and j, following (4).
3) For each subject i find the top k nearest subjects with highest S_{ij} and store js in NS_i.
4) Train GP$_i$ with all frames from $SQ_i, [1..m]$ as $(+1)$ instances and randomly sample equal number of frames from $SQ_{j,[1..m]}, j \in NS_i$ as (-1) instances.

Generalization step:

5) Use GP$_i$ to label each sequence in $SQ_{j,[1..m]}, j \notin NS_i$, for each frame f if GP$_i(f) > 0$ (i.e., mislabelled) add it to $GenL_i$ list to be retrained to GP$_i$.
6) Update GP$_i$ with all frames f in $GenL_i$ as (-1) instances.

In the specialization step, for each GP model, frames of the training sequences for the target identity are used as $(+1)$ instances. The (-1) instances are randomly sub-sampled from the sequences belonging to the k nearest subjects to the target identity, as determined by the EASR similarity (Fig. 2). The goal of specialization step is to force GP to learn distinctive features that separate each subject from its nearest neighbours.

The generalization step provides more (-1) instances to the GP model in areas of the problem space that the model is unable to correctly identify such instances, thus improving the generalizability of the GP model. The generalization step also minimizes the effect of noisy frames in the $(+1)$ instances. For example, consider the first 3 frames of sequence J1 shown at the top of Fig. 1. These frames do not provide any useful information for identifying the subject in that video. In the initial training of the GP model for sequence J1, these frames are provided as $(+1)$ instances, which misleads the model to classify any similar noisy frame as $(+1)$. In the generalization step, such noisy frames belonging to the rest of training sequences are detected (Fig. 3 shows a selection of these detected frames when training the model for J1). These frames are then used as new (-1) instances to update the GP model. This process helps to cancel out the effect of noisy frames in the $(+1)$ instances.

Now that the GP models have been constructed consulting the EASR suggestions, the next stage is to make predictions based on the models.

D. Identification Process: a Hierarchical Approach

In this section, we discuss our proposed hierarchical approach for aggregating the predictions of the two modules, namely EASR module and GP module, to come up with the most accurate prediction of identity for a probe video sequence. Fig. 4-left illustrates the flowchart of the proposed approach.

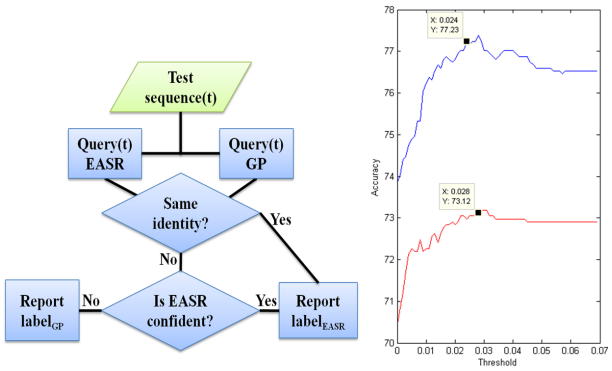

Fig. 4. Flowchart for the identification process (left); Exploring the effect of minimum cut-off for EASR confidence (τ) on accuracy (right)

Clearly, if predictions of both modules agree on the same identity, that prediction is reported. Otherwise, the hierarchical approach is used as follows: First, we give priority to the EASR module since it is more noise tolerant. We trust EASR-based prediction when it identifies the probe video sequence p by a clear winner; that is, when difference in similarity of p and the winner S_{pw} versus p and the closest next candidate S_{pc} as derived by (4) is above a pre-defined threshold τ. If the constraint for the cut-off is not satisfied, it indicates that the EASR module is not confident in its prediction, therefore the label generated by the GP module is reported as the final predicted identity.

