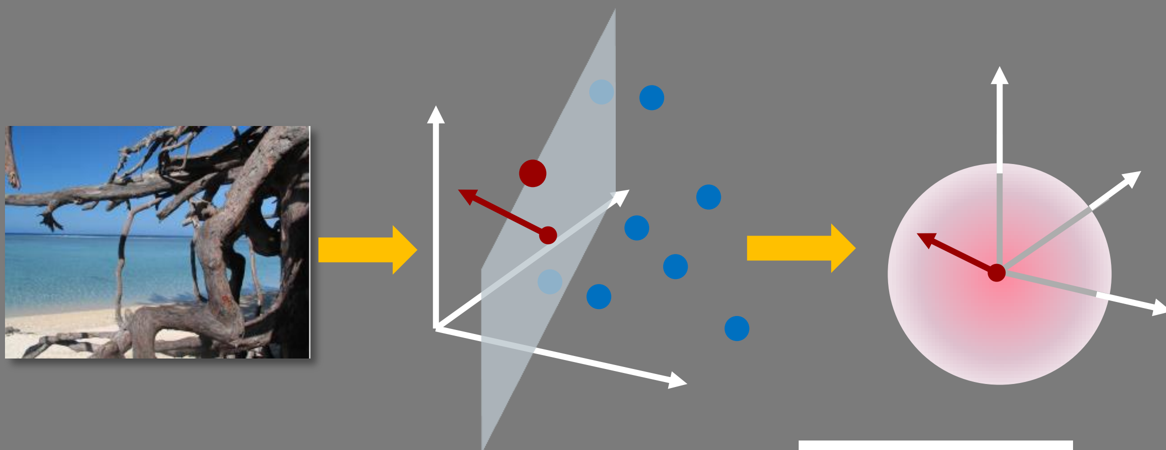
on learned visual embedding

patrick pérez



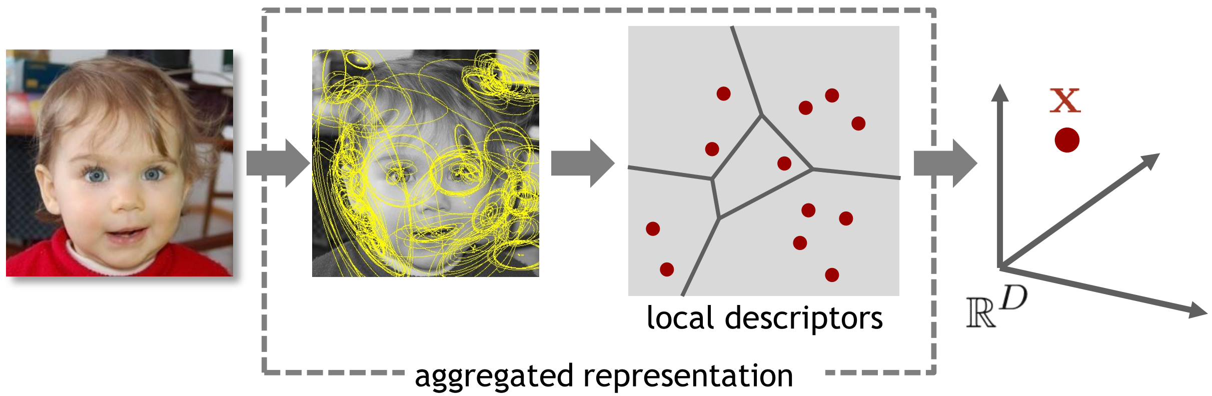
Allegro Workshop

Inria Rhônes-Alpes

22 July 2015

Vector visual representation

- Fixed-size image *representation* $\mathbf{x} \in \mathbb{R}^D$
  - High-dim $(100 \sim 100{,}000)$
  - Generic, unsupervised: BoW, FV, VLAD / DBM, SAE
  - Generic, supervised: learned aggregators / CNN activations
  - Class-specific, e.g. for faces: landmark-related SIFT, HoG, LBP, FV

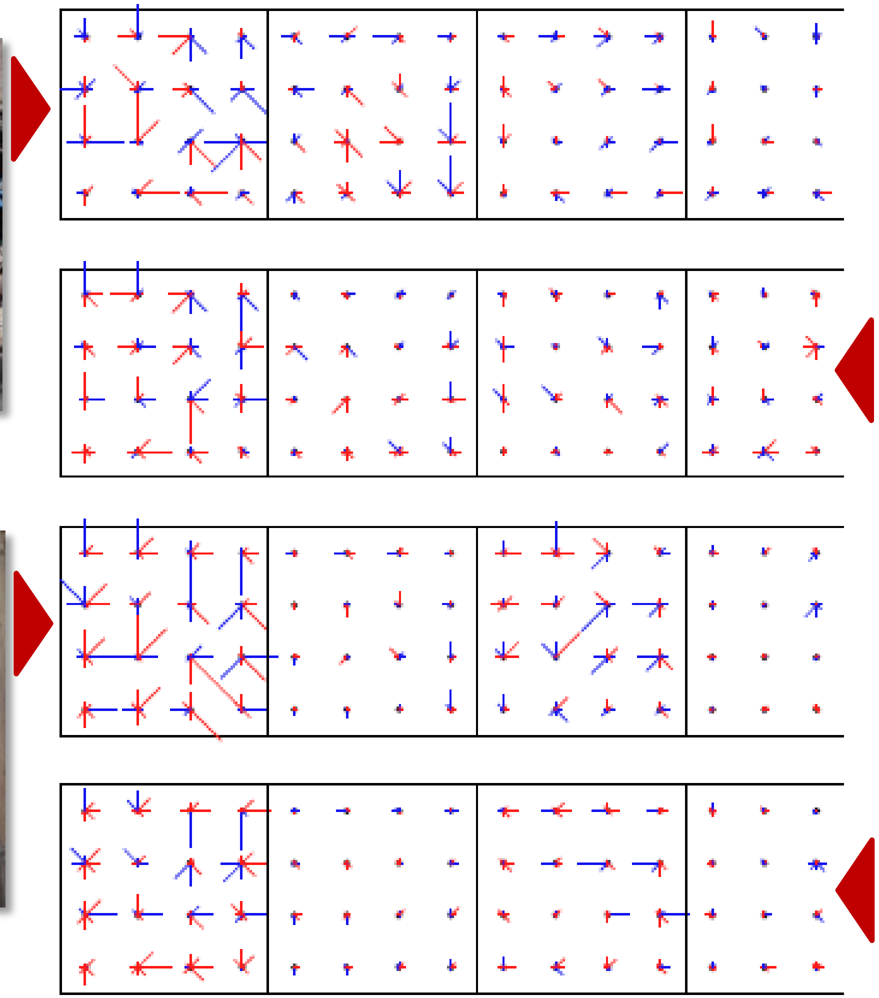


- Key to “compare” images and fragments, with built-in invariance
  - Verification (1-to-1)
  - Search (1-to-*N*)
  - Clustering (*N*-to-*N*)
  - Recognition (1-to-*K*)

VLAD: vector of locally aggregated descriptors

- C SIFT-like blocks, $D = 128 \times C$

$$\mathbf{x} = (\quad \mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_C)$$

