Análise Forense de Documentos Digitais

Prof. Dr. Anderson Rocha

anderson.rocha@ic.unicamp.br

http://www.ic.unicamp.br/~rocha

Reasoning for Complex Data (RECOD) Lab.
Institute of Computing, Unicamp

Av. Albert Einstein, 1251 – Cidade Universitária
CEP 13083-970 • Campinas/SP – Brasil

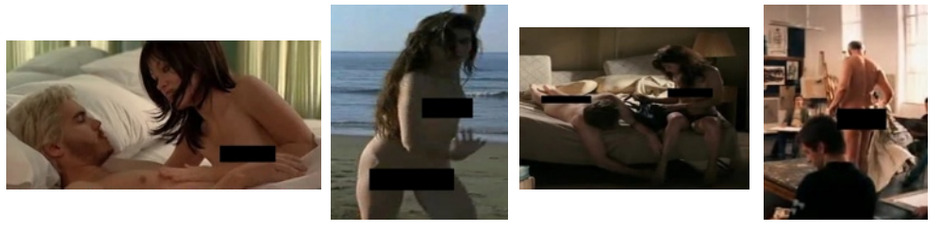
Outline

Outline

- The pornography vs. nude detection problem
- Related work
- Skin-based techniques and how they work
- Visual Words Techniques and how they work
- Working with Videos
- Space-temporal Visual Words
- Extending the Solutions to Violence Detection and General Action Recognition

Introduction

- The slides herein are a composition of many contributions of several authors
  - Michael J. Jones, James M. Rehg, Li Fei-Fei, Ana Paula Lopes, Sandra E. F. de Avila, Anderson N. A. Peixoto, Rodrigo S. Oliveira, Marcelo de M. Coelho, Arnaldo Araújo, Fillipe Dias, Guillermo Cámara Chávez, Eduardo Valle and many others.
- I deeply thank all of them for making their resources available



Introduction

Introduction

- The importance of pornography detection is attested by the large literature on the subject.
- Moreover, web filter is essential to avoid offensive content, such as adult or pornographic material.
- There are some industry software that block web sites with offensive content (e.g. CyberPatrol1, NetNanny2, K9 Web Protection3)

Introduction

- Also, there are software products which scan a computer for pornographic content (e.g. SurfRecon4, Porn Detection Stick5, PornSeer Pro6).
- The latter pornography-detection software is readily available for evaluation purposes.
- NuDetective (Brazilian Federal Police)

Introduction

State-of-the-Art

State-of-the-art

- Most work regarding the detection of pornographic material has been done for the image domain
- The vast majority of those works is based on the detection of human skin. For example, [Fleck et al., 1996; Forsyth and Fleck, 1996, 1997, 1999] proposed to detect skin regions in an image and match them with human bodies by applying geometric grouping rules.

State-of-the-art

- [Jones and Rehg, 2002] focused on the detection of human skin by constructing RGB color histograms from a large database of skin and non-skin pixels, which allows to estimate the "skin probability" of a pixel based on its color.
- [Rowley et al., 2006] used Jones and Rehg' skin color histograms in a system installed in Google's Safe Search.
- [Zuo et al., 2010] proposed a patch-based skin color detection that verifies whether all the pixels in a small patch correspond to human skin tone.

State-of-the-art

- Few methods have explored other possibilities.
- Bag-of-Words models [Sivic and Zisserman, 2003] have been employed for many complex visual classification tasks, including pornography detection in images and videos
- [Deselaers et al., 2008] first proposed a BoW model to filter pornographic images, which greatly improved the efficiency of the identification of pornographic images.

State-of-the-art

- Lopes et al. developed a BoW-based approach, which used the Hue-SIFT color descriptor, to classify images [Lopes et al., 2009b] and videos [Lopes et al., 2009a] of pornography.
- [Ulges and Stahl, 2011] introduced a color-enhanced visual word features in YUV color space to classify child pornography.

State-of-the-art

- [Steel, 2012] proposed a pornographic images recognition method based on visual words, by using mask-SIFT in a cascading classification system.
- Those previous works have explored only bags of static features.

State-of-the-art

- Very few works have explored spatiotemporal features or other motion information for detection of pornography
- [Tong et al., 2005] proposed a method to estimate the period of a signal to classify periodic motion patterns.
- [Endeshaw et al., 2008] developed a fast method for detection of indecent video content using repetitive movement analysis.