Fig. 4-right shows the overall accuracy of our method for different values of τ in two different experiments (described later). The left side of Fig. 4 shows that accuracy drops when we rely too much on the EASR module (low τ values). On the other hand, relying too much on the GP module has the same effect. The value of τ for each dataset, is selected based on cross-validation over the training set (the selected points are highlighted in the graphs in Fig. 4-right). As similarity of two EASRs falls between zero and one, $max(\tau) = 1$. However, in our experiments, τ is much smaller (always less than 0.05).

IV. Experiments

In this section, we briefly describe the datasets and evaluation settings for our experiments.

A. Datasets

Three publicly available benchmark datasets were used for evaluation: Honda/UCSD [14], CMU-MoBo [9], and the more challenging YouTube Celebrities [12] datasets.

Honda/UCSD dataset is a collection of 59 videos recorded from 20 subjects in order to form a common ground for assessment of different face identification algorithms. Each subject has at least 2 videos (except for one subject).

CMU-MoBo dataset contains video sequences of 25 subjects performing four different walking activities on a treadmill. Following the literature, the subject with fewer than four walking patterns is excluded from the dataset, thus only the first 24 subjects are used.

YouTube Celebrities dataset (YTC) is a collection of real-world videos from YouTube website featuring 47 celebrities. The videos are noisy, low resolution, and demonstrate large variations in illumination, pose, expression, and other uncontrolled conditions. For each subject there are 3 video clips, where each clip is divided into several sequences of unequal resolution and duration. There is a total number of 1910 sequences, all encoded in MPEG4 at 25fps rate.

B. Evaluation Settings

In this section, we describe the procedure for preparation of the training and testing data. We followed the common settings used in the literature to allow for fair comparison.

Face tracking: It is a common practice to first track and crop faces from each frame and only pass the subjects' faces to the recognizer. Similar to the related work, Viola-Jones method [21] is used for extracting faces in the Honda/UCSD and CMU-MoBo[3] datasets. For the YTC dataset, the Viola-Jones algorithm fails to detect faces in a number of sequences. Thus, following Hu et al. [10] we use the Incremental learning for Visual Tracking (IVT) algorithm [20]. IVT returns the face area in all frames of all 1910 sequences, however, some may not represent a correct face (see Fig. 3).

Resolution: All the cropped faces are resized to an equal resolution. Images in Honda/UCSD dataset are resized to 20×20 pixels, CMU-MoBo to 40×40 pixels, and YTC dataset to 20×20 pixels (20×20 resolution was selected to reduce the computational cost).

Features: To experiment with different feature types, we use histogram equalized intensity levels for the Honda/UCSD dataset, Local Binary Pattern (LBP) codes [17] for the CMU-MoBo dataset, and Histogram of Oriented Gradient (HOG) descriptors [6] for the YTC dataset.

Train/Test image-set arrangement: For the Honda/UCSD dataset we randomly select 20 sequences (one video per subject) for training and the rest for testing. It should be noted that, there is an alternative evaluation setting for the Honda/UCSD dataset, which uses a *predefined* set of 20 sequences (one video per subject) for training without any random permutations. Since recent algorithms (e.g., RNP and ISCRC) achieve 100% accuracy with this predefined setting, we use the random setting which provides more variation in order to have a more meaningful comparison. For the CMU-MoBo dataset we also randomly select 24 sequences, one video per subject for training and the rest for testing.

For the YTC dataset we perform 5-fold cross-validation, following the evaluation protocol used by Hu et al. [10]. Sequences of each subject are sequentially partitioned (no prior shuffling) into 5 folds, where each fold contains exactly 9 sequences (from 3 clips) with minimal overlap between folds. In each fold, 1 clip is randomly selected as the train data (3 sequences) and the other 2 clips are used as the test data (6 sequences). It is important to mention that there is another evaluation setting for the YTC dataset first used by Wang et al. [24]. In this setting, for every subject in each fold, 9 sequences (3 per clip) is randomly selected; 3 sequences (1 per clip) for training, and the rest for testing. It is easier for different methods to identify the subject with the second setting, because there is one sequence from each clip in the training set, which factors out differences in appearance of the subject in different clips. For this reason, we believe that the first setting is closer to real world scenarios, thus we adopted the protocol used by Hu et al. in [10]. For all three datasets we report accuracy results for the full length sequences as well as truncated sequences that only contain the first 50 frames of the sequence. All evaluations are done using 5-fold cross-validation.