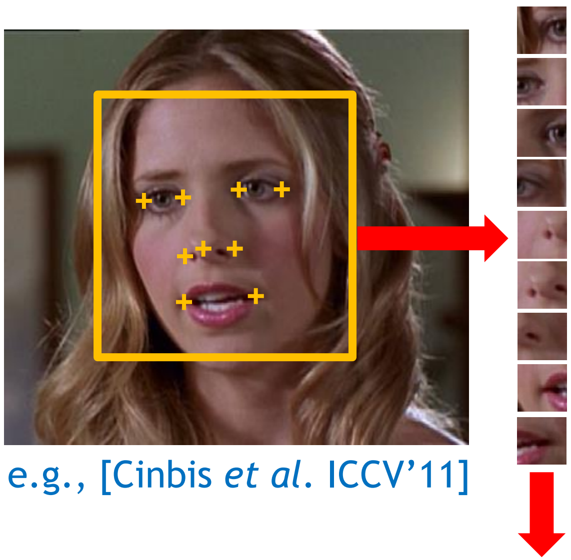


[Jégou *et al*. CVPR'10]

technicolor

Face representation

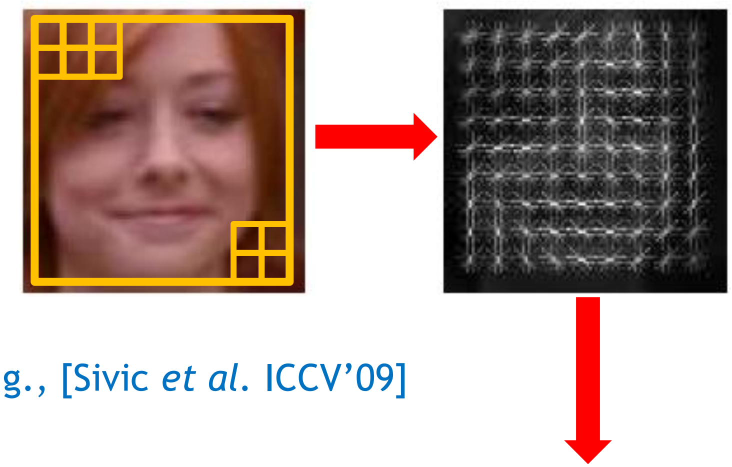
- Sparse representation
  - Layout of facial landmarks
  - Multi-scale descriptor of facial landmarks



e.g., [Cinbis *et al*. ICCV'11]

$$\mathbf{x} \in \mathbb{R}^{9\times3\times128}$$

- Dense representation
  - Fixed grid of overlapping blocks
  - SIFT/HOG/LBP block description
  - Fisher and CNN variants
  - Landmarks still useful to normalize



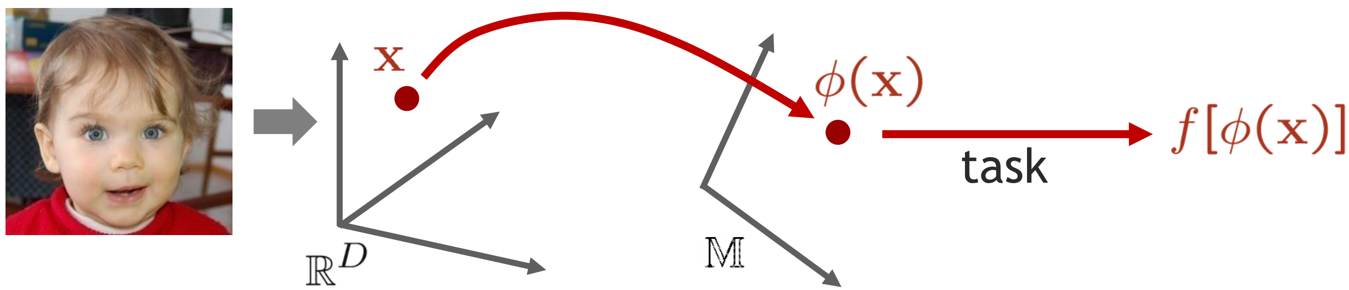
e.g., [Sivic *et al*. ICCV'09]

$$\mathbf{x} \in \mathbb{R}^{9\times9\times4\times6}$$

technicolor

Embedding visual representation

- Further encoding $\phi(\mathbf{x}) \in \mathbb{M}$ to
  - Reduce complexity and memory
  - Improve discriminative power
  - Specialize to specific tasks

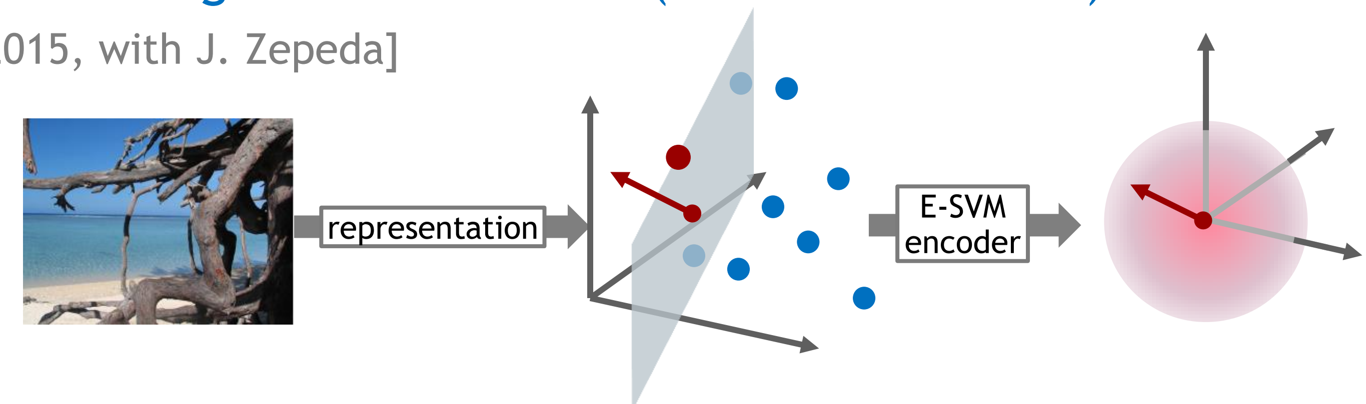


- Various types (possibly combined)
  - Discrete (Hamming, VQ, **PQ**): $\mathbb{M} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\},\ K = 2^B$
  - Linear (PCA, **metric learning**): $\mathbb{M} = \mathbb{R}^E,\ E < D$
  - Non-linear (**K-PCA**, spectral, NMF, SC): $\mathbb{M} \subset \mathbb{R}^E$

Outline

- Explicit embedding for visual search
  [JMIV 2015, with A. Bourrier, H. Jégou, F. Perronin and R. Gribonval]

- E-SVM encoding for visual search (and classification)
  [CVPR 2015, with J. Zepeda]



- Multiple metric learning for face verification
  [ACCV 2014, CVPR-w 2015, with G. Sharma and F. Jurie]

technicolor

Euclidean (approximate) search

- Nearest neighbor (1NN) search in $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^{N} \subset \mathbb{R}^D$

$$\arg\min_{\mathbf{x}\in\mathcal{X}} d(\mathbf{q},\mathbf{x}) \text{ or } \arg\max_{\mathbf{x}\in\mathcal{X}} s(\mathbf{q},\mathbf{x})$$

- Euclidean case

$$\arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{q}-\mathbf{x}\|_2^2 \text{ or } \arg\max_{\mathbf{x}\in\mathcal{X}} \langle\mathbf{q},\mathbf{x}\rangle$$

- Euclidean *approximate* NN (a-NN) for large scale
  - Discrete embedding efficient to search with: binary hashing or VQ
  - Product Quantization (PQ) [Jégou 2010]: asymmetric fine grain search

$\phi(\mathbf{x}) = [\phi_1(\mathbf{x}_1), \cdots, \phi_R(\mathbf{x}_R)],\ \phi_r : \mathbb{R}^{D/R} \mapsto \mathbb{M}_r \subset \mathbb{R}^{D/R}$

