Fisher Kernel based Relevance Feedback for Multimodal Video Retrieval

Ionuț MIRONICĂ[1]	Bogdan IONESCU[1,2]	Jasper Uijlings[2]	Nicu Sebe[2]

[1]LAPI – University Politehnica of Bucharest, Romania
[2]MHUG DISI – University of Trento, Italy

ACM International Conference on Multimedia Retrieval,
ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013

Problem Statement

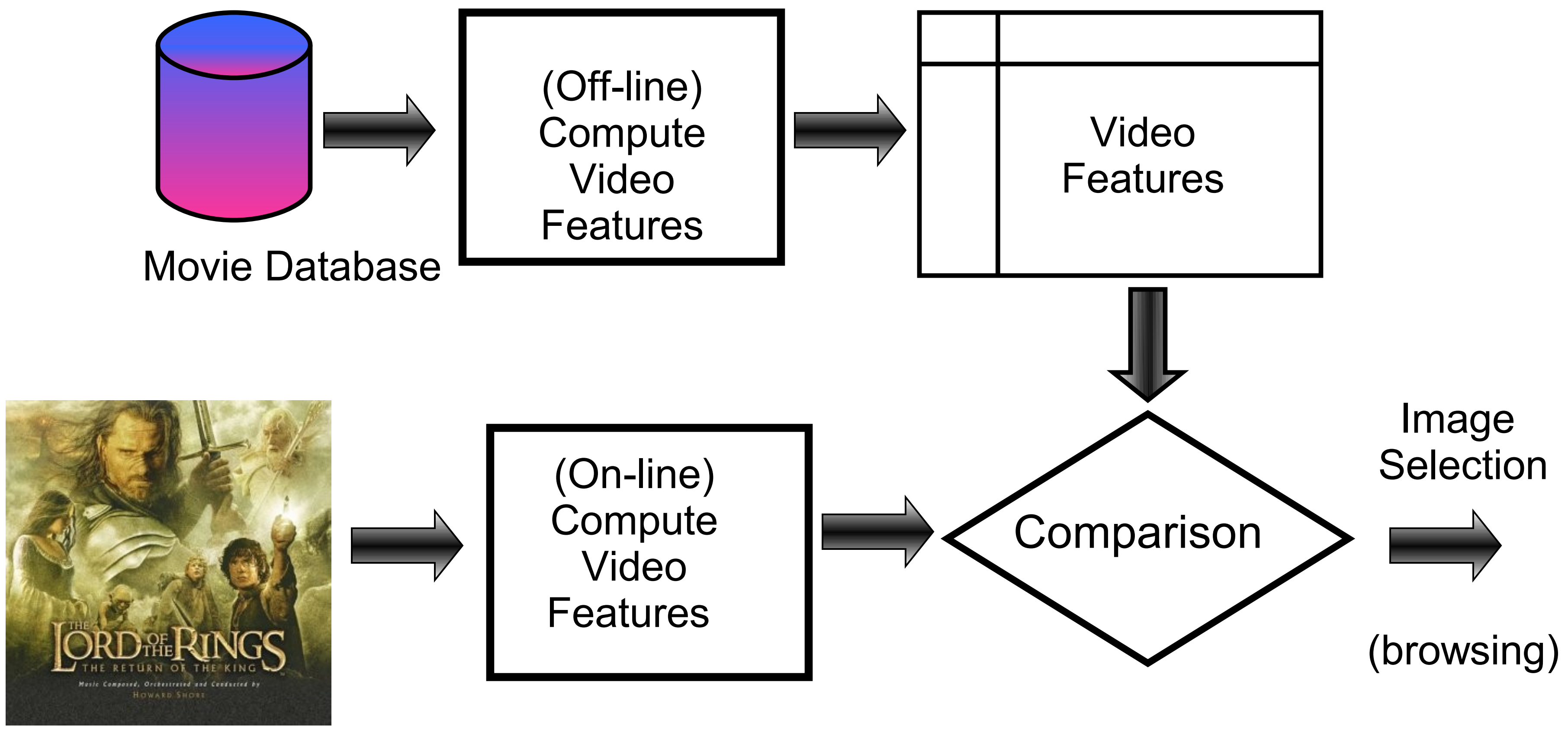
Concepts

- Content Based Video Retrieval
- Genre Retrieval
- Query by Example

Query Database

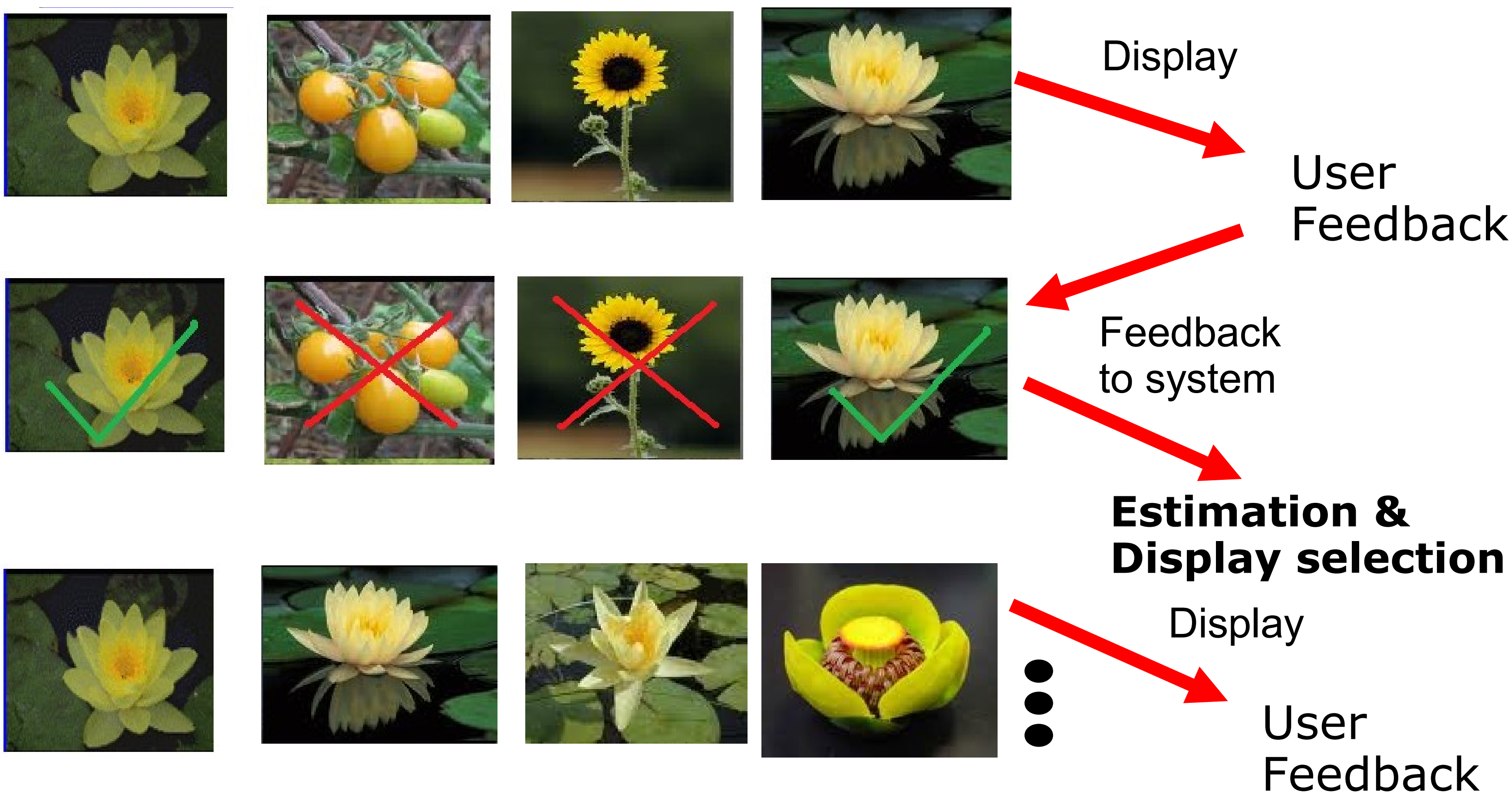
Movie Sample

CBVR System


Movie Database
(Off-line)
Compute
Video
Features
Video
Features
THE LORD OF THE RINGS
THE RETURN OF THE KING
(On-line)
Compute
Video
Features
Comparison
Image
Selection
(browsing)

