Face Recognition in Movie Trailers via Mean Sequence Sparse Representation-based Classification

Enrique G. Ortiz, Alan Wright, and Mubarak Shah
Center for Research in Computer Vision, University of Central Florida, Orlando, FL
eortiz@cs.ucf.edu, alanwright@knights.ucf.edu, shah@crcv.ucf.edu

Abstract

This paper presents an end-to-end video face recognition system, addressing the difficult problem of identifying a video face track using a large dictionary of still face images of a few hundred people, while rejecting unknown individuals. A straightforward application of the popular ℓ^1-minimization for face recognition on a frame-by-frame basis is prohibitively expensive, so we propose a novel algorithm Mean Sequence SRC (MSSRC) that performs video face recognition using a joint optimization leveraging all of the available video data and the knowledge that the face track frames belong to the same individual. By adding a strict temporal constraint to the ℓ^1-minimization that forces individual frames in a face track to all reconstruct a single identity, we show the optimization reduces to a single minimization over the mean of the face track. We also introduce a new Movie Trailer Face Dataset collected from 101 movie trailers on YouTube. Finally, we show that our method matches or outperforms the state-of-the-art on three existing datasets (YouTube Celebrities, YouTube Faces, and Buffy) and our unconstrained Movie Trailer Face Dataset. More importantly, our method excels at rejecting unknown identities by at least 8% in average precision.

1. Introduction

Face Recognition has received widespread attention for the past three decades due to its wide-applicability. Only recently has this interest spread into the domain of video, where the problem becomes more challenging due to the person's motion and changes in both illumination and occlusions. However, it also has the benefit of providing many samples of the same person, thus providing the opportunity to convert many weak examples into a strong prediction of the identity.

As video search sites like YouTube have grown, video content-based search has become increasingly necessary. For example, a capable retrieval system should return all videos containing specific actors upon a user's request. On sites like YouTube, where a cast list or script may not be available, the visual content is the key to accomplishing this task successfully. The main drawback is the availability of annotated video face tracks.

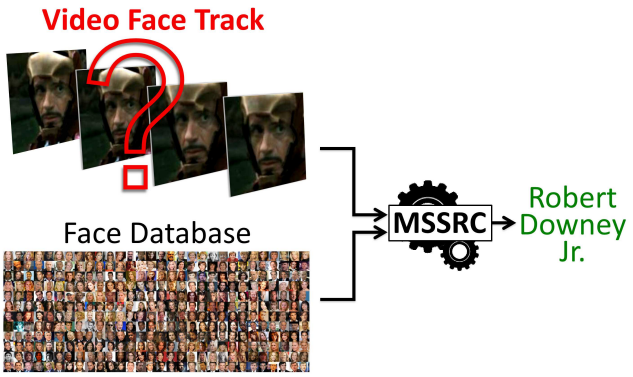


Figure 1. This paper addresses the difficult problem of identifying a video face track using a large dictionary of still face images of a few hundred people, while rejecting unknown individuals.

With the advent of social networking and photo-sharing, computer vision tasks on the Internet have become increasingly fascinating and viable. This avenue is one little exploited by video face recognition. Although large collections of annotated individuals in videos are not freely available, collecting data of annotated still images is easily doable, as witnessed by datasets like Labeled Faces in the Wild (LFW) [12] and Public Figures (PubFig) [16]. Due to wide availability, we employ large databases of still images to recognize individuals in videos, as depicted in Figure 1.

Existing video face recognition methods tend to perform classification on a frame-by-frame basis and later combining those predictions using an appropriate metric. A straight-forward application of ℓ^1-minimization in this fashion is very computationally expensive. In contrast, we propose a novel method, Mean Sequence Sparse Representation-based Classification (MSSRC), that performs a joint optimization over all faces in the track at once. Though this seems expensive, we show that this optimiza-