State-of-the-art

- [Jansohn et al., 2009] introduced a framework that combines keyframe-based methods with a statistical analysis of MPEG-4 motion vectors.
- [Valle et al., 2012] compared the use of several features, including spatiotemporal local descriptors, in the pornography detection.

Parallel

Parallel

- There is a vast amount of multimedia data nowadays
- Filtering improper multimedia material by its visual content is needed

Parallel

- Skin detectors
  - Precise skin detection is not a trivial task
  - Generic geometrical model vs. Various body poses
- BoVF representation...
  - ...has great success in object recognition tasks and...
  - ...is robust to several variations and occlusion

Statistical Color Models with Applications in Skin Detection

Michael J. Jones and James M. Rehg
Cambridge Research Laboratory
Compaq Computer Corporation
One Cambridge Center

Springer Intl. Journal of Computer Vision, Vol. 46, No. 1, 2002

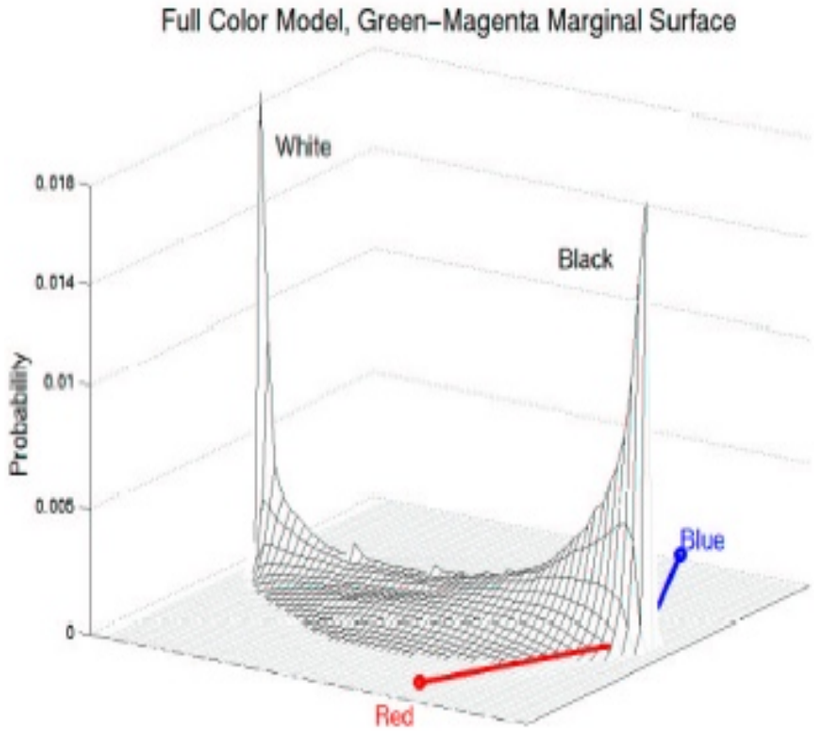
Introduction

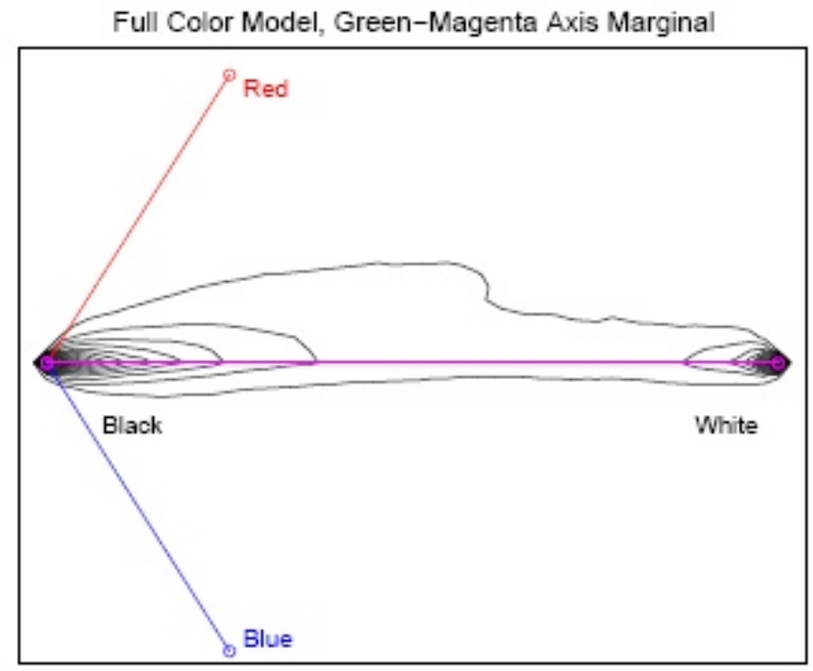
- The paper describes the construction of skin and non-skin statistical color models;
- Uses 1 billion training pixels from Web images;
- Shows a skin classifier with accuracy of 88%;
- System used to detected naked people on Web images;