Relevance Feedback

- **Relevance feedback** uses user-input to improve results



Relevance Feedback

State-of-the-Art algorithms:

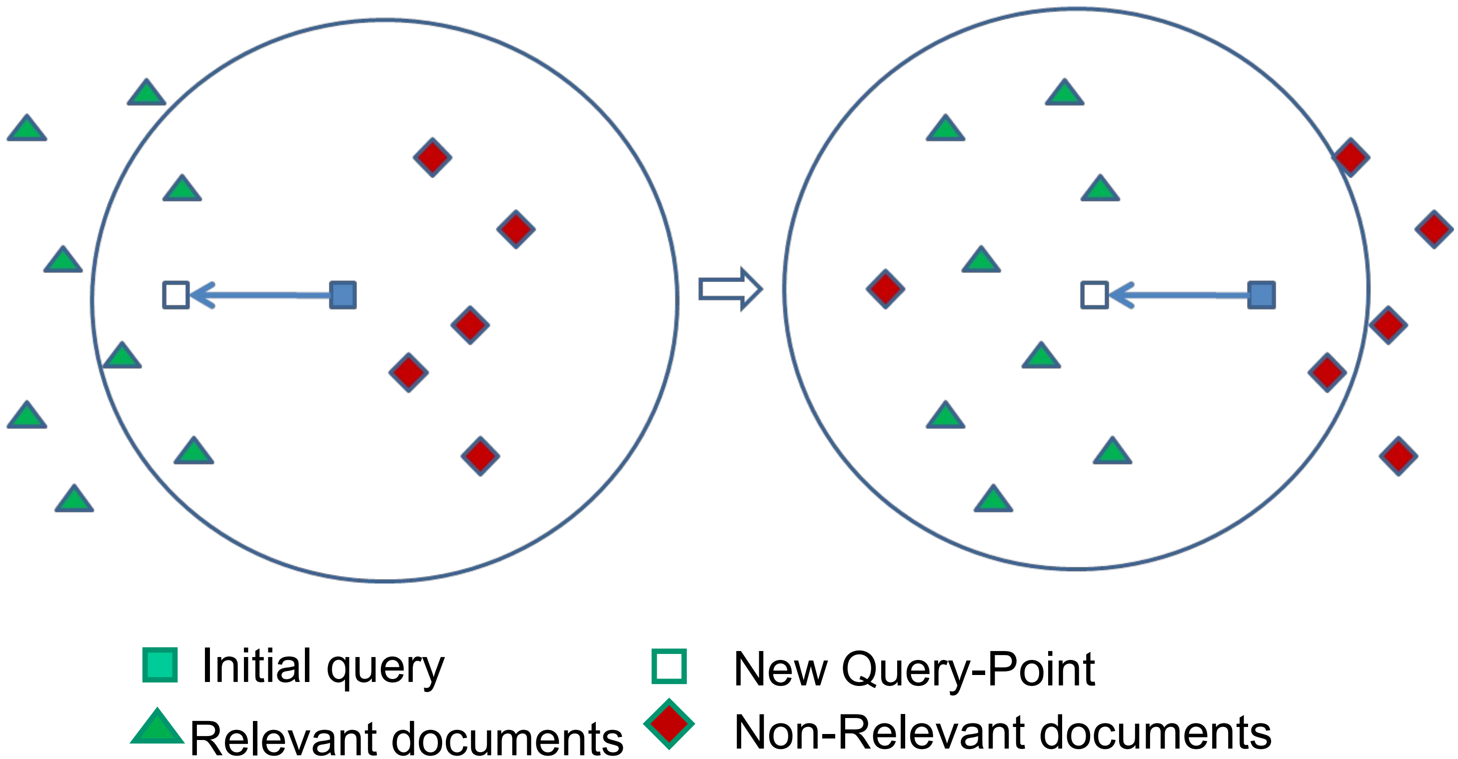
- Rocchio Algorithm
- Feature Relevance Estimation (FRE)
- Train Classifier (SVM)

This paper:

- Fisher Kernel representation + SVM

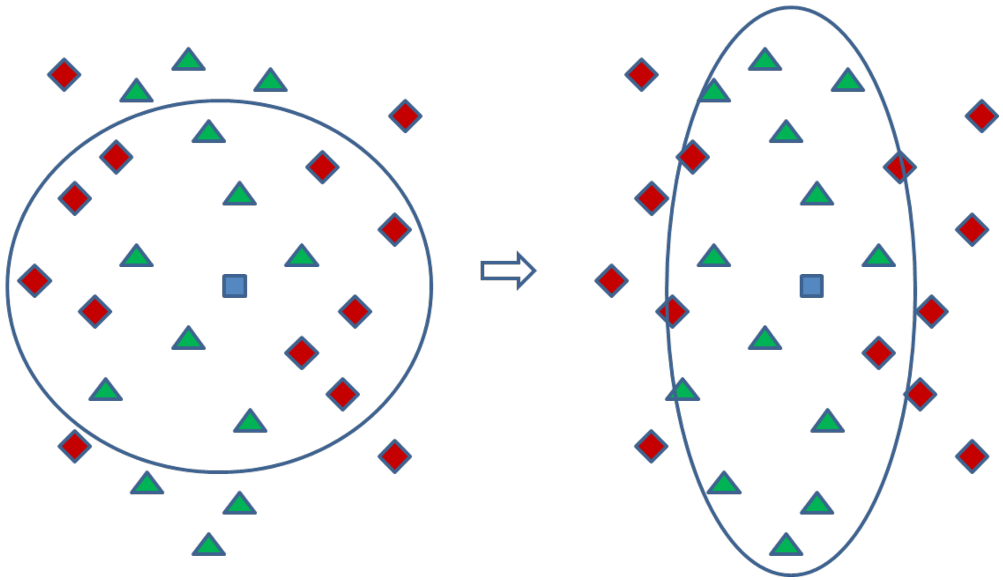
Rocchio Algorithm

- Update the query feature.