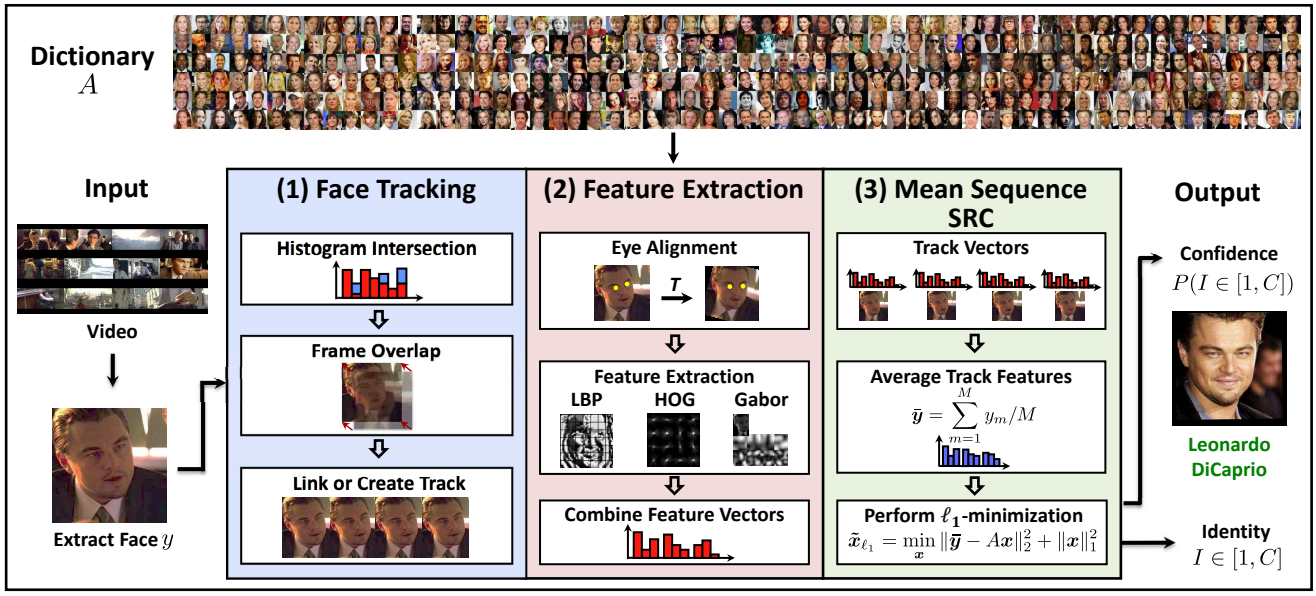


Figure 2. Video Face Recognition Pipeline. With a video as input, we perform face detection and track a face throughout the video clip. Then we extract, PCA, and concatenate three features, Gabor, LBP, and HOG. Finally, we perform face recognition using our novel algorithm MSSRC with an input face track and dictionary of still images.

tion reduces to a single ℓ^1-minimization over the mean face track, thus reducing a many classification problem to one with inherent computational and practical benefits.

Our proposed method aims to perform video face recognition across domains, leveraging thousands of labeled, still images gathered from the Internet, specifically the PubFig and LFW datasets, to perform face recognition on real-world, unconstrained videos. To do this we collected 101 movie trailers from YouTube and automatically extracted and tracked faces in the video to create a dataset for video face recognition (http://vfr.enriquegortiz.com). Furthermore, we explore the often little-studied, open-universe scenario in which it is important to recognize and reject unknown identities, *i.e*. we identify famous actors appearing in movie trailers while rejecting background faces that represent unknown extras. We show our method outperforms existing methods in precision and recall, exhibiting the ability to better reject unknown or uncertain identities.

The contributions of this paper are summarized as follows: (1) We develop a fully automatic end-to-end system for video face recognition, which includes face tracking and recognition leveraging information from both still images for the known dictionary and video for recognition. (2) We propose a novel algorithm, MSSRC, that performs video face recognition using an optimization leveraging all of the available video data. (3) We show that our method matches or outperforms the state-of-the-art on three existing datasets (YouTube Faces, YouTube Celebrities, and Buffy) and our unconstrained Movie Trailer Face Dataset.

The rest of this paper is organized as follows: Section 2 discusses the related work on video face recognition. Then Section 3 describes our entire framework for video face recognition from tracking to recognition. Next, in Section 4, we describe our unconstrained Movie Trailer Face Dataset. Section 5 exhaustively evaluates our method on existing video datasets and our new dataset. Finally, we end with a summary of conclusions and future work in Section 6.

2. Related Work

For a complete survey of video-based face recognition refer to [18]; here we focus on an overview of the most related methods. Current video face recognition techniques fall into one of three categories: key-frame based, temporal model based, and image-set matching based.

Key-frame based methods generally perform a prediction on the identity of each key-frame in a face track followed by a probabilistic fusion or majority voting to select the best match. Due to the large variations in the data, key-frame selection is crucial in this paradigm [4]. Zhao *et al*.'s [25] work is most similar to us in that they use a database with still images collected from the Internet. They learn a model over this dictionary by learning key faces via clustering. These cluster centers are compared to test frames using a nearest-neighbor search followed by majority, probabilistic voting to make a final prediction. We, on the other hand, use a classification scheme that enhances robustness by finding an agreement amongst the individual frames in a single optimization.

Temporal model based methods learn the temporal, facial dynamics of the face throughout a video. Several methods employ Hidden Markov Models (HMM) for this end, *e.g*. [14]. Most related to us, Hadid *et al*. [10] uses a still image training library by imposing motion information on it to train an HMM and Zhou *et al*. [26] probabilistically generalizes a still-image library to do video-to-video matching. Generally training these models is prohibitively expensive, especially when the dataset size is large.