Histogram Color Models

- Images are organized in two sets:
  - Generic Training Set;
    - Used to compute a general histogram density;
  - Classifier Training Set;
    - Used to build the skin and non-skin models;
    - Manually separated into subsets containing skin and not containing skin;
    - Skin pixels are manually labeled;

General Color Model

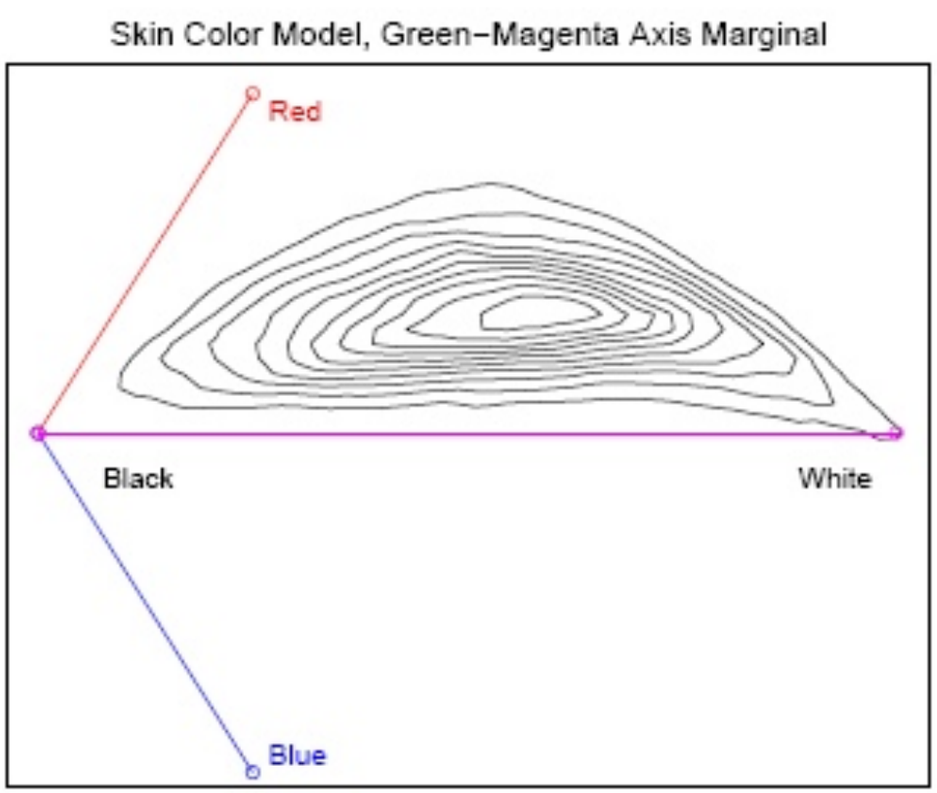


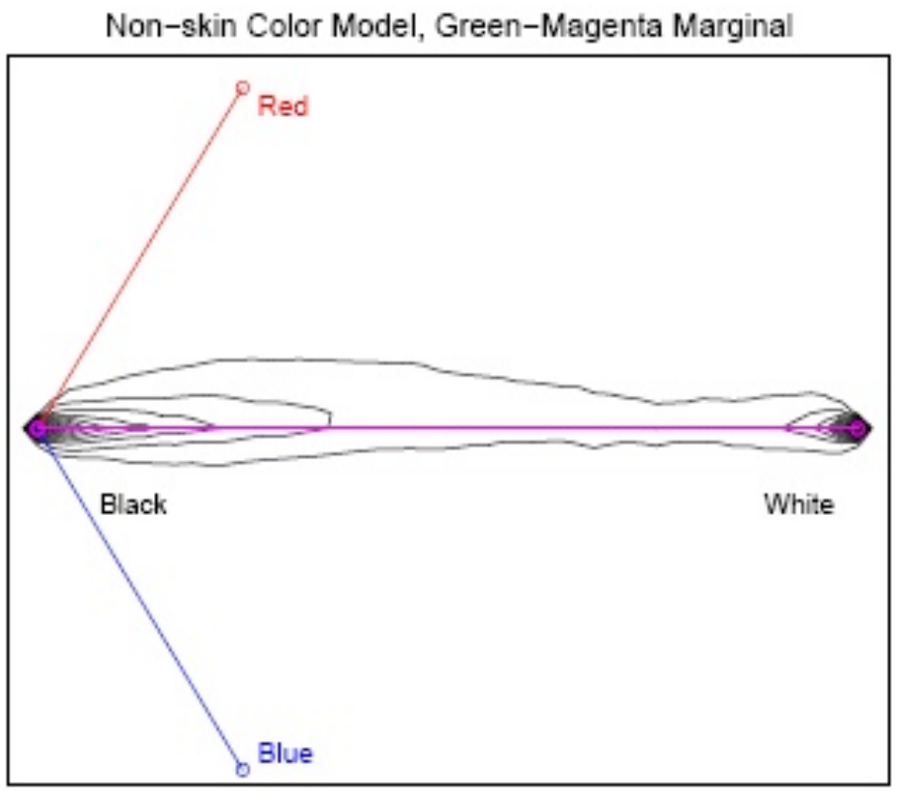


Observations:

1. Most colors fall on or near the gray line;
2. Black and white are by far the most frequent colors, with white occuring slightly more frequently;
3. There is a marked skew in the distribution toward the red corner of the color cube.

Skin and Non-skin Color Models





Observations:

1. Non-skin model is the general model without skin pixels (10% of pixels);
2. There is a significant degree of separation between the skin and non-skin models;