Feature Relevance Estimation (FRE)

- Reweigh features according to variance of relevant documents

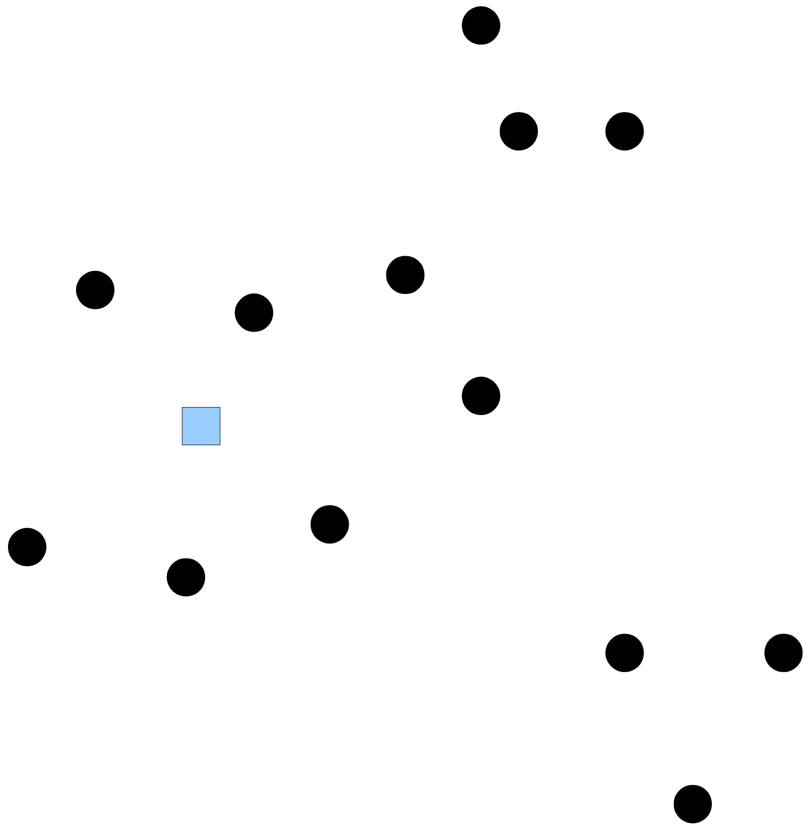


- Initial query
- Relevant documents
- Non-Relevant documents

$$Dist(X,Y)=\sqrt{\frac{\sum_{i=1}^{d}(X_i-Y_i)^2}{\sigma_i}}$$

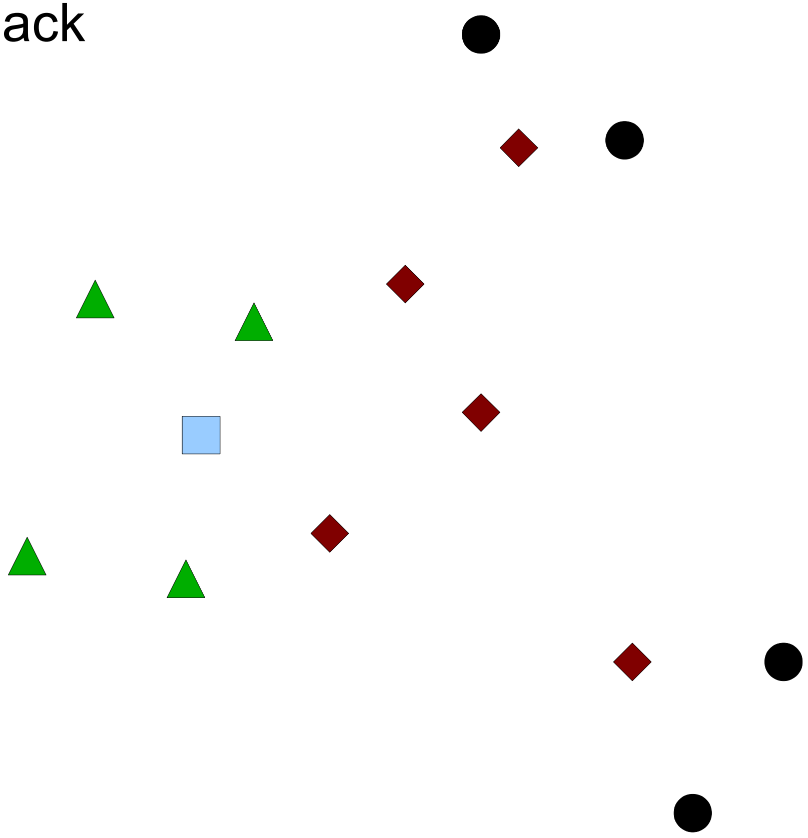
Train Classifier (SVM)

- Initial retrieval



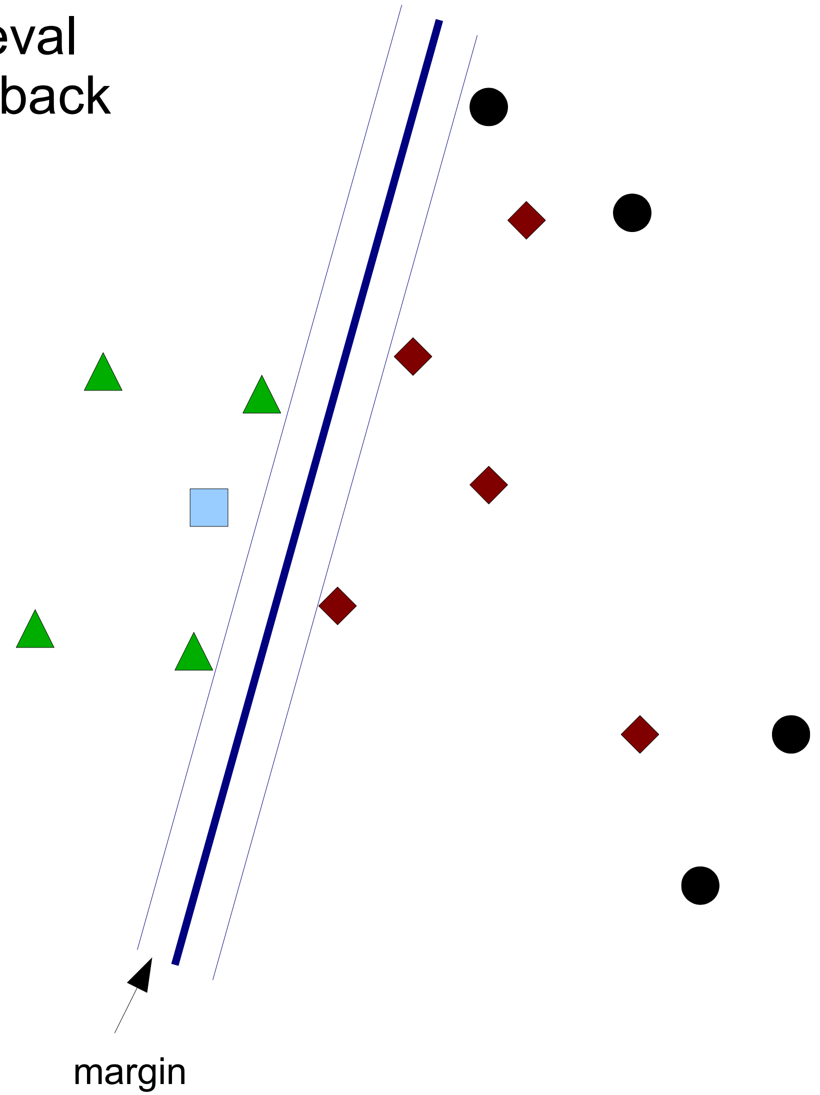
Train Classifier (SVM)

- Initial retrieval
- User feedback



Train Classifier (SVM)