$B = R \times B_s$ bits code with sub-quantizers on 2^{B_s} values

$\arg\min_{\mathbf{x}\in\mathcal{X}} \sum_{r=1}^{R} \|\mathbf{q}_r - \phi_r(\mathbf{x}_r)\|_2^2$

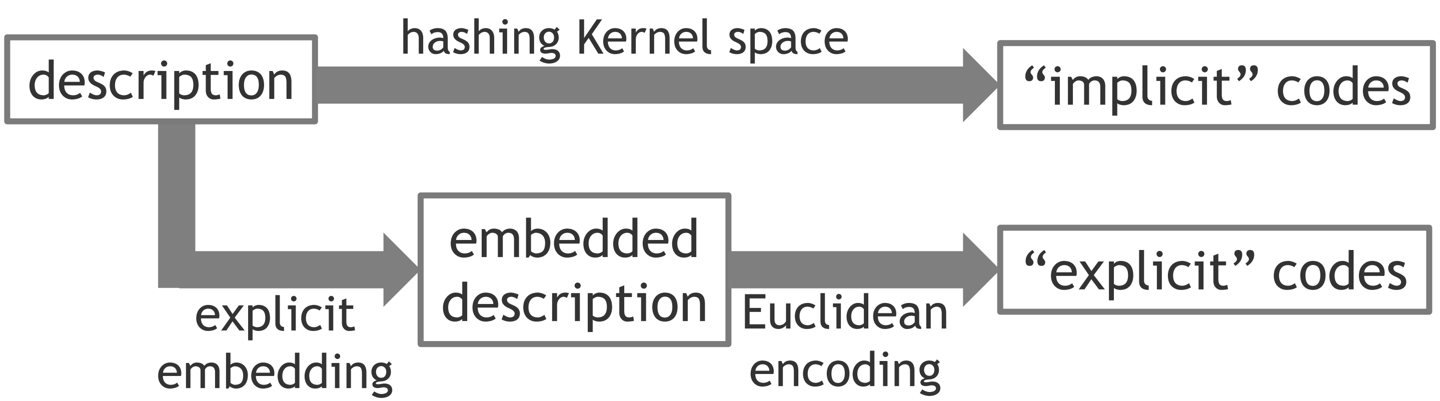
$D \times 2^{B_s}$ distances and $(R-1) \times N$ sums for search

technicolor

Beyond Euclidean

- Other (di)similarities
  - χ^2 and histogram intersection (HI) kernels
  - Data-driven kernels

  Appealing but costly
- Fast approximate search with Mercer kernels?
  - Exploiting of kernel trick to transport techniques to implicit space
  - Inspiration from classification with *explicit embedding*
    [Vedaldi and Zisserman, CVPR'10][Perronnin *et al*. CVPR'10]



The implicit path

- Kernelized Locality Sensitive Hashing (KLSH)
  [Kulis and Grauman ICCV'09]
  - Random draw of directions within RKHS subspace spanned by implicit maps of a random subset of input vectors
  - Hashing function computed thanks to kernel trick

- Random Maximum Margin Hashing (RMMH)
  [Joly and Buisson CVPR'11]
  - Each hashing function is a kernel SVM learned on a random subset of input vectors (one half labeled +1, the other -1)

$$h(\mathbf{x}) = \mathsf{sign}\Big(\sum_{m=1}^{M} y_m \alpha_m K(\mathbf{x}, \mathbf{z}_m) + b\Big)$$

  - Outperforms KLSH

technicolor

Explicit embedding

- Data-independent
  - Truncated expansions or Fourier sampling
  - Restricted to certain kernels (e.g., additive, multiplicative)
- Generic data-driven: Kernel PCA (KPCA) and the like
  - Mercer kernel *K* to capture similarity
  - Learning subset $\mathcal{Z} = \{\mathbf{z}_1 \cdots \mathbf{z}_M\}$
  - Low-rank approximation of kernel matrix $\mathbf{K} = [K(\mathbf{z}_i, \mathbf{z}_j)] \succeq 0$

$$\mathbf{K} = U\Lambda U^\top \approx U_E \Lambda_E U_E^\top, \quad D < E \ll M$$

$$\phi(\mathbf{z}_m) = \Lambda_E^{\frac{1}{2}} U_E^\top$$

$$\phi(\mathbf{x}) = \Lambda_E^{-\frac{1}{2}} U_E^\top \mathbf{k}, \quad \mathbf{k} = [K(\mathbf{x}, \mathbf{z}_m)]_{m=1}^{M}$$

$$\phi_e(\mathbf{x}) = \lambda_e^{-\frac{1}{2}} \langle \mathbf{u}_e, \mathbf{k} \rangle = \lambda_e^{-1} \sum_{m=1}^{M} K(\mathbf{x}, \mathbf{z}_m) \phi_e(\mathbf{z}_m)$$

technicolor

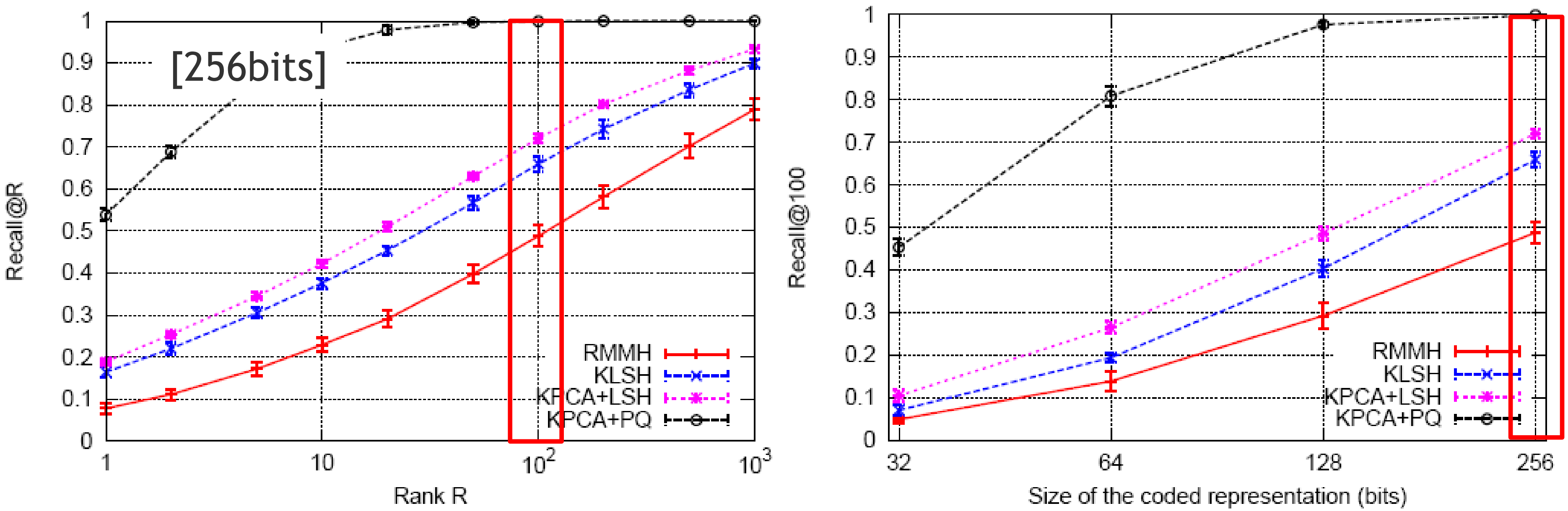
NN and a-NN search with KPCA

- Exact search
  - KPCA encoding
  - Exact Euclidean 1NN search
  - Bound computation
  - Most similar item is in short list truncated with bounds
- Approximate search
  - KPCA encoding
  - Euclidean a-kNN search with PQ
  - Similarity re-ranking of short list

