A Compact and Discriminative Face Track Descriptor

Omkar M Parkhi **Karen Simonyan** **Andrea Vedaldi** **Andrew Zisserman**

Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

Objective

same different

B. Murray **S. Gellar** **E. Taylor**

Goal: Recognise and verify face identities in very large video collections

Video Fisher Vector Faces

A novel, discriminative, efficient, and very compact face track descriptor:

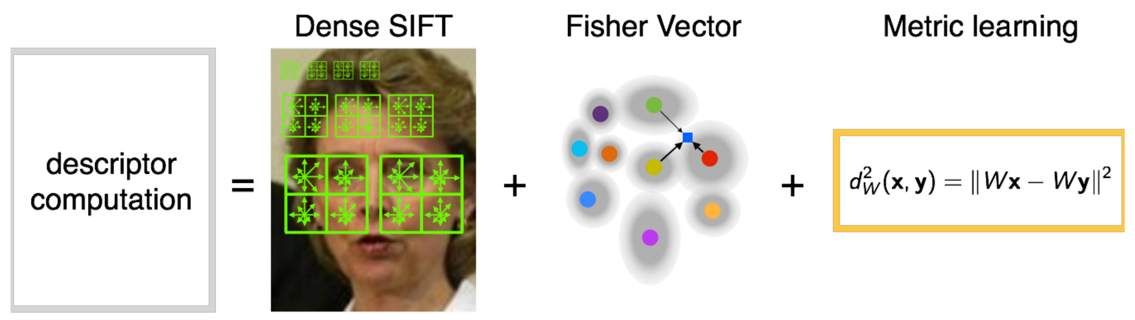
- Robust to face misalignments
- A single descriptor per track
- Compact: low dimensional & binarised

Video Fisher Vector Faces (VF²)

Fisher Vector Faces applied to face tracks:

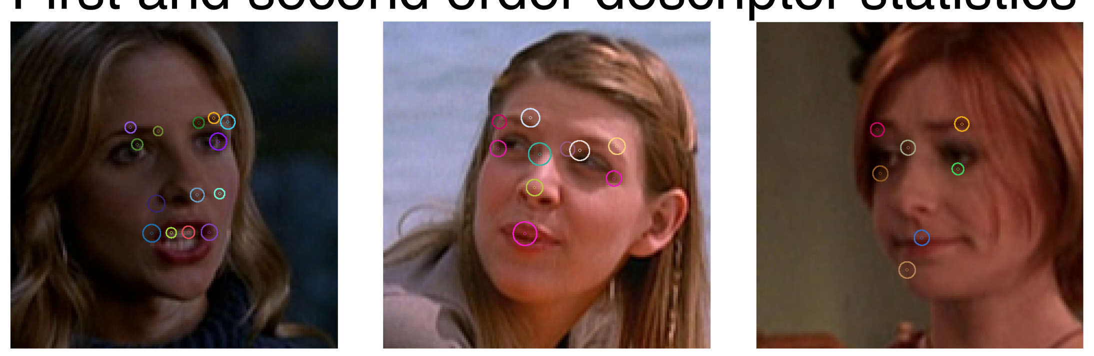
- **Video pooling**: one easy-to-use descriptor per track
- **Jittered pooling**: efficient data augmentation
- **Binarisation**: extreme compression
- **Hard-assignment fisher vector:** 6 times faster

Fisher Vector Faces (FVF)



A powerful single-frame face descriptor:

- Dense sampling of local descriptors (SIFT) with spatial (x,y) augmentation
- Fisher Vector encoding
  - Gaussian Mixture Model codebook
  - First and second order descriptor statistics