- Initial retrieval
- User Feedback
- Train SVM



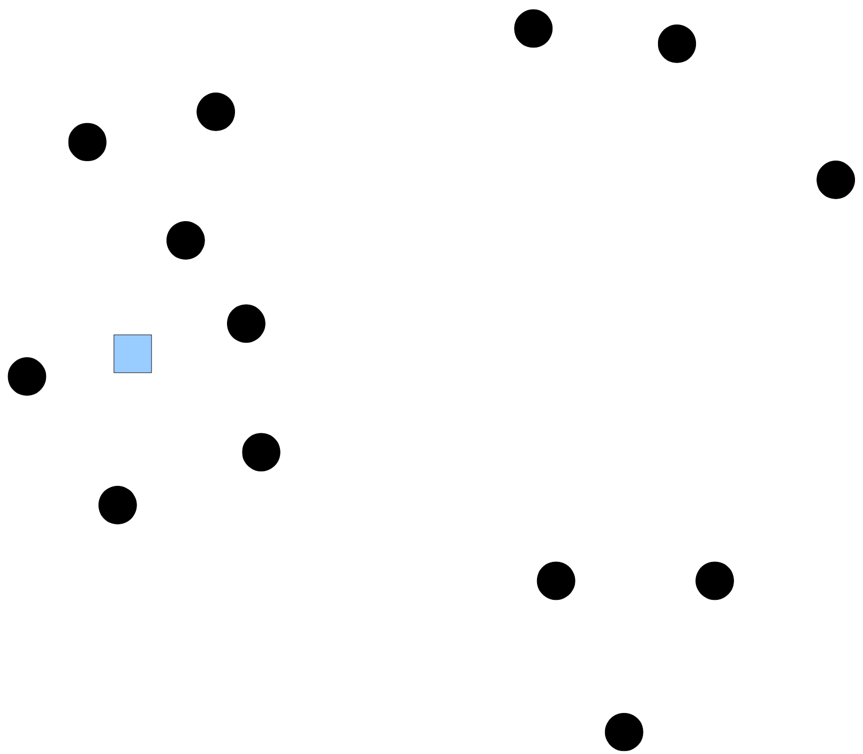
Fisher Kernel (FK)

- represents a signal as the gradient of the probability density function that is a learned generative model of that signal

[Jaakkola and Haussler 1999, Perronnin and Dance 2007]

Fisher Kernel (FK)

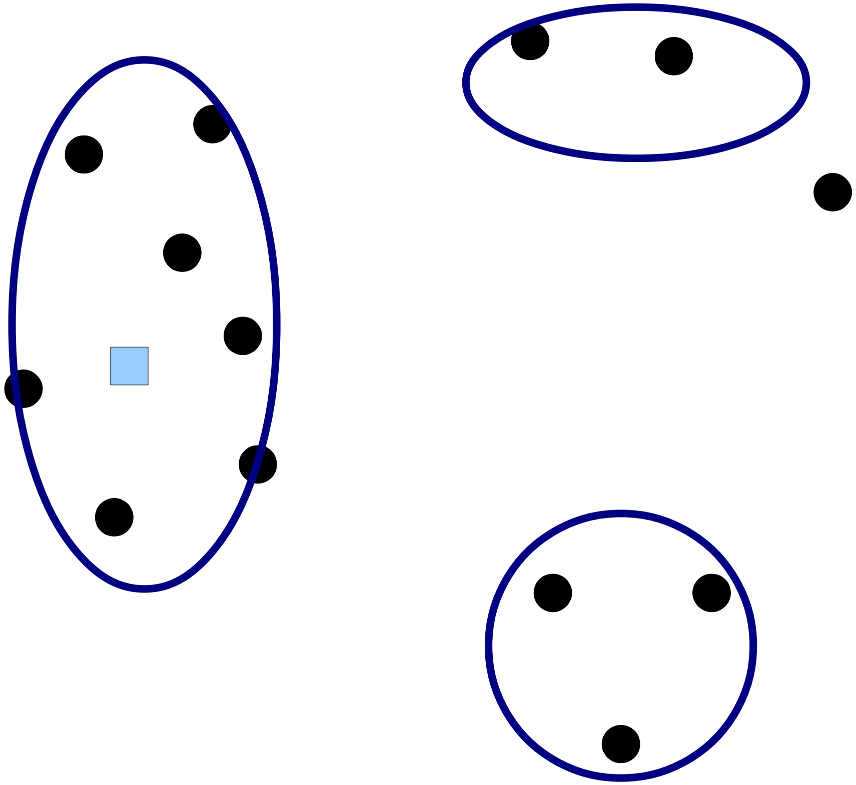
- represents a signal as the gradient of the probability density function that is a learned generative model of that signal



[Jaakkola and Haussler 1999, Perronnin and Dance 2007]

Fisher Kernel (FK)

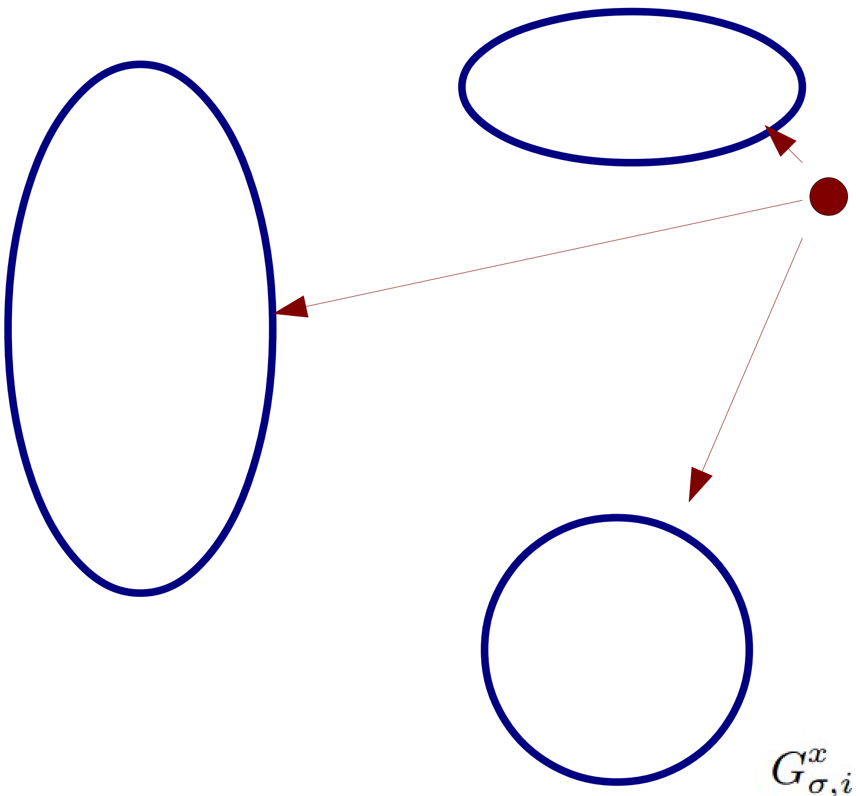
- represents a signal as the gradient of the probability density function that is a learned generative model of that signal



[Jaakkola and Haussler 1999, Perronnin and Dance 2007]

Fisher Kernel (FK)

- represents a signal as the gradient of the probability density function that is a learned generative model of that signal



$$G_{\mu,i}^{x} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i}$$

$$G_{\sigma,i}^{x} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]$$

[Jaakkola and Haussler 1999, Perronnin and Dance 2007]

Fisher Kernel based RF

- Use a video document as a „query by example”

- The system returns top *k* documents (*k=20*)

- Use these document to train an Gaussian Mixture Model (GMM)
- Compute the Fisher Kernels for all the features (reshape the features space)
- Apply Fisher Kernel Normalization