technicolor

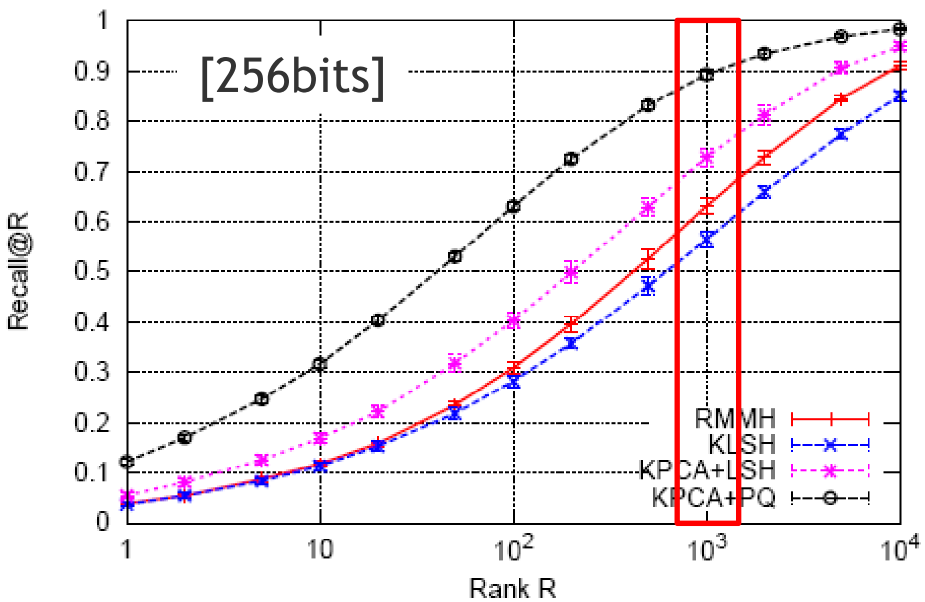
Experiments

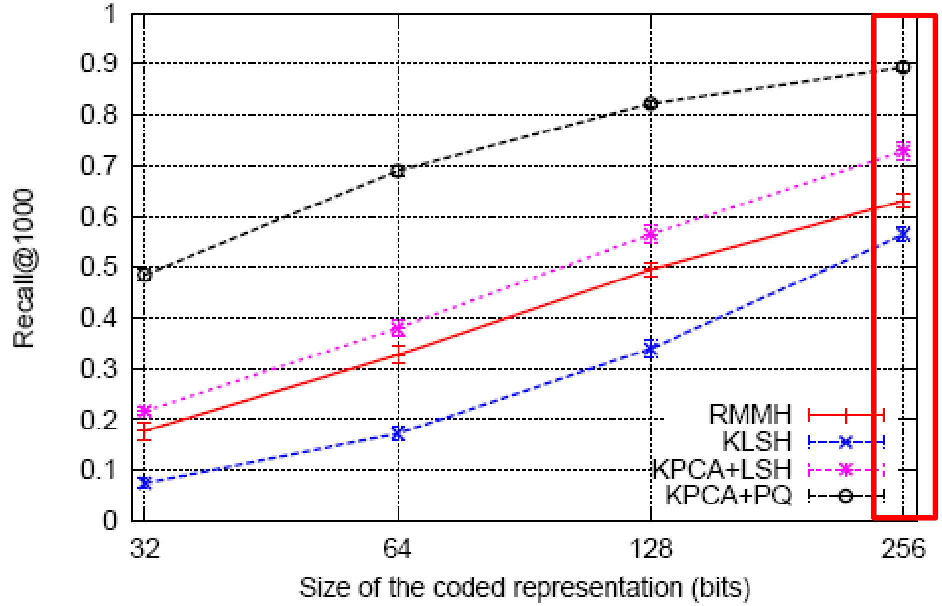
- 1NN local descriptors search
  - N=1M SIFT (D=128), $K=\chi^2$, M=1024, E=128,
  - Tested also: KPCA+LSH (binary search in explicit space)



Experiments

- 1NN image search
  - N=1.2M images BoW (D=1000), K=χ^2, M=1024, E=128
  - Tested also: KPCA+LSH (binary search in explicit space)





technicolor

Discriminative encoding with E-SVM

- Boost discriminative power of representation
  - Extract what is “unique” about image (representation) relative to all others

- Method
  - Exemplar-SVM (E-SVM) [Malisiewicz 2012] to encode visual representation
  - Symmetrical encoding even for asymmetric problems
  - Recursive encoding

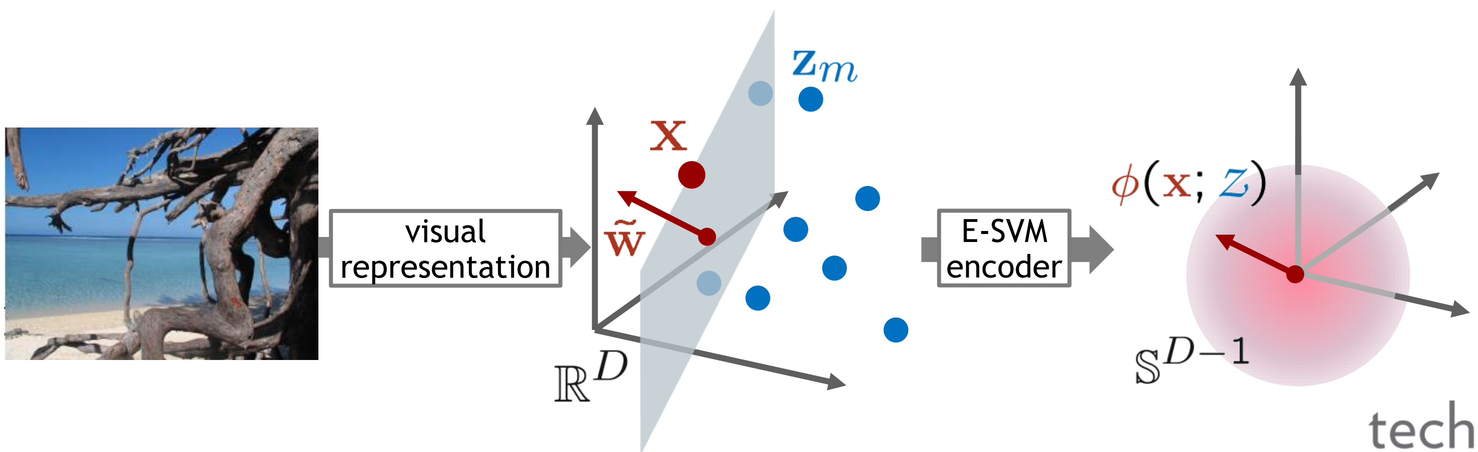
- Application: search and classification

Method

- Large "generic" set of images $\mathcal{Z} = \{\mathbf{z}_m\}_{m=1}^{M} \subset \mathbb{R}^D$

- Exemplar-SVM
$$\begin{aligned}\tilde{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^D} \Big[& \frac{1}{2}\|\mathbf{w}\|_2^2 \\ & + \alpha_+ \max(0, 1 - \mathbf{x}^\top \mathbf{w}) \\ & + \alpha_- \sum_{m=1}^{M} \max(0, 1 + \mathbf{z}_m^\top \mathbf{w}) \Big]\end{aligned}$$