Image-set matching based methods allows the modeling of a face track as an image-set. Many methods, like [24], perform a mutual subspace distance where each face track is modeled in their own subspace from which a distance is computed between each. They are effective with clean data, but these methods are very sensitive to the variations inherent in video face tracks. Other methods take a more statistical approach, like [5], which used Logistic Discriminant-based Metric Learning (LDML) to learn a relationship between images in face tracks, where the inter-class distances are maximized. LDML is very computationally expensive and focuses more on learning relationships within the data, whereas we directly relate the test track to the training data.

Character recognition methods have been very popular due to their application to movies and sitcoms. [8, 19] perform person identification, where they use all available information, e.g. clothing appearance and audio, to identify the cast rather than the facial information alone. Another [3] used a small user selected sample of characters in the given movie to do a pixel-wise Euclidean distance to handle occlusion. While others [2], use a manifold for known characters which successfully clusters input frames. While character recognition is suitable for a long-running series, the use of clothing and other contextual clues are not helpful in the task of identifying actors between movies, TV shows, or non-related video clips. In these scenarios, our approach of focusing on the facial recognition aspect from still images is more adept in unconstrained environments.

Still-Image based literature is vast, but a popular approach is Wright *et al*.'s [23] Sparse Representation-based Classification (SRC), in which they present the principle that a given test image can be represented by a linear combination of images from a large dictionary of faces. The key concept is enforcing sparsity, since a test face can be reconstructed best from a small subset of the large dictionary, *i.e*. training faces of the same class. A straight-forward adaptation of this method would be to perform estimation on each frame and fuse results probabilistically, similarly to key-frame based methods. However, ℓ^1-minimization is known to be computationally expensive, thus we propose a constrained optimization with the knowledge that the images within a face track are of the same person. We show that imposing this fact reduces the problem to computing a single ℓ^1-minimization over the average face track.

3. Video Face Recognition Pipeline

In this section, we describe our end-to-end video face recognition system. First, we detail our algorithm for face tracking based on face detections from video. Next, we chronicle the features we use to describe the faces and handle variations in pose, lighting, and occlusion. Finally, we derive our optimization for video face recognition that classifies a video face track based on a dictionary of still images.

3.1. Face Tracking

Our method performs the difficult task of face tracking based on face detections extracted using the high-performance SHORE face detection system [15] and generates a face track based on two metrics. To associate a new detection to an existing track, our first metric determines the ratio of the maximum sized bounding box encompassing both face detections to the size of the larger bounding box of the two detections. The formulation is as follows:

$$d_{spatial} = \frac{w * h}{\max(h_1 * w_1, h_2 * w_2)}, \qquad (1)$$

where (x_1, y_1, w_1, h_1) and (x_2, y_2, w_2, h_2) are the (x, y) location and the width and height of the previous and current frames respectively. The overall width w and height h are computed as $w = \max(x_1 + w_1, x_2 + w_2) - \min(x_1, x_2)$ and $h = max(y_1 + h_1, y_2 + h_2) - \min(y_1, y_2)$. Intuitively, this metric encodes the dimensional similarity of the current and previous bounding boxes, intrinsically considering the spatial information.

The second tracking metric takes into account the appearance information via a local color histogram of the face. We compute the distance as a ratio of the histogram intersection of the RGB histograms with 30 bins per channel of the last face of a track and the current detection to the total summation of the histogram bins:

$$d_{appearance} = \sum_{i=1}^{n} \min(a_i, b_i) / \sum_{i=1}^{n} a_i + b_i, \qquad (2)$$

where a and b are the histograms of the current and previous face. We compare each new face detection to existing tracks; if the location and appearance metric is similar, the face is added to the track, otherwise a new track is created. Finally, we use a global histogram for the entire frame, encoding scene information, to detect scene boundaries and impose a lifespan of 20 frames of no detection to end tracks.

3.2. Feature Extraction

Because real-world datasets contain pose variations even after alignment, we use three fast and popular local features: Local Binary Patterns (LBP) [1], Histogram of Oriented Gradients (HOG) [7], and Gabor wavelets [17]. More features aid recognition, but at a higher computational cost.