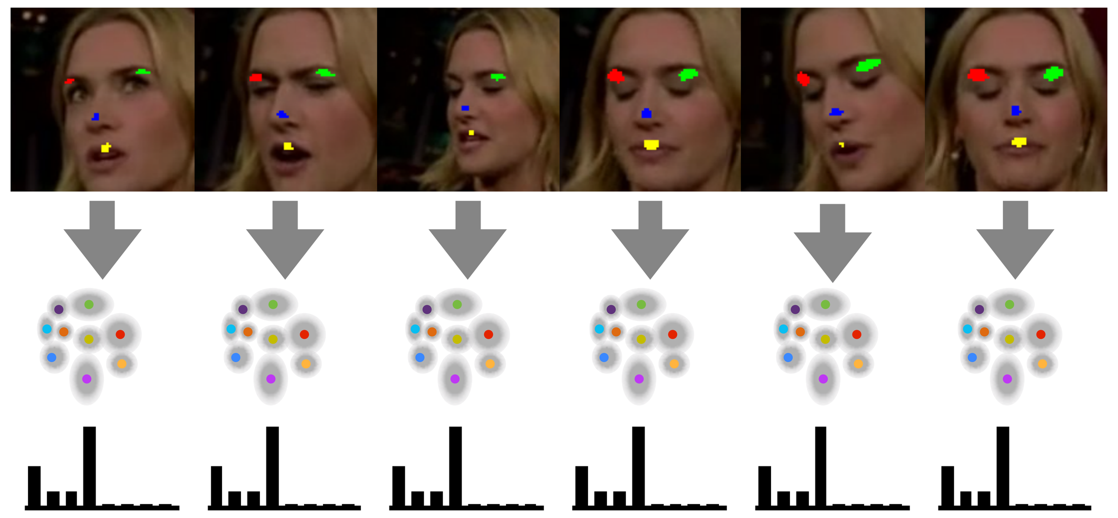


- Discriminative low-rank Mahalanobis metric

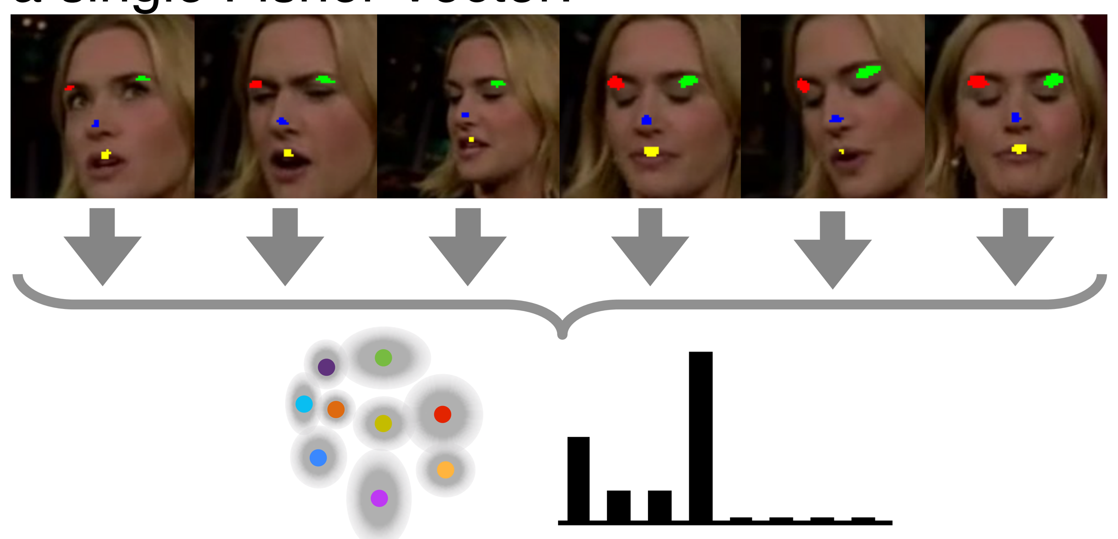
Video and Jittered Pooling

Video Pooling

Conventional face track descriptors compute one vector for each frame:



A face track is then a collection of descriptors that need to be either combined or jointly compared. Instead we pool all frames in a single Fisher Vector:



Jittered Pooling

Pooling can be extended to jittered versions of the data, such as horizontal flips:

Learn to Compare & Compress

Goal: learn to simultaneously compare and compress descriptors.