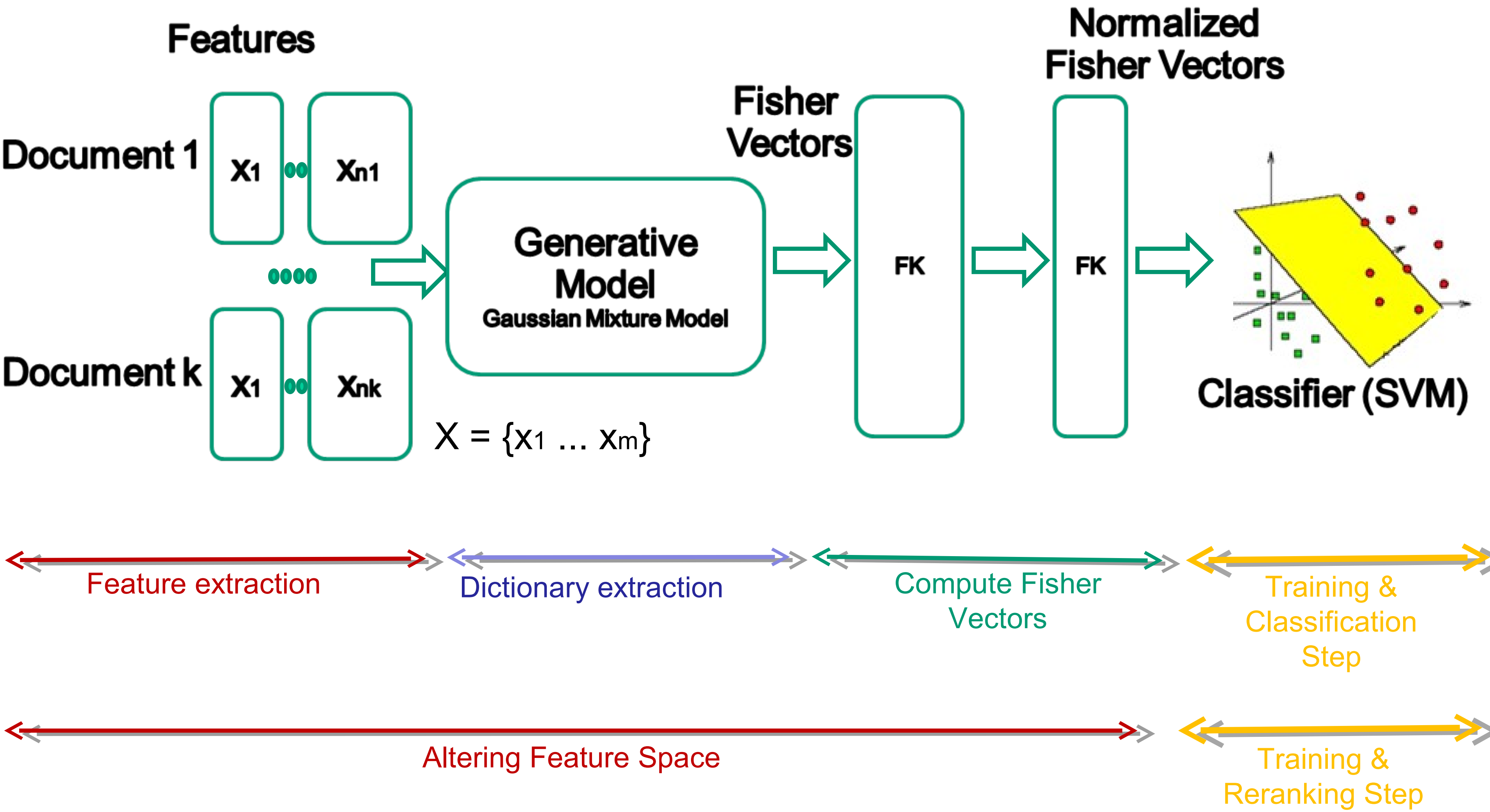
Fisher Kernel based RF

- The user selects the relevant documents

- The system is trained with these samples (SVM classifiers)
- The top N documents are reranked using the classifier confidence level (N=2000)

- If the results are not satisfactory the use can continue with more Relevance Feedback iterations)

Fisher Kernel Framework

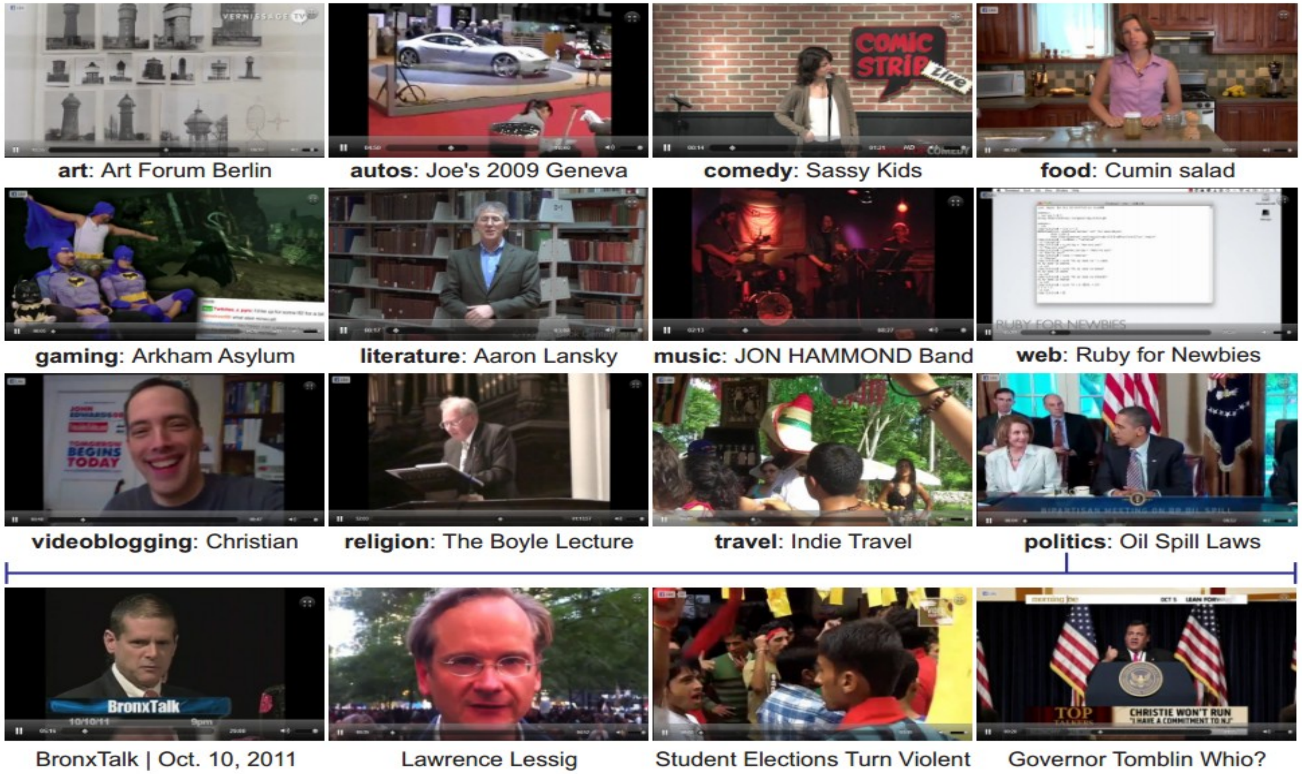


Experimental Setup

MediaEval 2012 Dataset

- 14,838 episodes from 2,249 shows ~ 3,260 hours of data
- split into Development and Test sets