Comparisons: We compare the identification rate of our proposed method against several relevant image-set based methods proposed in the recent years (namely, MSM, MDA, AHISD/CHISD, SANP, RNP, MSSRC, JSR, and ISCRC in chronological order). Except for JSR, for all other methods we used the code provided by the authors adjusted with their suggested parameter values. For JSR we did not have access to the code thus report the results provided by the authors. However, it should be noted that the evaluation settings for JSR are different than what we are using in this paper – they used the Wang et al. setting for the YTC dataset, and 30×30 resolution for both YTC and CMU-MoBo datasets.

To make the comparisons fair, we used the same feature type for training all algorithms (i.e., intensity levels for Honda/UCSD, LBP for CMU-MoBo, and HOG for YTC). Interestingly, this enhancement led to improved accuracy for all algorithms (including the older algorithms such as SANP) on the YTC dataset compared to the results reported in the original papers. Also, it must be noted that the original evaluation of RNP was done only on 29 subjects for the YTC dataset, and the results obtained in [29] are higher than the results obtained on the full dataset. Additionally, MSSRC method comes with its own face tracking algorithm which was disabled here, since the aim of this paper is to compare the recognition power of different algorithms, therefore, we use the same tracking algorithm in all evaluations.

V. Results and Discussion

In this section, we summarize the results of the experiments described above. First, we present and discuss the performance results in terms of identification rate for the proposed method (EASR+GP) as well as the most successful methods in the literature (as listed above). Then, we provide running time comparison for EASR+GP vs. the other methods. Performance results on each of the three benchmark datasets is derived by exactly following the protocol described in the Evaluation Settings section. This protocol is the same as that in the related work.We perform Welch's t-test [26] to check whether the improvement in performance of the proposed method is statistically significant compared to the best performance of the contender methods. Outcomes of

[3]In this work, we have directly used the pre-processed version of CMU-Mobo dataset provided by the authors of [3]. The pre-processing procedure includes face tracking, resolution, and feature extraction.

TABLE I

IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON THREE DATASETS (MEAN ± STANDARD DEVIATION)

Dataset		**Honda/UCSD**		**CMU-MoBo**		**YouTube Celebrities**	
Method	**Year**	50	All	50	All	50	All
MSM	1998	87.69 ± 6.12	90.26 ± 2.15	92.50 ± 2.71	97.22 ± 1.70	70.57 ± 5.33	65.82 ± 4.56
MDA	2009	87.69 ± 2.81	96.41 ± 1.40	84.17 ± 6.56	95.28 ± 2.88	64.26 ± 3.76	69.22 ± 4.90
AHISD	2010	88.21 ± 3.89	83.59 ± 3.89	92.50 ± 2.71	95.56 ± 2.48	69.43 ± 4.16	63.83 ± 3.24
CHISD	2010	86.15 ± 2.92	91.28 ± 2.29	92.50 ± 2.71	98.61 ± 1.39	67.73 ± 5.09	69.65 ± 4.59
SANP	2011	87.18 ± 6.01	96.41 ± 2.29	92.50 ± 2.71	99.17 ± 0.76	67.59 ± 5.71	73.40 ± 3.18
RNP	2013	88.21 ± 4.66	93.33 ± 2.29	92.50 ± 2.71	98.33 ± 1.16	69.50 ± 5.30	73.48 ± 3.65
MSSRC	2013	91.28 ± 1.40	93.85 ± 2.29	91.11 ± 3.49	98.33 ± 1.52	70.78 ± 3.48	72.20 ± 3.52
ISCRC	2014	92.31 ± 4.44	95.38 ± 1.15	$\mathbf{94.44 \pm 2.20}^{\dagger}$	$\mathbf{99.44 \pm 0.76}^{\dagger}$	66.38 ± 4.73	70.71 ± 3.14
EASR+GP		$\mathbf{94.87 \pm 4.05}$	$\mathbf{99.49 \pm 1.15}^{\star}$	$\mathbf{93.61 \pm 2.71}$	$\mathbf{98.89 \pm 1.16}$	$\mathbf{73.12 \pm 3.11}$	$\mathbf{77.23 \pm 3.81}^{\diamond}$