Algorithm 1 Mean Sequence SRC (MSSRC)

1. **Input:** Training gallery A, test face track $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M]$, and sparsity weight parameter λ.
2. Normalize the columns of A to have unit ℓ^2-norm.
3. Compute mean of the track $\bar{\boldsymbol{y}} = \sum_{m=1}^{M} \boldsymbol{y}_m / M$ and normalize to unit ℓ^2-norm..
5. Solve the ℓ^1-minimation problem

$$\tilde{\boldsymbol{x}}_{\ell_1} = \arg\min_{\boldsymbol{x}} \|\bar{\boldsymbol{y}} - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$$

6. Compute residual errors for each class $j \in [1, C]$

$$r_j(\bar{\boldsymbol{y}}) = \|\bar{\boldsymbol{y}} - A_j x_j\|_2$$

7. **Output**: identity I and confidence $P(I|\bar{\boldsymbol{y}})$

$$I(\bar{\boldsymbol{y}}) = \arg\min_j r_j(\bar{\boldsymbol{y}})$$

$$P(I \in [1, C]|\bar{\boldsymbol{y}}) = \frac{C \cdot \max_j \|x_j\|_1 / \|\tilde{\boldsymbol{x}}\|_1 - 1}{C - 1}$$

Before feature extraction, all images are first eye-aligned using eye locations from SHORE and normalized by subtracting the mean, removing the first order brightness gradient, and performing histogram equalization. Gabor wavelets were extracted with one scale $\lambda = 4$ at four orientations $\theta = \{0°, 45°, 90°, 135°\}$ with a tight face crop at a resolution of 25x30 pixels. A null Gabor filter includes the raw pixel image (25x30) in the descriptor. The standard $\text{LBP}^{U2}_{8,2}$ and HOG descriptors are extracted from 72x80 loosely cropped images with a histogram size of 59 and 32 over 9x10 and 8x8 pixel patches, respectively. All descriptors were scaled to unit norm, dimensionality reduced with PCA to 1536 dimensions each, and zero-meaned.

3.3. Mean Sequence Sparse Representation-based Classification (MSSRC)

Given a test image $\boldsymbol{y}$ and training set A, we know that the images of the same class to which $\boldsymbol{y}$ should match is a small subset of A and their relationship is modeled by $\boldsymbol{y} = A\boldsymbol{x}$, where $\boldsymbol{x}$ is the coefficient vector relating them. Therefore, the coefficient vector $\boldsymbol{x}$ should only have non-zero entries for those few images from the same class and zeros for the rest. Imposing this sparsity constraint upon the coefficient vector $\boldsymbol{x}$ results in the following formulation:

$$\hat{\boldsymbol{x}}_{\ell_1} = \arg\min_x \|\boldsymbol{y} - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1, \tag{3}$$

where the ℓ^1-norm enforces a sparse solution by minimizing the absolute sum of the coefficients.

The leading principle of our method is that all of the images $\boldsymbol{y}$ from the face track $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M]$ belong to the same person. Because all images in a face track belong to the same person, one would expect a high degree of correlation amongst the sparse coefficient vectors $\boldsymbol{x}_j \forall j \in [1 \ldots M]$, where M is the length of the track. Therefore, we can look for an agreement on a single coefficient vector $\boldsymbol{x}$ determining the linear combination of training images A that make up the unidentified person. In fact, with sufficient similarity between the faces in a track, one might expect nearly the same coefficient vector to be recovered for each frame. This provides the intuition for our approach: we enforce a single coefficient vector for all frames. Mathematically, this means the sum squared residual error over the fames should be minimized. We enforce this constraint on the ℓ^1 solution of Eqn. 3 as follows:

$$\tilde{\boldsymbol{x}}_{\ell_1} = \arg\min_{\boldsymbol{x}} \sum_{m=1}^{M} \|\boldsymbol{y}_m - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \tag{4}$$

where we minimize the ℓ^2 error over the entire image sequence, while assuming the coefficient vector $\boldsymbol{x}$ is sparse and the same over all of the images.

Focusing on the first part of the equation, more specifically the ℓ^2 portion, we can rearrange it as follows:

$$\begin{aligned}\sum_{m=1}^{M} \|\boldsymbol{y}_m - A\boldsymbol{x}\|_2^2 &= \sum_{m=1}^{M} \|\boldsymbol{y}_m - \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}} - A\boldsymbol{x}\|_2^2 \\ &= \sum_{m=1}^{M} (\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|_2^2 + 2(\boldsymbol{y}_m - \bar{\boldsymbol{y}})^T(\bar{\boldsymbol{y}} - A\boldsymbol{x}) + \ldots \\ &\qquad \|\bar{\boldsymbol{y}} - A\boldsymbol{x}\|_2^2), \end{aligned} \tag{5}$$

where $\bar{\boldsymbol{y}} = \sum_{m=1}^{M} \boldsymbol{y}_m / M$. However,

$$\begin{aligned}&\sum_{m=1}^{M} 2(\boldsymbol{y}_m - \bar{\boldsymbol{y}})^T(\bar{\boldsymbol{y}} - A\boldsymbol{x}) \\ &= 2\left(\textstyle\sum_{m=1}^{M} \boldsymbol{y}_m - M\bar{\boldsymbol{y}}\right)(\bar{\boldsymbol{y}} - A\boldsymbol{x}) \\ &= 0(\bar{\boldsymbol{y}} - A\boldsymbol{x}) = 0.\end{aligned}$$