  5,288 for development / 9,550 for test



Experimental Setup

MediaEval 2012 Dataset

- **26 Genre labels**

1000 art	1001 autos_and_vehicles
1002 business	1003 citizen_journalism
1004 comedy	1005 conferences_and_other_events
1006 **default_category**	1007 documentary
1008 educational	1009 food_and_drink
1010 gaming	1011 health
1012 literature	1013 movies_and_television
1014 music_and_entertainment	1015 personal_or_auto-biographical
1016 politics	1017 religion
1018 school_and_education	1019 sports
1020 technology	1021 the_environment
1022 the_mainstream_media	1023 travel
1024 videoblogging	1025 web_development_and_sites

Experimental Setup

- Precision – Recall Chart

$$\text{Precision} = \frac{\text{Number of returned relevant documents}}{\text{Total number of returned documents}}$$

$$\text{Recall} = \frac{\text{Number of returned relevant documents}}{\text{Total number of relevant documents}}$$

- Mean Average Precision

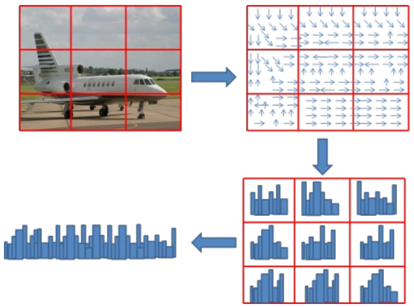
  - summarize rankings from multiple queries by averaging average precision

Simulated Relevance Feedback (one round)

Visual Descriptors

- Global MPEG 7 related descriptors:

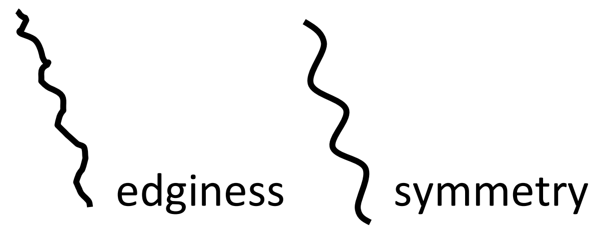
  Local Binary Pattern (LBP), Autocorrelogram,
  Color Coherence Vector (CCV), Edge Histogram (EHD)
  Color Layout Pattern(CLD), Scalable Color Descriptor
  classic color histogram (hist), color moments

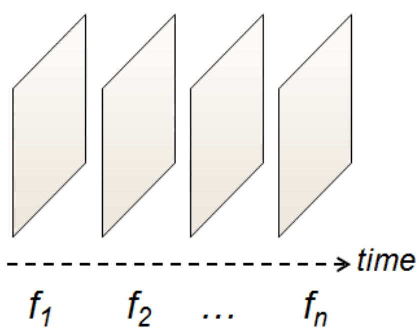


- Global HoG

- Global Structural Descriptors

  describe contour properties and
  appearance parameters [IJCV, C. Rasche'10]

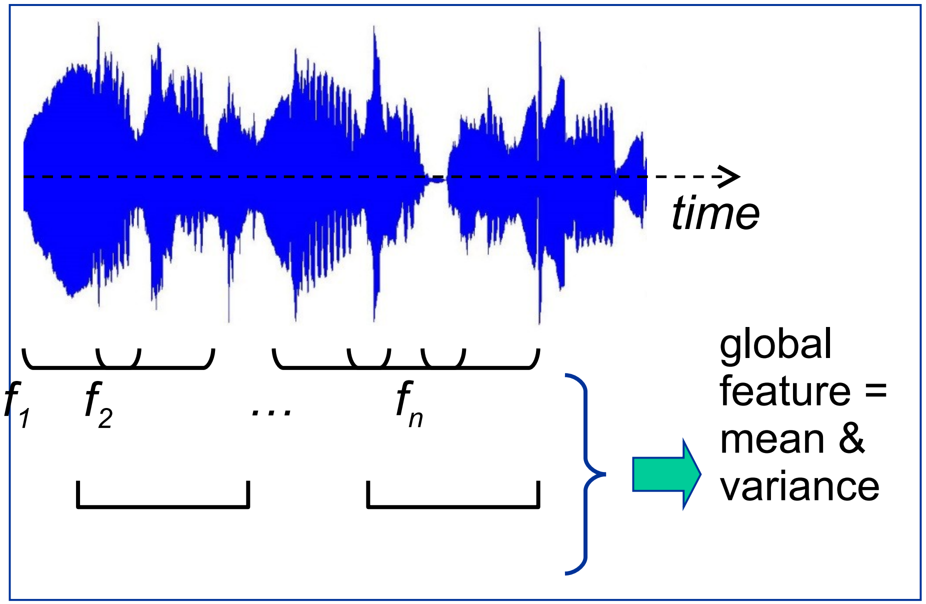




Global Descriptors = mean and variance over all frame descriptors

Audio Descriptors

Standard Audio Features



- Zero-Crossing Rate,
- Linear Predictive Coefficients,
- Line Spectral Pairs,
- Mel-Frequency Cepstral Coefficients
- spectral centroid, flux, rolloff, and kurtosis,

variance of each feature over a certain window.

[B. Mathieu et al., Yaafe toolbox, ISMIR'10]

Block-based Audio Features

- Spectral Pattern, delta Spectral Pattern,
- variance delta Spectral Pattern,
- Logarithmic Fluctuation Pattern ,
- Correlation Pattern, Local Single Gaussian model
- Spectral Contrast Pattern,
- Mel-Frequency Cepstral Coefficients

[Klaus Seyerlehner et al., MIREX'11, USA]

Text Descriptors

Text source: ASR [Lamel, et al., ICNL'08]

TF-IDF descriptors (Term Frequency-Inverse Document Frequency)

1. extract root words

2. remove terms <5%-percentile of the frequency distribution

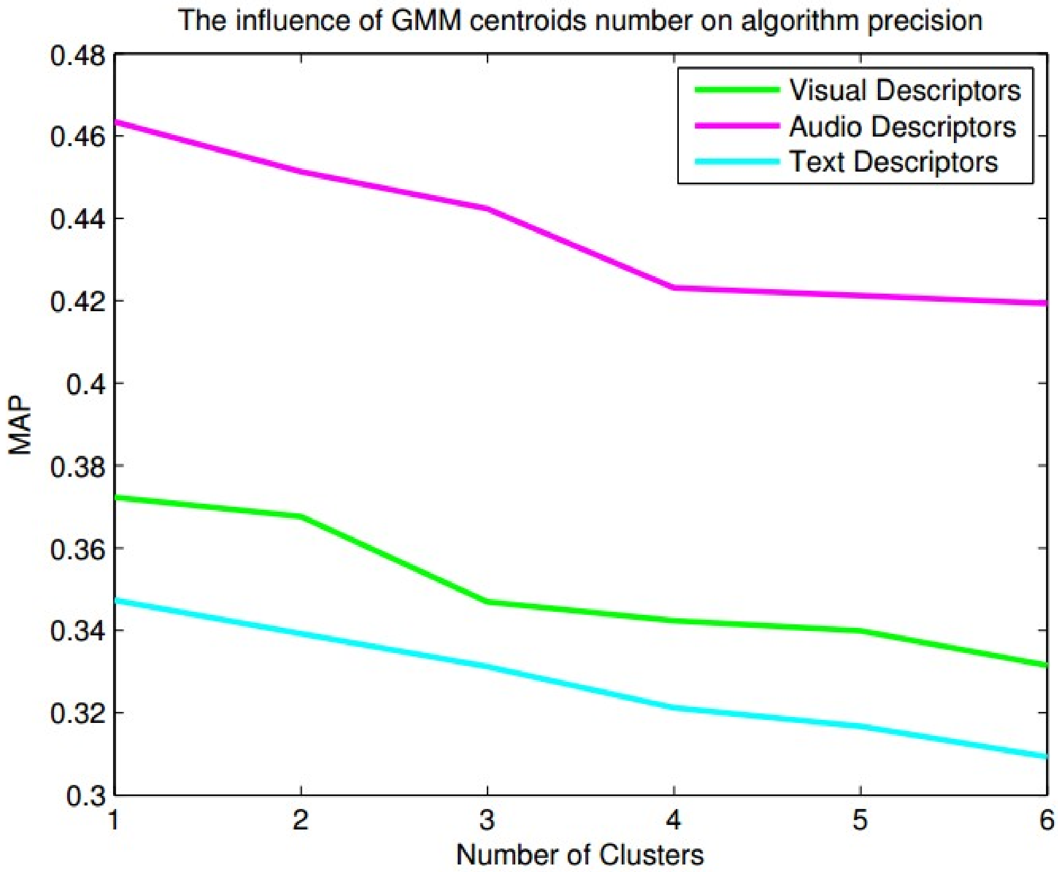
3. select term corpus: retaining for each genre class m terms (e.g. m = 150 for ASR) with the highest $\chi 2$ values that occur more frequently than in complement classes