Method: discriminative **low-rank metric learning**, parametrised by the projection W:

$$d^2_W(\phi_i,\phi_j) = \|W\phi_i - W\phi_j\|_2^2$$
$$= (\phi_i - \phi_j)^\top W^\top W(\phi_i - \phi_j)$$

We also test joint similarity-metric learning:

$$d^2_{V,W}(\phi_i,\phi_j) = (\phi_i - \phi_j)^\top W^\top W(\phi_i - \phi_j) - \phi_i V^\top V \phi_j$$

Objective Function and Learning

$$\min_{V,W} \sum_{i,j} \max\left[1 - \underset{\uparrow\,\text{label}}{y_{ij}}\left(\underset{\uparrow\,\text{bias}}{b} - d^2_{V,W}(\phi_i,\phi_j)\right)\right]$$

Non-convex functions optimised using SGD. Large reduction in dimensionality without performance loss (68K → 128).

Binarisation

Goal: further reduce memory footprint.

Method: Parseval Tight Frame Expansion

1. Start with m-dimensional descriptors ψ
2. Sample a random $n \times n$ matrix M with $n > m$
3. Decompose $M = QR$
4. $U \leftarrow$ first m columns of Q
5. Binarisation sign($U\psi$) has q bits only

Typical use case: compress 128-D float descriptors (4096 bit) down to 1024 bits without accuracy loss (4× reduction).

Experiments

- Excellent performance with small training sets
- Cross-task and cross-dataset transfer

Face Verification on YouTube Faces

- Restricted: train on only pre-specified pairs
- Unrestricted: use any pair

Parameter Tuning

	Method	Proj. Dim.	EER
1	Image Pool. (soft assignment FV)	128	17.3
2	Video Pool. (soft assignment FV)	128	15.0
3	Video Pool.	128	16.2
4	Video Pool. + jitt.	128	14.2
5	Video Pool.	256	16.9
6	Video Pool.	512	17.0
7	Video Pool.	1024	17.0
8	Video Pool. + binar. 1024 bit	128	15.0
9	Video Pool. + binar. 2048 bit	128	15.0
10	Video Pool. + binar. 1024 bit + jitt.	128	13.4
11	Video Pool. + joint sim.	128 x 2	14.4
12	Video Pool. + joint sim. + flip	128 x 2	13.0
13	Video Pool. + joint sim. + jitt.	**128 x 2**	**12.3**

Comparison with the State of the Art

	Method	EER
1	MGBS & SVM -	21.2
2	APEM Fusion	21.4
3	STFRD & PMML	19.9
4	VSOF & OSS (Adaboost)	20.0
5	Our VF2 (restricted)	16.1
6	Our VF2 (restricted & flip)	14.9
7	Our VF2 (unrestricted & flip)	13.0
8	**Our VF2 (unrestricted & jitt.)**	**12.3**
9	**DeepFace (additional training data)**	**8.6**

Face Verification on INRIA Buffy Dataset

- 327 test tracks from 3 episodes of Buffy
- Training set doesn't contain identities

	Method	Feat. Dim.	EER
1	Cinbis et al.	3.5K	42.50
2	Our VF2 (GMM trained on Buffy) & Flip	68K	30.11
3	Cinbis et al. (trained on LFW)	-	36.20
4	Cinbis et al. (trained on Buffy)	-	30.00
5	Our VF2 (trained on YTF) + joint sim + flip	128 x 2	25.77
6	**Our VF2 (trained on YTF) + binar. 2048 bit + flip**	**128**	**21.90**

Face Classification on Oxford Buffy

- 7 episodes from season 5 of "Buffy the Vampire Slayer"
- Training data obtained from alignment of transcripts and subtitles

	GMM & Proj-n train set.	Proj-n.	Bin-n.	Avg. AP
1	Sivic *et al.* (RBF-MKL)			0.81
2	Sivic *et al.* (Average Kernel)			0.79
3	Sivic *et al.* (Average Kernel; ours)			0.80
4	Buffy	none	none	0.81
5	Youtube Faces	none	none	0.80
6	**Youtube Faces + jitt.**	**1024**	**none**	**0.86**
7	Youtube Faces	1024	2048 bits	0.82

Acknowledgement: This work was supported by ERC grant VisRec no. 228180 and EU Project AXES ICT-269980