- Final encoding $\phi(\mathbf{x}; \mathcal{Z}) = \dfrac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|_2}$



technicolor

Method

- E-SVM learning: stochastic gradient (SGD) with Pegasos
- Recursive encoding (RE-SVM)

$$\mathbf{w}^{(1)} = \phi(\mathbf{x}; \mathcal{Z}) \qquad \mathcal{Z}^{(1)} = \phi(\mathcal{Z}; \mathcal{Z})$$
$$\mathbf{w}^{(k+1)} = \phi(\mathbf{w}^{(k)}; \mathcal{Z}^{(k)}) \quad \mathcal{Z}^{(k+1)} = \phi(\mathcal{Z}^{(k)}, \mathcal{Z}^{(k)})$$

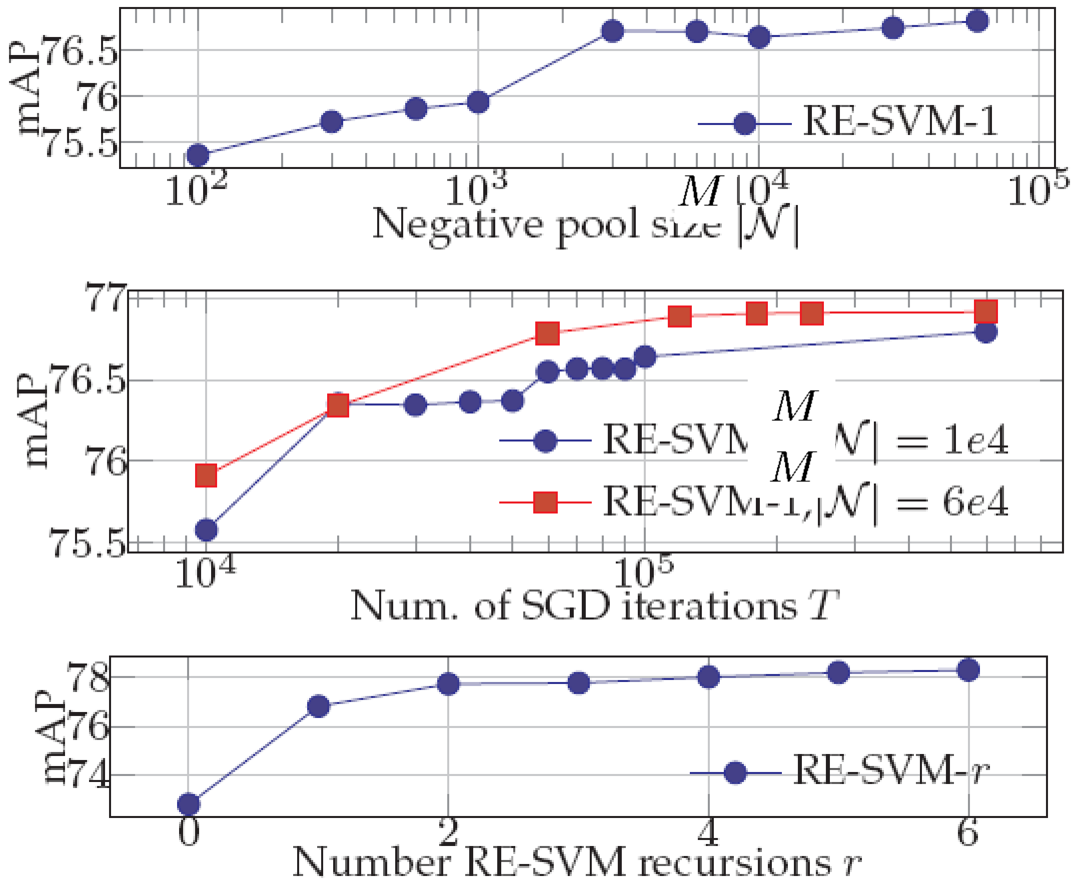
- Image search: symmetrical embedding
  - Query and database codes: $\mathbf{w}_0$ and $\{\mathbf{w}_n\}_{n=1}^N$
  - Cosine similarity: $\langle \mathbf{w}_0, \mathbf{w}_n \rangle$

- Classification: learn and run classifier on E-SVM codes

technicolor

Image search

- *Holiday* dataset, VLAD-64 (*D*=8192)



$M = 60,000$

$T = 10^5$

technicolor

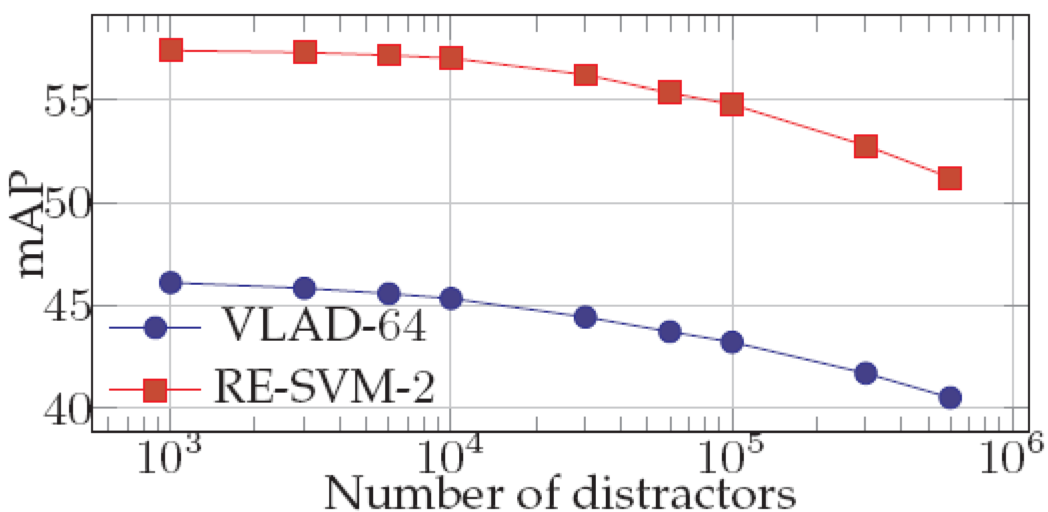
Image search

- *Holiday* and *Oxford* datasets

$47 \rightarrow 4$

$92 \rightarrow 29$

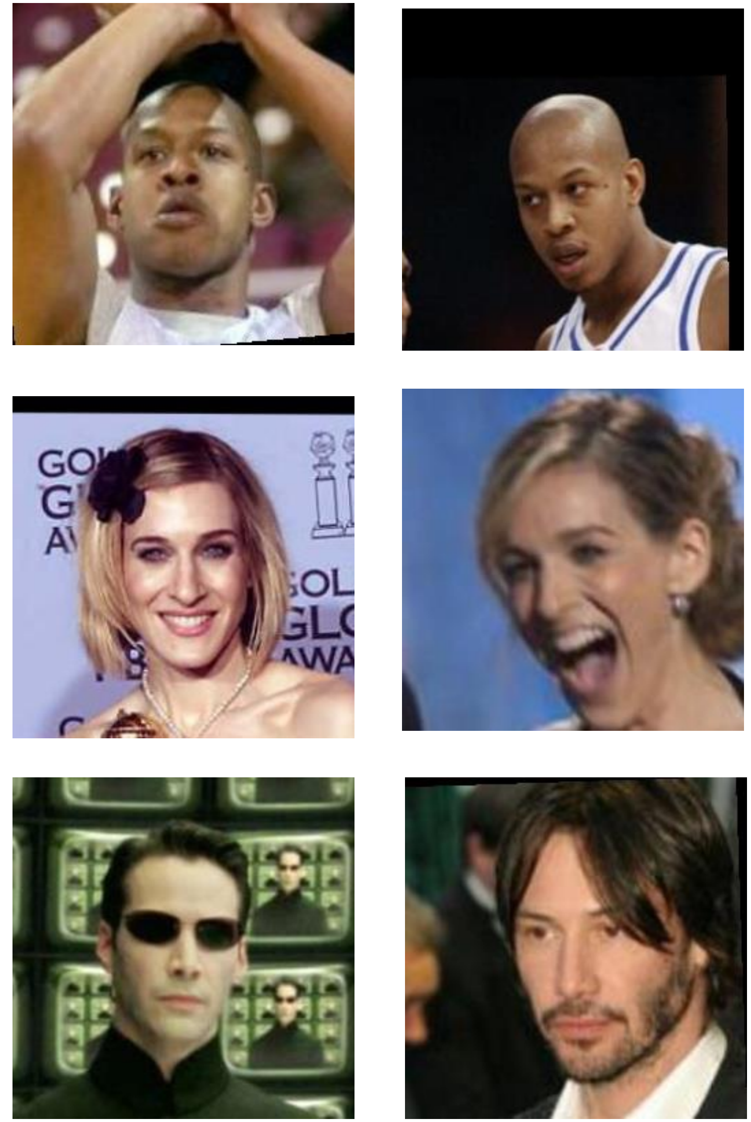
	Holidays	Oxford
VLAD-64 [1]	72.7	46.3
VLAD-64 + RE-SVM-1	77.5	55.5
VLAD-64 + RE-SVM-2	**78.3**	**57.5**
CNN [2]	68.2	40.6
CNN [2] + RE-SVM-2	71.8	44.6