Notes:
$\star$ indicates statistically significant improvement of accuracy compared to the second best result at $\alpha = 0.05$,
$\diamond$ indicates statistically significant improvement of accuracy compared to the second best result at $\alpha = 0.1$, and
$\dagger$ indicates no significant difference between EASR+GP and the best result in the rest of the column (statistically).

the significance tests are described along with the summary of performance results.

Table I summarizes *Mean* ± *Standard Deviation* of the identification rates for different methods in the literature on Honda/UCSD, CMU-MoBo, and YouTube Celebrities datasets for both the truncated sequences (only the first 50 frames), as well as the full length sequences.

On the CMU-MoBo dataset, the proposed method achieved a slightly lower identification rate compared to ISCRC (less than 1%). However, based on the statistical analysis, there is no significant difference between the results. It should be noted that the Honda/UCSD and CMU-MoBo datasets are commonly used as benchmarks and considered as easier recognition tasks since most of the algorithms in the literature have already achieved above 90% accuracy. Therefore, there is not much room for improvement. However, we believe that the results on the most challenging dataset, YouTube Celebrities, can rank different algorithms in terms of performance and efficiency.

For the YTC dataset, the EASR+GP approach achieved significantly better results and improved state-of-the-art by $\approx 4\%$ for the full length sequences. It also achieves the highest accuracy for the truncated sequences, as reflected in Table I. The superior results of the proposed method can be attributed to its capability of handling extremely noisy samples in the YTC dataset more efficiently compared to the rest of the methods in the literature. It is worth mentioning that a simple ensemble of GP binary classifiers without employing the specialization – generalization learning strategy performed poorly on YTC, testifying to the merit of our approach.

As mentioned before, the competing methods also benefited from using HOG features, especially the top two performers, namely RNP and SANP. When HOG features are used, the average accuracy of RNP and SANP algorithms increase by over 8% compared to the reported accuracies in [29] and [10] that used intensity levels as features.

The identification rate for JSR on the YTC dataset with full length sequences is 73.7% as reported in [5]. This accuracy is only 0.2% higher than the best contender reported here (it does not affect the result of the significance test though). However, as mentioned before, the evaluation setting used in [5] is different and the results are reported for 30×30 resolution, therefore we did not include this result in Table I.

Running time comparison: We also report the average computation time of all methods in experiments on the YTC dataset for the truncated sequences (with 50 frames). All the timing results are reported based on running Matlab codes provided by the authors of each algorithm on a machine with an Intel Xeon E5-2603 (1.8 GHz) processor and 40 Gigabytes of RAM. We report the average online identification time (in seconds) for one sequence (Table II). We also provide the total offline training time (in seconds) for methods that required training including the EASR+GP method. While our proposed method requires an initial offline training of the models (over 70% of this time is used for training the GP models), it is important to note that the offline training time is a one-time only overhead. For comparison, SANP will require an extra 160 seconds for identifying only 10 test sequences compared to our method. Also, adding a new subject to the gallery requires far less training time, since only one new model needs to be constructed.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a representation structure for video sequences called Ensemble of Abstract Sequence

TABLE II

AVERAGE COMPUTATION TIME (SECONDS) OF DIFFERENT METHODS ON THE YTC DATASET WITH TRUNCATED SEQUENCES (50 FRAMES). T1: TOTAL OFFLINE TRAINING TIME. T2: AVERAGE ONLINE TESTING TIME FOR ONE SEQUENCE.