Thus, Eq. 5 becomes:

$$\begin{aligned}&\sum_{m=1}^{M} \|\boldsymbol{y}_m - A\boldsymbol{x}\|_2^2 \\ &= \sum_{m=1}^{M} \|\boldsymbol{y}_m - \bar{\boldsymbol{y}}\|_2^2 + M\|\bar{\boldsymbol{y}} - A\boldsymbol{x}\|_2^2,\end{aligned} \tag{6}$$

where the first part of the sum is a constant. Therefore, we obtain the final simplification of our original minimization:

$$\begin{aligned}\tilde{\boldsymbol{x}}_{\ell_1} &= \arg\min_{\boldsymbol{x}} \sum_{m=1}^{M} \|\boldsymbol{y}_m - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1^2 \\ &= \arg\min_{\boldsymbol{x}} M\|\bar{\boldsymbol{y}} - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \\ &= \arg\min_{\boldsymbol{x}} \|\bar{\boldsymbol{y}} - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \end{aligned} \quad (7)$$

where M, by division, is absorbed by the constant weight λ. By this sequence, our optimization reduces to the ℓ^1-minimization of $\boldsymbol{x}$ for the mean face track $\bar{\boldsymbol{y}}$.

This conclusion, that enforcing a single, consistent coefficient vector $\boldsymbol{x}$ across all images in a face track Y is equivalent to a single ℓ^1-minimization over the average of all the frames in the face track, is key to keeping our approach robust yet fast. Instead of performing M individual ℓ^1-minimizations over each frame and classifying via some voting scheme, our approach performs a single ℓ^1-minimization on the mean of the face track, which is not only a significant speed up, but theoretically sound. Furthermore, we empirically validate in subsequent sections that our approach outperforms other forms of temporal fusion and voting amongst individual frames.

Finally, we classify the average test track $\bar{\boldsymbol{y}}$ by determining the class of training samples that best reconstructs the face from the recovered coefficients:

$$I(\bar{\boldsymbol{y}}) = \min_j r_j(\bar{\boldsymbol{y}}) = \min \|\bar{\boldsymbol{y}} - A_j x_j\|_2, \quad (8)$$

where the label $I(\bar{\boldsymbol{y}})$ of the test face track is the minimal residual or reconstruction error $r_j(\bar{\boldsymbol{y}})$ and x_j is the recovered coefficients from the global solution $\tilde{\boldsymbol{x}}_{\ell_1}$ that belong to class j. Confidence in the determined identity is obtained using the Sparsity Concentration Index (SCI), which is a measure of how distributed the residuals are across classes:

$$SCI = \frac{C \cdot \max_j \|x_j\|_1 / \|\tilde{\boldsymbol{x}}\|_1 - 1}{C - 1} \in [0, 1], \quad (9)$$

ranging from 0 (the test face is represented equally by all classes) to 1 (the test face is fully represented by one class).

4. Movie Trailer Face Dataset

Existing datasets do not capture the large-scale identification scope we wish to evaluate. The YouTube Celebrities Dataset [14] has unconstrained videos from YouTube, however they are very low quality and only contain 3 unique videos per person, which they segment. The YouTube Faces Dataset [22] and Buffy Dataset [5] also exhibit more challenging scenarios than traditional video face recognition datasets, however YouTube Faces is geared towards face verification, same vs. not same, and Buffy only contains 8 actors; thus, both are ill-suited for the large-scale face identification of our proposed video retrieval framework.

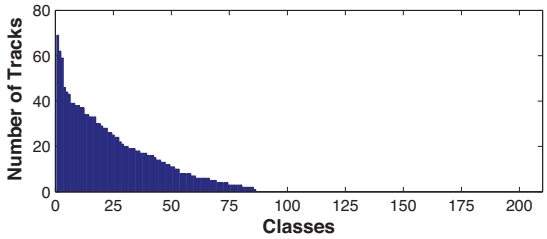


Figure 3. The distribution of face tracks across the identities in PubFig+10.

We built our Movie Trailer Face Dataset using 101 movie trailers from YouTube from the 2010 release year that contained celebrities present in the supplemented PublicFig+10 dataset. These videos were then processed to generate face tracks using the method described above. The resulting dataset contains 4,485 face tracks, 65% consisting of unknown identities (not present in PubFig+10) and 35% known. The class distribution is shown in Fig. 3 with the number of face tracks per celebrity in the movie trailers ranging from 5 to 60 labeled samples. The fact that half of the public figures do not appear in any of the movie trailers presents an interesting test scenario in which the algorithm must be able to distinguish the subject of interest from within a large pool of potential identities.