technicolor

Face verification

- Given 2 face images: **Same person?**
  - Persons unseen before
- Various types of supervision for learning
  - Named faces (provide +/- pairs)
  - Tracked faces (provide + pairs)
  - Simultaneous faces (provide – pairs)
- *Labelled Faces in the Wild* (LFW)
  - +13,000 faces; +4,000 persons
  - 10-fold testing with 300 +/- pairs per fold
  - Restricted setting: only pair information for training
  - Unrestricted setting: name information for training



technicolor

Linear metric learning

- Powerful approach to face verification
- Learning Mahalanobis distance in input space $\mathbb{R}^D$, via $M \succeq 0$

$$d^2_M(\mathrm{x}, \mathrm{x}') = (\mathrm{x} - \mathrm{x}')^\top M (\mathrm{x} - \mathrm{x}')$$

- Typical training data: $\mathcal{T} = \{(\mathrm{x}, \mathrm{x}', y_{\mathrm{xx}'})\} \subset \mathbb{R}^{2D} \times \{-1, +1\}$
  - +/- pairs should become close/distant
- Verification of new faces: $y_{\mathrm{xx}'} = \mathrm{sign}(1 - d^2_M(\mathrm{x}, \mathrm{x}'))$
- Several approaches
  - Large margin nearest neighbor (LMNN) [Weinberger *et al*. NIPS'05]
  - Information theoretic metric learning (ITML) [Davis et al. ICML'07]
  - Logistic Discriminant Metric Learning (LDML) [Guillaumin et al. ICCV'09]
  - Pairwise Constrained Component Analysis (PCCA) [Mignon & Jurie, CVPR'12]

Low-rank metric learning

- Very high dimension (in range 1,000 ~100,000)
  - Prohibitive size of Mahalanobis matrix
  - Scarcity of training data
- Low-rank Mahalanobis metric learning: $M = L^\top L,\ L \in \mathbb{R}^{E\times D},\ E \ll D$

$$\begin{aligned} d_L^2(\mathbf{x}, \mathbf{x}') &= (\mathbf{x} - \mathbf{x}')^\top M (\mathbf{x} - \mathbf{x}') \\ &= \|L\mathbf{x} - L\mathbf{x}'\|_2^2 \end{aligned}$$

  - Learn linear projection (dim. reduction) and metric
- Minimize loss over training set

$$\min_{L,b} \sum_{(\mathbf{x},\mathbf{x}',y_{\mathbf{xx}'})\in\mathcal{T}} \text{loss}[d_L^2(\mathbf{x}, \mathbf{x}'), y_{\mathbf{xx}'}; b]$$

  - Rank fixed by cross-validation
- Proposed: extension to **latent variables** and **multiple metrics**

technicolor

Losses

- Probabilistic logistic loss

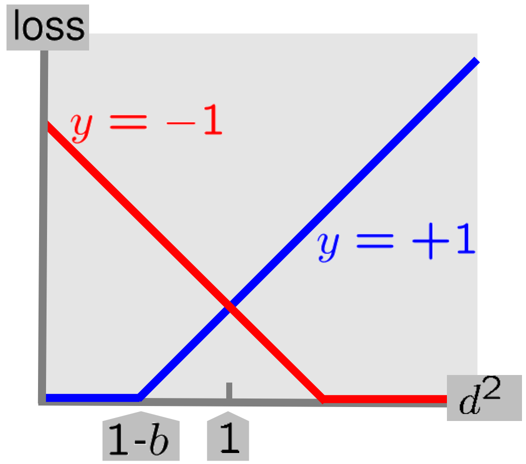
$$1 + y_{\mathbf{x}\mathbf{x}'} \tanh[-\frac{1}{2}(d_L^2(\mathbf{x},\mathbf{x}') - b)]$$

- Generalized logistic loss

$$\frac{1}{\beta}\log(1 + \exp[\beta y_{\mathbf{x}\mathbf{x}'}(d_L^2(\mathbf{x},\mathbf{x}') - b)])$$

- Hinge loss

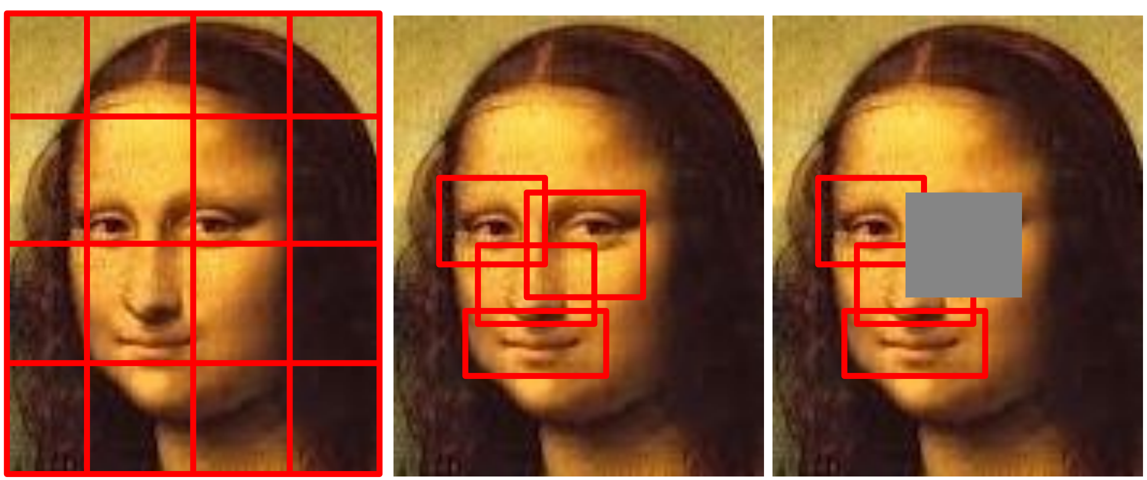
$$\max\left[0, 1 - y_{\mathbf{x}\mathbf{x}'}(b - d_L^2(\mathbf{x},\mathbf{x}'))\right]$$



Expanded parts model

- Expanded parts model
  [Sharma *et al*. CVPR’13]
  for human attributes
  and object/action recog.