4. for each document the **TF-IDF** values are computed.

Results

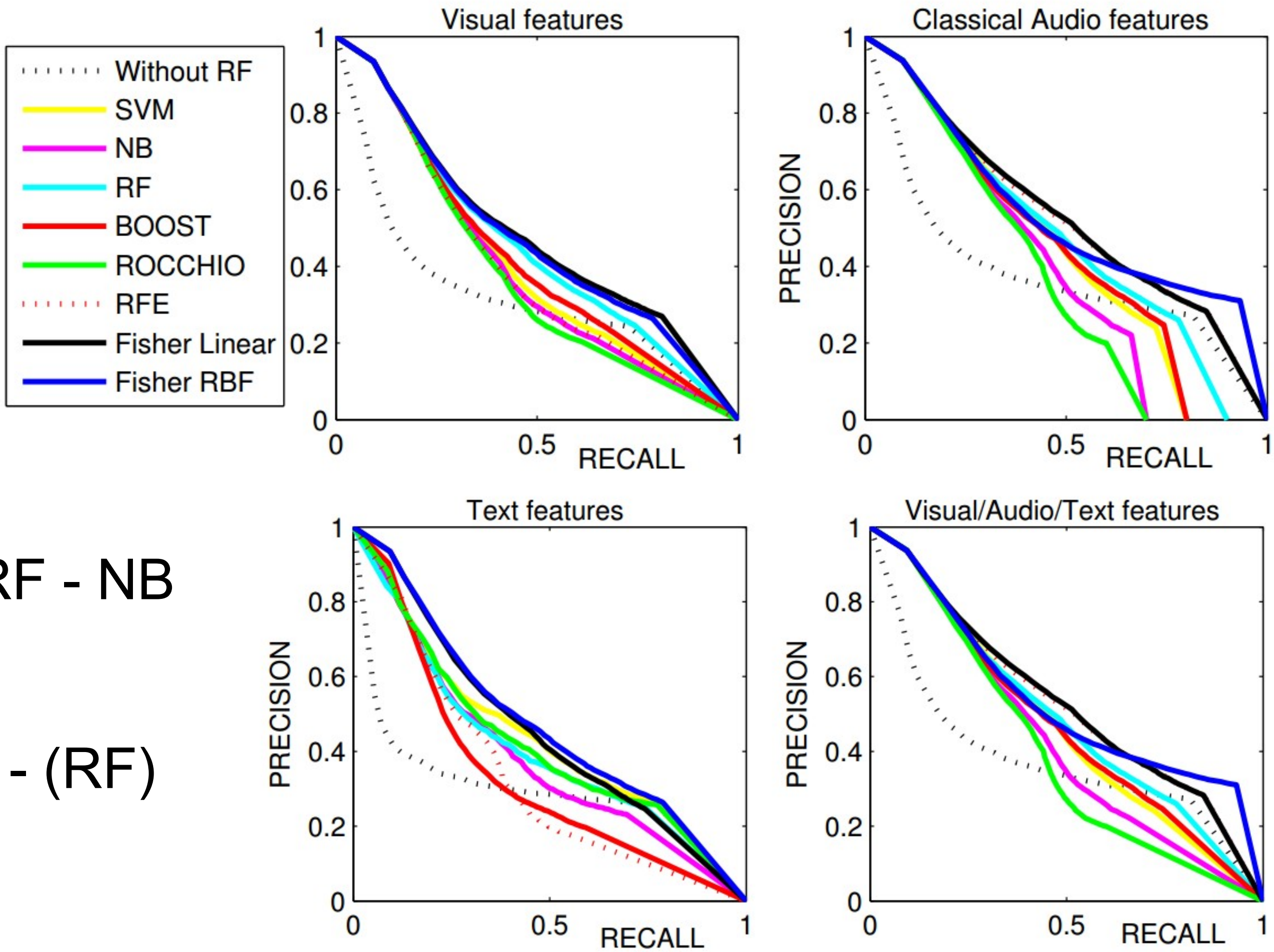
Optimizing Fisher Kernel

- number of GMM Centroids



Results

Comparison to State-of-the-Art algorithms



- Rocchio
- Nearest Neighbor RF - NB
- Boost RF
- SVM RF
- Random Forest RF - (RF)
- Relevance Feature Estimation - (RFE)

Results

Comparison to State-of-the-Art algorithms

Feature	Without RF	Rocchio	NB	Boost	SVM	Random Forest	RFE	FK Linear	FK RBF
HoG	17.18	25.57	24.18	26.72	26.49	26.89	27.50	29.46	**29.59**
Structural	14.82	21.96	23.73	23.63	24.62	24.69	23.91	**26.28**	23.96
MPEG 7	25.97	30.88	34.09	32.55	32.90	36.85	31.93	40.50	**40.80**
All Visual	26.18	32.98	34.25	35.99	36.08	**42.28**	32.43	41.33	42.23
Standard Audio	29.26	32.71	34.88	32.88	38.58	40.46	44.32	44.80	**46.34**
Block Based Audio	21.23	35.39	35.22	39.87	31.46	33.41	31.93	**43.96**	43.69
Text	20.40	32.55	26.91	26.93	34.70	34.70	25.82	34.84	**35.14**
All features	30.29	37.91	39.88	38.88	40.93	45.31	44.93	46.43	**46.80**

(MAP values)