5. Experiments

In this section, we first compare our tracking method to a standard method used in the literature. Then, we evaluate our video face recognition method on three existing datasets, YouTube Faces, YouTube Celebrities, Buffy. We also evaluate several algorithms, including MSSRC (ours), on our new Movie Trailer Face Dataset, showing the strengths and weaknesses of each and thus proving experimentally the validity of our algorithm.

5.1. Tracking Results

To analyze the quality of our automatically generated face tracks, we ground-truthed five movie trailers from the dataset: 'The Killer Inside', 'My Name is Khan', 'Biutiful', 'Eat, Pray, Love', and 'The Dry Land'. Based on tracking literature [13], we use two CLEAR MOT metrics, Multiple Object Tracking Accuracy and Precision (MOTP and MOTA), for evaluation that better consider issues faced by trackers than standard accuracy, precision, or recall. The MOTA tells us how well the tracker did overall in regards to all of the ground-truth labels, while the MOTP appraises how well the tracker performed on the detections that exist in the ground-truth.

Video		Method	
		KLT [8]	Ours
'The Killer Inside'	MOTP	68.93	69.35
	MOTA	42.88	42.16
'My Name is Khan'	MOTP	65.63	65.77
	MOTA	44.26	48.24
'Biutiful'	MOTP	61.58	61.34
	MOTA	39.28	43.96
'Eat Pray Love'	MOTP	56.98	56.77
	MOTA	34.33	35.60
'The Dry Land'	MOTP	64.11	62.70
	MOTA	27.90	30.15
Average	MOTP	**63.46**	63.19
	MOTA	37.73	**40.02**

Table 1. Tracking Results. Our method outperforms the KLT-based [8] method in terms of MOTA by 2%.

Method	Accuracy $\pm$ SE	AUC	EER
MBGS [22]	75.3 ± 2.5	82.0	26.0
MSSRC (Ours)	$\mathbf{75.3 \pm 2.2}$	**82.9**	**25.3**

Table 2. YouTube Faces Dataset. Results for top performing video face verification algorithm MBGS and our competitive method MSSRC. Note: MBGS results are different from those published, but they are the output of default settings in their system.

Although our goal is not to solve the tracking problem, in Table 1 we show our results compared to a standard face tracking method. The first column shows a KLT-based method [8], where the face detections are associated based on a ratio of overlapping tracked features, and the second shows our method. Both methods are similarly precise, however our metrics have a larger coverage of total detections/tracks by 2% in MOTA with a 3.5x speedup. Results are available online.

5.2. YouTube Faces Dataset

Although face identification is the focus of our paper, we evaluated our method on the YouTube Faces Dataset [22] for face verification (same/not same), to show that our method can also work in this context. To the best of our knowledge, there is only one paper [9], that has done face verification using SRC, however it was not in the context of video face recognition, but that of still images from LFW. The YouTube Faces Dataset consists of 5,000 video pairs, half same and half not. The videos are divided into 10 splits each with 500 pairs. The results are averaged over the ten splits, where for each split one is used for testing and the remaining nine for training. The final results are presented in terms of accuracy, area under the curve, and equal error rate. As seen in Table 4, we obtain competitive results with the top performing method MBGS [22], within 1% in terms of accuracy, and MSSRC even surpasses it in terms of area under the curve (AUC) by just below 1% with a lower equal error rate by 0.7%. We perform all experiments with the same LBP data provided by [22] and a τ value of 0.0005.

Method	Accuracy (%)
HMM [14]	71.24
MDA [20]	67.20
SANP [11]	65.03
COV+PLS [21]	70.10
UISA [6]	74.60
MSSRC (Ours)	**80.75**

Table 3. YouTube Celebrities Dataset. We outperform the best reported result by 6%.

Method	Accuracy (%)
LDML [5]	85.88
MSSRC (Ours)	**86.27**

Table 4. Buffy Dataset. We obtain a slight gain in accuracy over the reported method.

5.3. YouTube Celebrities Dataset

The YouTube Celebrities Dataset [14] consists of 47 celebrities (actors and politicians) in 1910 video clips downloaded from YouTube and manually segmented to the portions where the celebrity of interest appears. There are approximately 41 clips per person segmented from 3 unique videos per actor. The dataset is challenging due to pose, illumination, and expression variations, as well as high compression and low quality. Using our tracker, we successfully tracked 92% of the videos as compared to the 80% tracked in their paper [14]. The standard experimental setup selects 3 training clips, 1 from each unique video, and 6 test clips, 2 from each unique video, per person. In Table 3, we summarize reported results on YouTube Celebrities, where we outperform the state-of-the-art by at least 6%.