- Objectives
  - Avoid fixed layout
  - Learn collection of **discriminative parts** and associated **metrics**
  - Leverage the model to handle **occlusions**



Expanded parts model

- Mine P discriminative parts and learn associated metrics $\mathcal{L} = \{L_p\}_{p=1}^P$
- Dissimilarity based on comparing $K < P$ best parts

$$d_{\mathcal{L}}^2(\mathbf{x}, \mathbf{x}') = \min_{\boldsymbol{\alpha} \in \{0,1\}^P} \sum_{p=1}^{P} \alpha_p \| L_p(\mathbf{x}_p - \mathbf{x}'_p) \|_2^2$$

$$\text{sb.t. } \|\boldsymbol{\alpha}\|_0 = K, \text{ and overlap}(\boldsymbol{\alpha}) < \theta$$

- Learning
  - Minimize **hinge loss**: greedy on parts + gradient descent on matrices
  - Prune down to P a large set of N random parts
  - Projections initialized by whitened PCA
  - Stochastic gradient: given annotated pair $(\mathbf{x}, \mathbf{x}', y_{\mathbf{x}\mathbf{x}'})$

$$\text{if } y_{\mathbf{x}\mathbf{x}'}(b - d_{\mathcal{L}}^2(\mathbf{x}, \mathbf{x}')) < 1$$

$\forall p \in$ support of $\boldsymbol{\alpha}^*$:

$$\partial_{L_p} \text{loss} = y_{\mathbf{x}\mathbf{x}'} L_p (\mathbf{x}_p - \mathbf{x}'_p)(\mathbf{x}_p - \mathbf{x}'_p)^\top$$

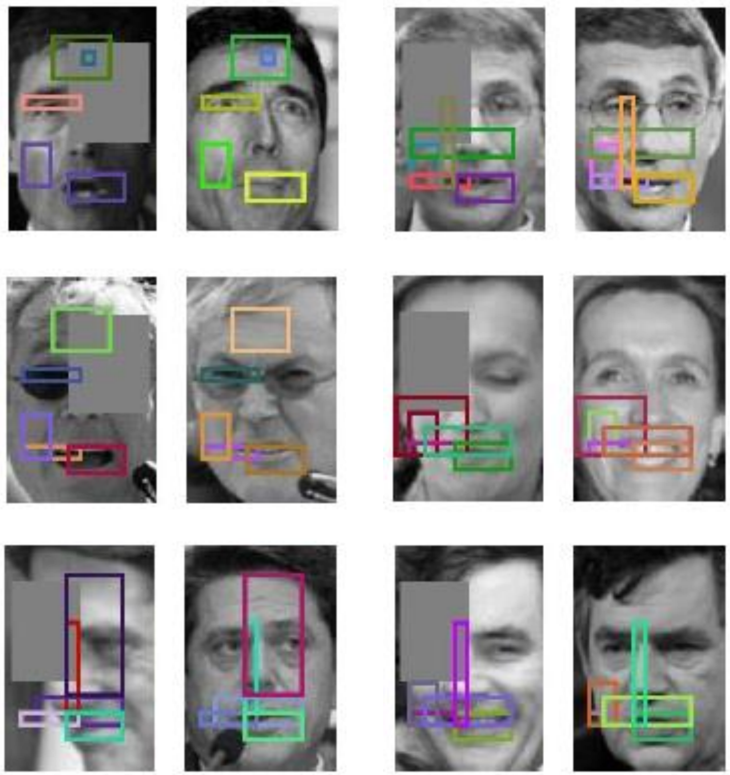
technicolor

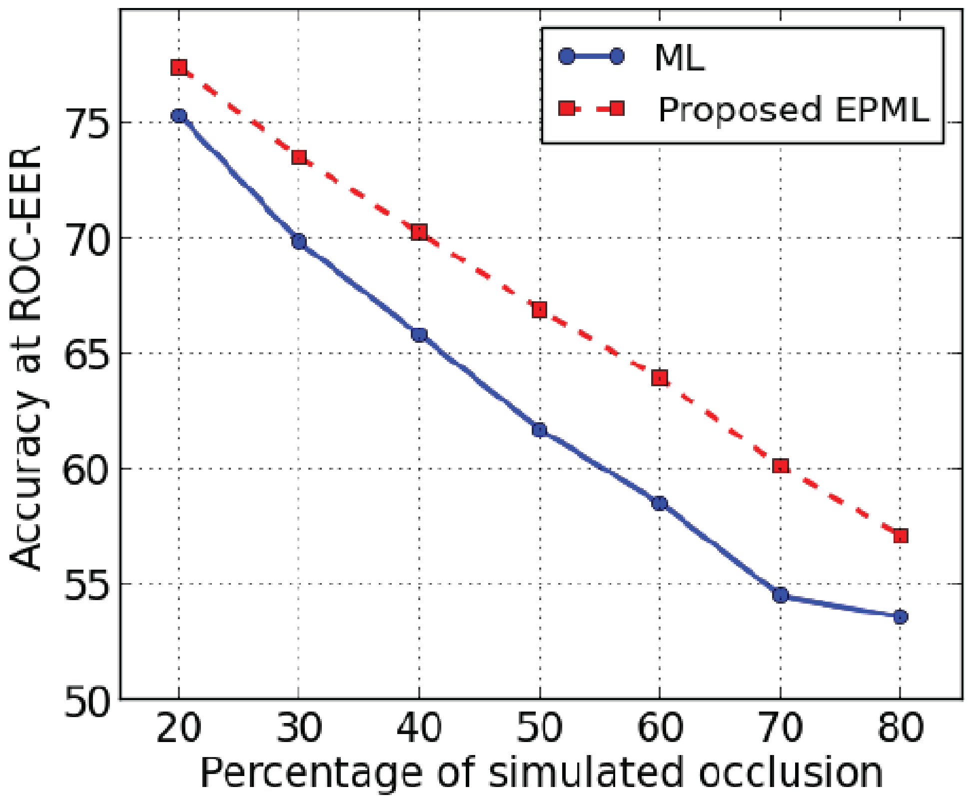
Experiments with occlusions

- LFW, unrestricted setting
  - $N = 500$, $P \sim 50$, $K = 20, D = 10k, E = 20$, 10^6 SGD iterations
  - Random occlusions $(20 - 80\%)$ at test time, on one image only

  - Focused occlusions

Experiments with occlusions





	Left eye	Right eye	Both eyes	Nose	Mouth	Nose + mouth
ML	75.5	73.4	61.7	78.0	77.3	73.5
EPML	78.9	77.0	69.2	79.1	78.5	75.5

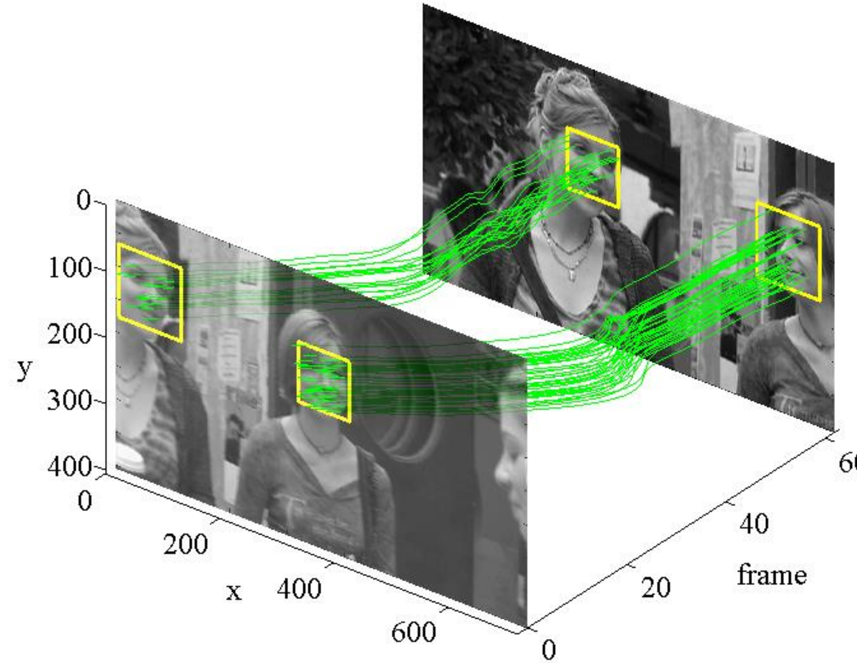
technicolor

Comparing face sets

- Given groups of single-person faces

$$\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^{N} \subset \mathbb{R}^D$$

e.g., labelled clusters, face **tracks**



[Everingham *et al*.BMVC'06]

- Comparing sets
  - Based on face pair comparison, i.e.