Results

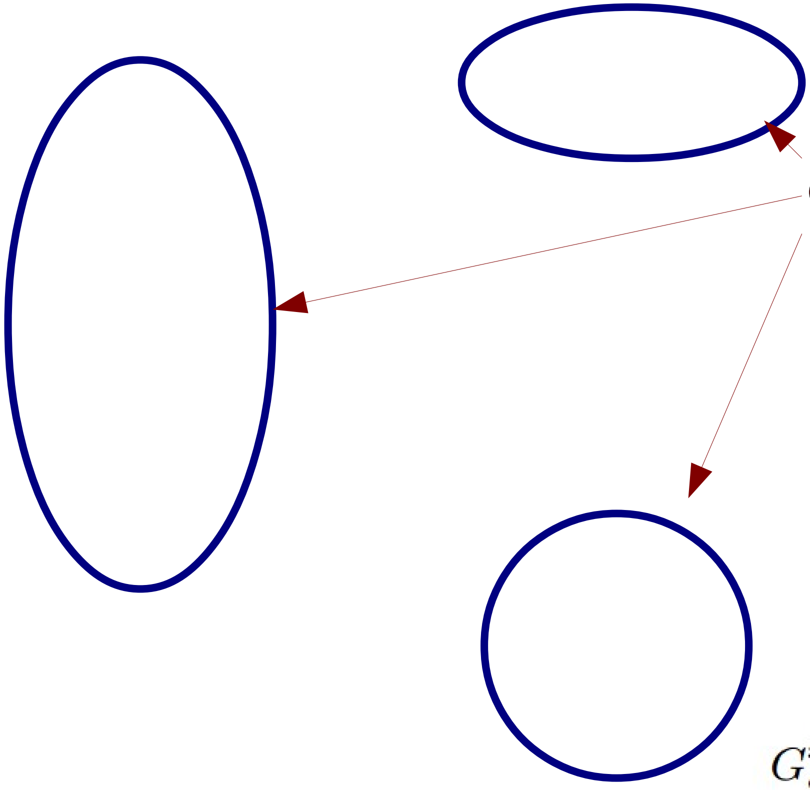
Fisher Kernel representation on *all* data or only *relevant* data

Feature	FK on all data	FK on relevant data
Visual features	34.02%	**38.23%**
Standard audio features	38.25%	**46.34%**
Text features	32.37%	**35.14%**

(MAP values)

Fisher Kernel for Temporal Variation

- represents a signal as the gradient of the probability density function that is a learned generative model of that signal



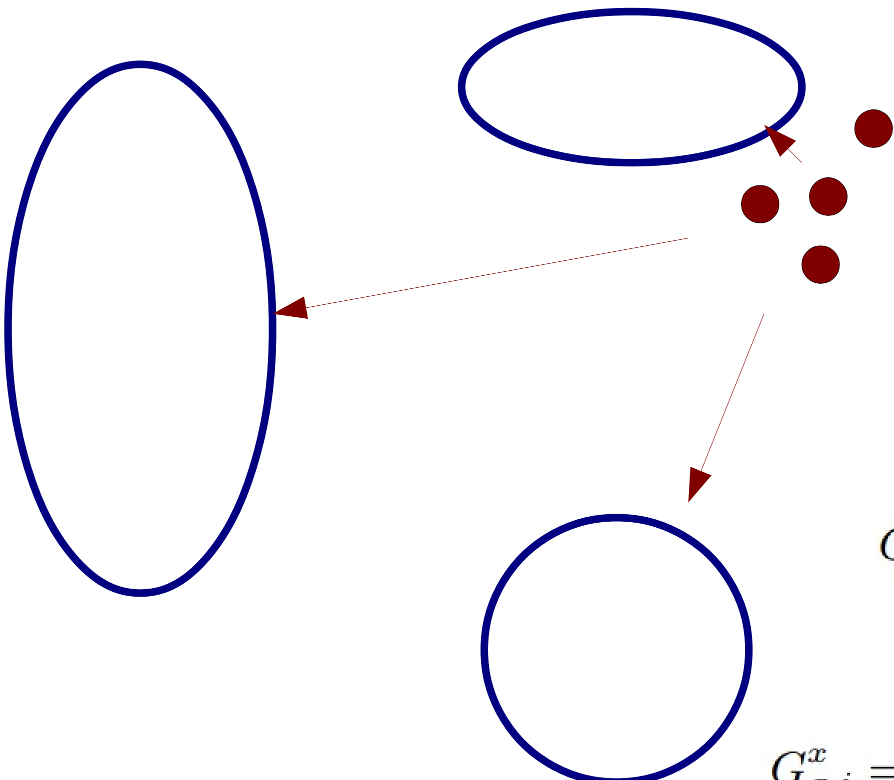
$$G^{x}_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i}$$

$$G^{x}_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]$$

[Jaakkola and Haussler 1999, Perronnin and Dance 2007]

Fisher Kernel for Temporal Variation

- represents a signal as the gradient of the probability density function that is a learned generative model of that signal



$$G^{x}_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i}$$

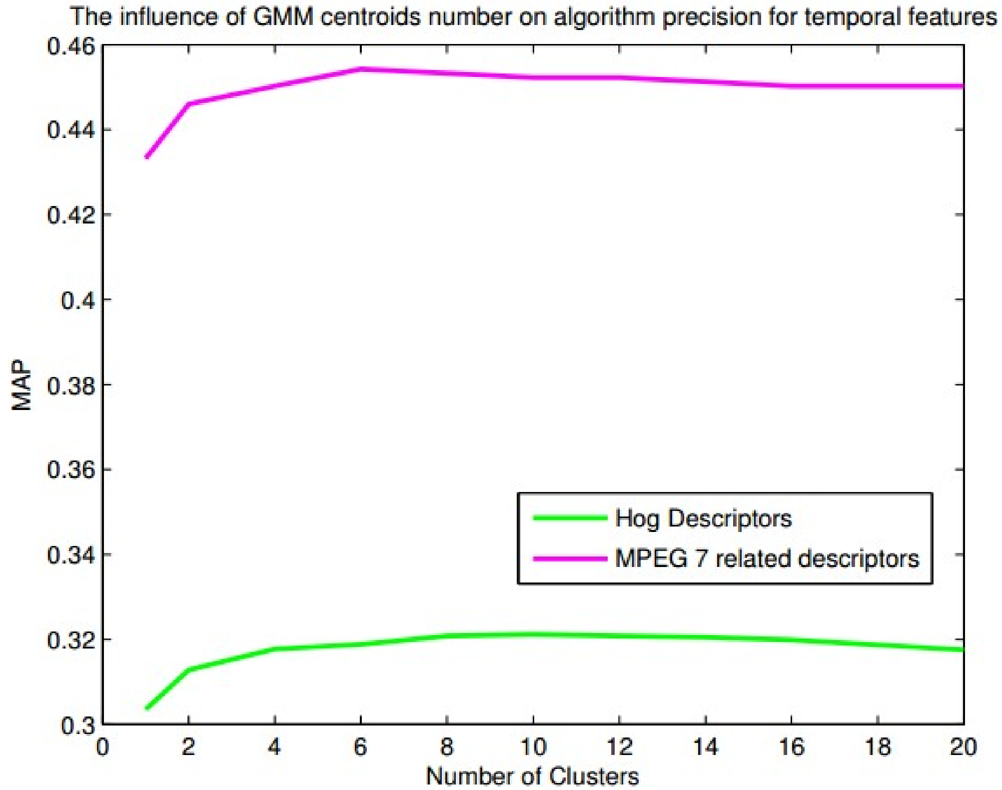
$$G^{x}_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]$$

[Jaakkola and Haussler 1999, Perronnin and Dance 2007]

Results

Frame Agreggation with Fisher Kernel

Number of GMM Centroids



Results

Frame Agreggation with Fisher Kernel

Feature	FKRF RBF (T=1)	Frame Aggregation with Fisher Kernel
HoG	29.59%	32.87%
MPEG-7	40.80%	45.43%

(MAP values)

almost 5% improvement for MPEG-7

Conclusions

- Relevance feedback is a powerful tool for improving content based video retrieval systems

- Fisher Kernel RF outperforms state-of-the-art.

- Modelling the temporal variation in video using the Fisher Kernel yields 5% improvement for MPEG-7 features