5.4. Buffy Dataset

The Buffy Dataset consists of 639 manually annotated face tracks extracted from episodes 9, 21, and 45 from different seasons of the TV series "Buffy the Vampire Slayer". They generated tracks using the KLT-based method [8] (available on the author's website). For features, we compute SIFT descriptors at 9 fiducial points as described in [5] and use their experimental setup with 312 tracks for training and 327 testing. They present a Logistic Discriminant-based Metric Learning (LMDL) method that learns a subspace. In their supervised experiments, they tried several classifiers with each obtaining similar results. However, using our classifier, there is a slight improvement.

Method	AP (%)	Recall (%)
NN	9.53	0.00
SVM	50.06	9.69
LDML [5]	19.48	0.00
L2	36.16	0.00
SRC (First Frame)	42.15	13.39
SRC (Voting)	54.88	23.47
MSSRC (Ours)	**58.70**	**30.23**

Table 5. Movie Trailer Face Dataset. MSSRC outperforms all of the non-SRC methods by at least 8% in AP and 20% recall at 90% precision.

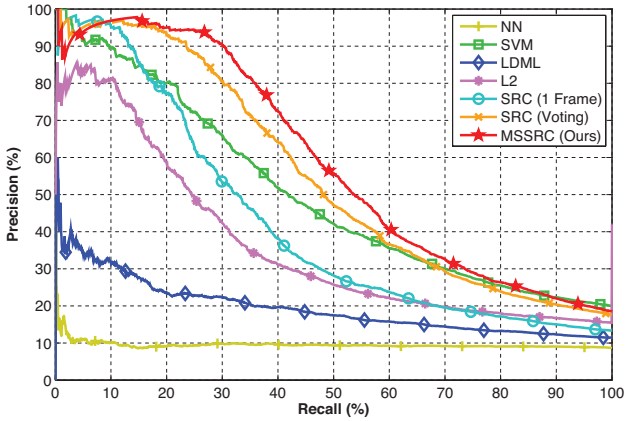


Figure 4. Precision vs. Recall for the Movie Trailer Face Dataset. MSSRC rejects unknowns or distractors better than all others.

5.5. Movie Trailer Face Dataset

In this section, we present results on our unconstrained Movie Trailer Face Dataset that allows us to test larger scale face identification, as well as each algorithms ability to reject unknown identities. In our test scenario, we chose the Public Figures (PF) [16] dataset as our training gallery, supplemented by images collected of 10 actors and actresses from web searches for additional coverage of face tracks extracted from movie trailers. We also cap the maximum number of training images per person in the dataset to 200 for better performance due to the fact that predictions are otherwise skewed towards the people with the most examples. The distribution of face tracks across all of the identities in the PubFig+10 dataset are shown in Fig. 3. In total, PubFig+10 consists of 34,522 images and our Movie Trailer Face Dataset has 4,485 face tracks, which we use to conduct experiments on several algorithms.

5.5.1 Algorithmic Comparison

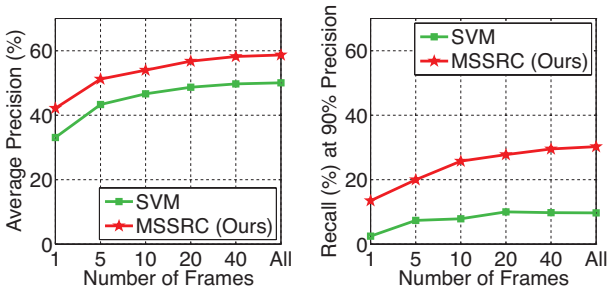
The tested methods include NN, LDML, SVM, L2, SRC, and our method MSSRC. For the experiments with NN, LDML, SVM, L2, and SRC, we test each individual frame of the face track and predict its final identity via probabilistic voting and its confidence is an average over the predicted distances or decision values. The confidence values are used to reject predictions to evaluate the precision and recall of the system. Note all MSSRC experiments are performed with a λ value of 0.01. We present results in terms of precision and recall as defined in [8].

Table 5 presents the results for the described methods on the Movie Trailer Face Dataset in terms of two measures, average precision and recall at 90% precision. NN performs very poorly in terms of both metrics, which explains why NN based methods have focused on finding "good" keyframes to test on. LMDL struggles with the larger number of training classes vs. the Buffy experiment with only 19.48% average precision. The L2 method performs surprisingly well for a simple method. We also tried Mean L2 with similar performance. The SVM and SRC based methods perform very closely at high recall, but not in terms of AP and recall at 90% precision with MSSRC outperforming SVM by 8% and 20% respectively. In Fig. 4, the SRC based methods reject unknown identities better than the others.