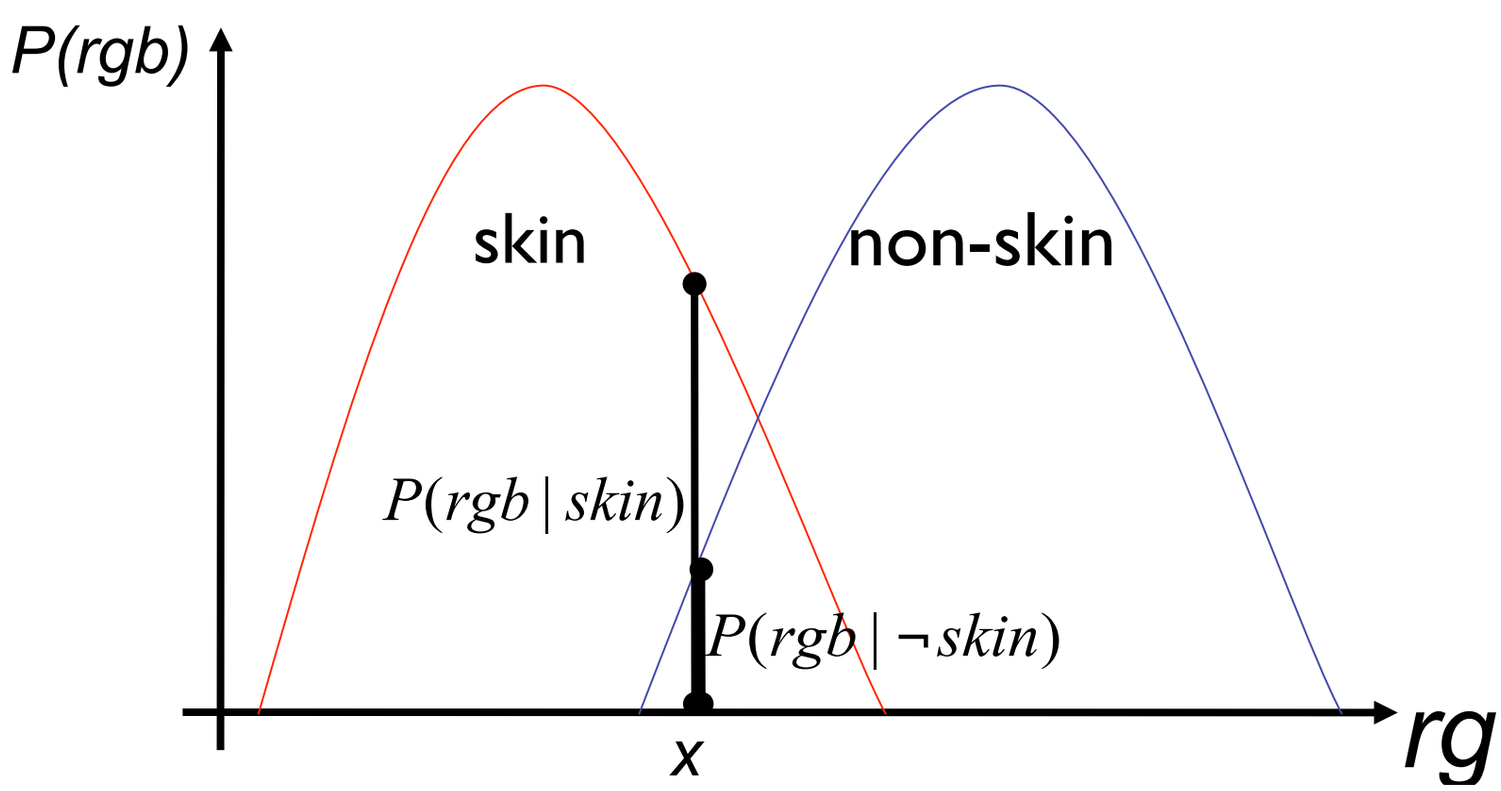
Skin Detection Using Color Models

- Given skin and non-skin histogram models we can construct a skin pixel classifier;
- Such a classifier is extremely useful in two contexts:
  - Detection and recognition of faces and figures;
  - Image indexing and retrieval;

Skin Detection Using Color Models

- a skin pixel classifier is derived through the standard likelihood ratio approach:

$$\frac{P(rgb \mid skin)}{P(rgb \mid \neg skin)} \geq \Theta$$



RECOD

Histogram-based Skin Classifier

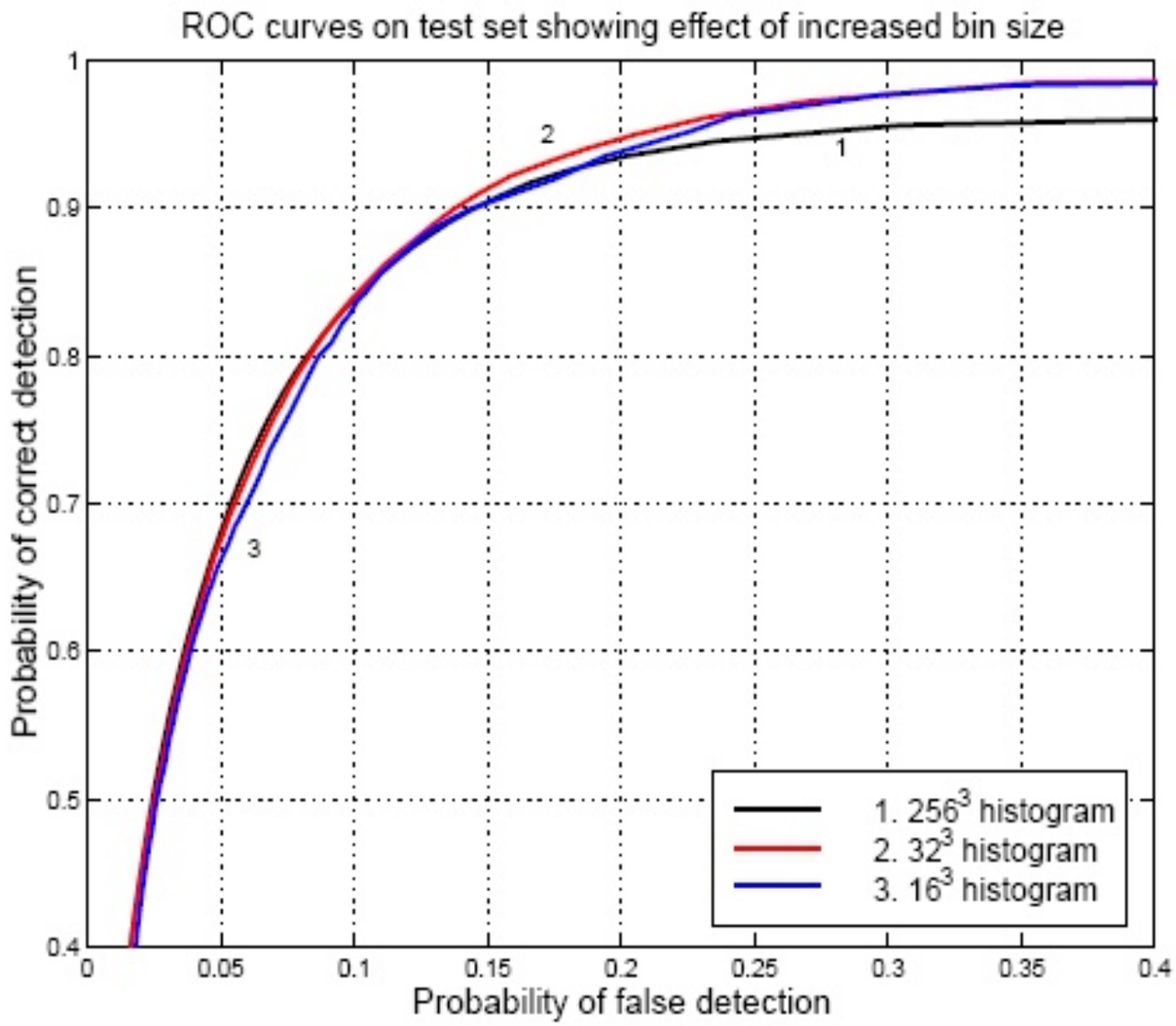
Histogram-based Skin Classifier

- Qualitative observations: $\theta = 0.4$;
  - The classifier does a good job of detecting skin in most of these examples;
  - In particular, the skin labels form dense sets whose shape often resembles that of the true skin pixels;
  - The detector tends to fail on highly saturated or shadowed skin;

Histogram-based Skin Classifier

- More qualitative observations:
  - The example photos also show the performance of the detector on non-skin pixels.
  - In photos such as the house (lower right) or flowers (upper right) the false detections are sparse and scattered.