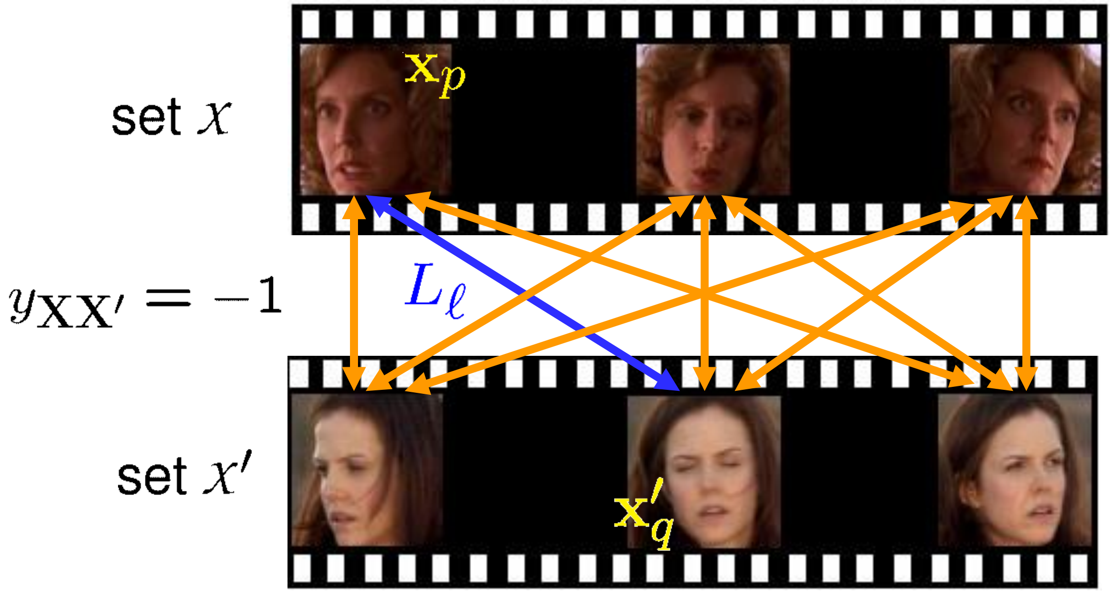
$$D_L^2(\mathcal{X}, \mathcal{X}') = \min_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{x}'\in\mathcal{X}'} d_L^2(\mathbf{x}, \mathbf{x}')$$

  - For face tracks: a single descriptor per track [Parkhi *et al*. CVPR' 14]

Learning multiple metrics

- Metrics associated to L mined types of cross-pair variations

$$D^2_{\mathcal{L}}(\mathcal{X}, \mathcal{X}') = \min_{(\ell,p,q)} \|L_\ell(\mathbf{x}_p - \mathbf{x}'_q)\|_2^2$$



- Learning from annotated set pairs $\mathcal{T} = \{(\mathcal{X}, \mathcal{X}', y_{\mathcal{X}\mathcal{X}'})\}$

$$\min_{\mathcal{L},b} \sum_{(\mathcal{X},\mathcal{X}',y_{\mathcal{X}\mathcal{X}'})\in\mathcal{T}} \text{loss}[D^2_{\mathcal{L}}(\mathcal{X}, \mathcal{X}', y_{\mathcal{X}\mathcal{X}'}); b]$$

technicolor

Learning multiple metrics

- Stochastic gradient: given annotated pair $(\mathcal{X}, \mathcal{X}', y_{\mathcal{X}\mathcal{X}'})$
  - Subsample the sets (to ensure variety of cross-pair variations)
  - Dissimilarity: $$D^2_{\mathcal{L}}(\mathcal{X}, \mathcal{X}') = \min_{(\ell,p,q)} \|L_\ell(\mathbf{x}_p - \mathbf{x}'_q)\|_2^2 = \|L_{\ell*}(\mathbf{x}_{p*} - \mathbf{x}'_{q*})\|_2^2$$
  - Sub-gradient of pair's **hinge** loss: if $y_{\mathcal{X}\mathcal{X}'}(b - \|L_{\ell*}(\mathbf{x}_{p*} - \mathbf{x}'_{q*})\|_2^2) < 1$

$$\partial_{L_{\ell*}} \text{loss} = y_{\mathcal{X}\mathcal{X}'} L_{\ell*}(x_p - x'_p)(x_p - x'_p)^\top$$

  - Projections initialized by whitened PCA computed on random subsets

technicolor

New dataset

- From 8 different series (inc. Buffy, Dexter, MadMen, etc.)
- 400 high quality labelled face tracks, 23M faces, 94 actors
- Wide variety of poses, attributes, settings
- Ready for metric learning and test (700 pos., 7000 neg.)

technicolor

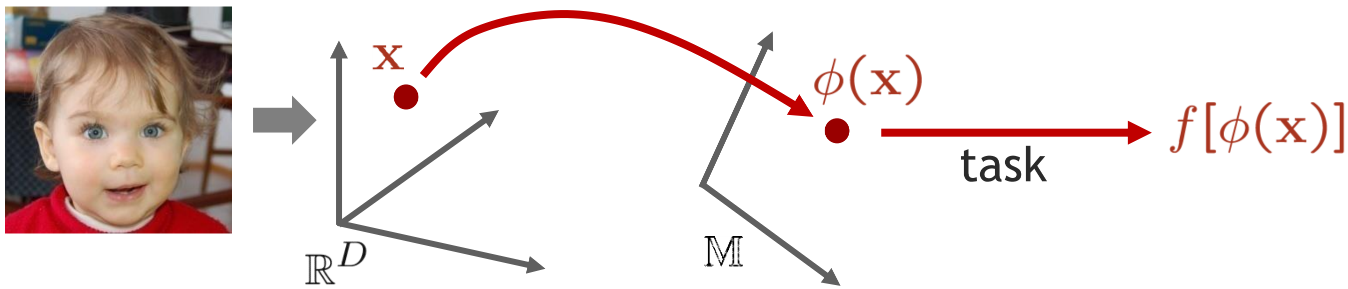
Comparing face tracks

- Parameters: $D \sim 14000,\ K = 3,\ 10^6$ SGD iterations

Method	Subspace dim. E	Aver. Precision known persons	Aver. Precision unknown persons
PCA+cosine sim + min-min	1000	24.8	20.4
PCA+cosine sim + min-min	100	21.4	20.2
Metric Learning + min-min	100	23.7	21.0
Latent ML (proposed)	(3X)33	**27.9**	**22.9**

Conclusion

- Learn embedding of visual description



  - Unsupervised learning of ϕ
  - Task-dependent supervised learning of (ϕ, f)
- Also for deep learning
  - 1-layer adaptation of CNN features for classification with linear SVM
  - Ad-hoc dim. reduction or learned with L1 regularization (Kulkarni et al. BMVC15)
  - Same performance as VGG-M 128 [Chatfield 2014], with 4x smaller codes