The straightforward application of SRC on a frame-by-frame basis and our efficient method MSSRC perform within 4% of each other, thus experimentally validating that MSSRC is computationally equivalent to performing standard SRC on each individual frame. Instead of computing SRC on each frame, which takes approximately 45 minutes per track, we reduce a face track to a single feature vector for ℓ^1-minimization (1.5 min/track). Surprisingly, MSSRC obtains better recall at 90% precision by 7% and 4% in average precision. Instead of fusing results after classification, as done on the frame by frame methods, MSSRC benefits in better rejection of uncertain predictions. In terms of timing, the preprocessing steps of tracking runs identically for SRC and MSSRC at 20fps and feature extraction runs at 30fps. For identification, MSSRC classifies at 20 milliseconds per frame, whereas SRC on a single frame takes 100 milliseconds. All other methods classify in less than 1ms, however with a steep drop in precision and recall.

5.5.2 Effect of Varying Track Length

The question remains, do we really need all of the images? To answer this question we select the first m frames for each track and test the two best performing methods from the previous experiments: MSSRC and SVM. Fig. 5 shows that at just after 20 frames performance plateaus, which is close to the average track length of 22 frames. Most importantly, the results show that using multiple frames is beneficial since moving from using 1 frame to 20 frames results in a 5.57% and 16.03% increase in average precision and recall at 90% precision respectively for MSSRC. Fur-



(a) Average Precision (b) Recall at 90% Precision

Figure 5. Effect of Varying Track Length. We see that performance levels out at about 20 frames (close to the average track length). MSSRC outperforms SVM by 8% in average in terms of AP.

thermore, Fig. 5 shows that the SVM's performance also increases with more frames, although MSSRC outperforms the SVM method in its ability to reject unknown identities.

6. Conclusions and Future Work

In this paper we have presented a fully automatic end-to-end system for video face recognition, which includes face tracking and identification leveraging information from both still images for the known dictionary and video for recognition. We propose a novel algorithm Mean Sequence SRC, MSSRC, that performs a joint optimization using all of the available image data to perform video face recognition. We finally showed that our method outperforms the state-of-the-art on real-world, unconstrained videos in our new Movie Trailer Face Dataset. Furthermore, we showed our method especially excels at rejecting unknown identities outperforming the next best method in terms of average precision by 8%. Video face recognition presents a very compelling area of research with difficulties unseen in still-image recognition. In the future, we would explore the effect of selecting key-frames, or less noisy frames. Furthermore, there is a whole area of domain transfer for transferring knowledge from the still-image domain to the videos.

Acknowledgement

We acknowledge Brian C. Becker, Niels da Vitoria Lobo, and Xin Li for their feedback and help.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 2006. 3

[2] O. Arandjelovic and R. Cipolla. Automatic Cast Listing in Feature-Length Films with Anisotropic Manifold Space. In *CVPR*, 2006. 3

[3] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *CVPR*, 2005. 3

[4] S. Berrani and C. Garcia. Enhancing face recognition from video sequences using robust statistics. *AVSS*, 2005. 2

[5] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. *ICCV*, 2011. 3, 5, 6, 7

[6] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for Video-Based Face Recognition. In *CVPR*, 2012. 6

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3

[8] M. Everingham and J. Sivic. Taking the bite out of automated naming of characters in TV video. *CVIU*, 2009. 3, 6, 7

[9] H. Guo, R. Wang, J. Choi, and L. S. Davis. Face verification using sparse representations. *CVPR Workshop*, 2012. 6

[10] A. Hadid and M. Pietikainen. From still image to video-based face recognition: an experimental analysis. *FG*, 2004. 3

[11] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, 2011. 6

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 1

[13] R. Kasturi, D. Goldgof, Padmanabhan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *TPAMI*, 2009. 5

[14] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008. 3, 5, 6

[15] C. Kueblbeck and A. Ernst. Face detection and tracking in video sequences using the modified census transformation. *JIVC*, 2006. 3

[16] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *TPAMI*, 2011. 1, 7

[17] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *TIP*, 2002. 3

[18] C. Shan. Face recognition and retrieval in video. *Video Search and Mining*, 2010. 2

[19] M. Tapaswi and M. Bäuml. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series. *CVPR*, 2012. 3

[20] R. Wang and X. Chen. Manifold Discriminant Analysis. In *CVPR*, 2009. 6

[21] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, 2012. 6

[22] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *TPAMI*, 2011. 5, 6

[23] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 2009. 3

[24] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *FG*, 1998. 3

[25] M. Zhao, J. Yagnik, H. Adam, and D. Bau. Large scale learning and recognition of faces in web videos. *FG*, 2008. 2

[26] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *CVIU*, 2003. 3