	MSM	**MDA**	**AHISD**	**CHISD**	**SANP**	**RNP**	**MSSRC**	**ISCRC**	**EASR+GP**
T1	N/A	24.21	N/A	N/A	N/A	1.18	N/A	22.31	156.18
T2	0.64	2.63	3.35	8.46	19.01	0.92	70.78	2.04	2.79

Note: N/A indicates online-only methods

Representatives (EASR) which is tuned to reduce the effect of noisy frames in a sequence, along with a learning scheme called specialization – generalization which tries to alleviate the effect of noisy frames. In this scheme, EASRs are employed to select the most informative sequences from the training data to support efficient learning. These sequences are then used to provide learning instances for training an ensemble of Gaussian Process (GP) models. Identification is done in a hierarchical manner using both the EASR similarities as well as the ensemble of GP binary classifiers.

Evaluation of the EASR+GP approach on Honda/UCSD, CMU-MoBo, and YouTube Celebrities datasets demonstrated better and promising performance of the proposed method compared to a host of other relevant methods including the state-of-the-art. The improvement is especially noticeable (and statistically significant) on the most challenging dataset (YouTube Celebrities). For this dataset we compared our work with an essentially enhanced version of the other algorithms, as using HOG descriptors instead of intensity levels as features improved their performance (compared to the identification rates provided in the original works).

We are planning to go beyond the face identification task and test the proposed method with datasets for other types of video based recognition tasks (e.g., object categorization). We are also working on an extended version of the proposed method which uses other kernels (e.g., pyramid match kernel [8]) for the Gaussian process models which can better take advantage of localized feature descriptors such as Scale-Invariant Feature Transform (SIFT) [15]. Another potential extension of this work is to use EASR as a general purpose filtering approach to improve other methods in the literature in terms of their resilience to noisy frames.

ACKNOWLEDGEMENTS

This work is supported by NSERC Discovery Grant (No. 261403-2011 RGPIN). The authors also thank Dr. Ruiping Wang for sharing the cropped faces of Honda/UCSD dataset. *Contribution of authors.* NH: design of the approach, implementation of the algorithm and the experiments, data analysis, writing and editing the manuscript; LC: design of the study and the approach, extracting the faces from YTC.

REFERENCES

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 581–588, 2005.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2567–2573, 2010.

[4] Y.-C. Chen, V. Patel, P. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision, ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 766–779. Springer Berlin Heidelberg, 2012.

[5] Z. Cui, H. Chang, S. Shan, B. Ma, and X. Chen. Joint sparse representation for video-based face recognition. *Neurocomputing*, 135(0):306 – 312, 2014.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[7] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Robotics Research*, pages 192–201. Springer, 2005.

[8] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *The Journal of Machine Learning Research*, 8:725–760, 2007.

[9] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA, June 2001.

[10] Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1992–2004, 2012.

[11] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[12] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[13] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1005–1018, 2007.

[14] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–313–I–320 vol.1, 2003.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[16] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[18] E. G. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3531–3538. IEEE, 2013.

[19] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[20] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[21] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[22] R. Wang and X. Chen. Manifold discriminant analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 429–436, 2009.

[23] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503, June 2012.

[24] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifold-manifold distance and its application to face recognition with image sets. *IEEE Transactions on Image Processing*, 21(10):4466–4479, 2012.

[25] T. Wang and P. Shi. Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, 30(13):1161–1165, 2009.

[26] B. L. Welch. The generalization of s̈tudent'sp̈roblem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[27] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *The Journal of Machine Learning Research*, 4:913–931, 2003.

[28] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proceedings of 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323, 1998.

[29] M. Yang, P. Zhu, L. Van Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–7, 2013.

[30] P. Zhu, W. Zuo, L. Zhang, S.-K. Shiu, and D. Zhang. Image set-based collaborative representation for face recognition. *Information Forensics and Security, IEEE Transactions on*, 9(7):1120–1132, 2014.