  - More problematic are images with wood or copper-colored metal such as the kitchen scene (upper left) or railroad tracks (lower left).
  - These photos contain colors which often occur in the skin model and are difficult to discriminate reliably.
  - This results in fairly dense sets of false positives.

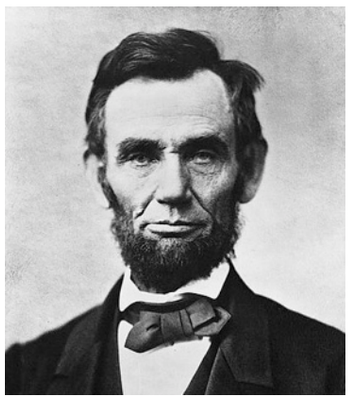
Histogram-based Skin Classifier



fraction of pixels labeled as skin that were classified correctly.

fraction of non-skin pixels which are mistakenly classified as skin.

Receiver Operating Characteristic Curves (ROC curves)

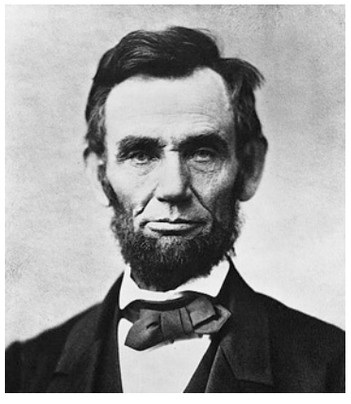


Original

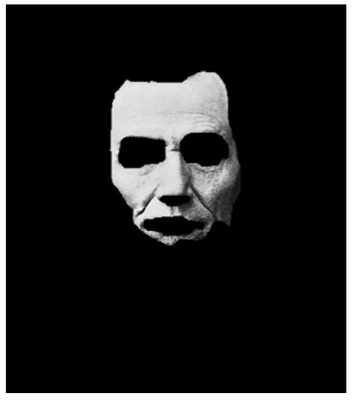
Θ = 7

Θ = 9

Receiver Operating Characteristic Curves (ROC curves)

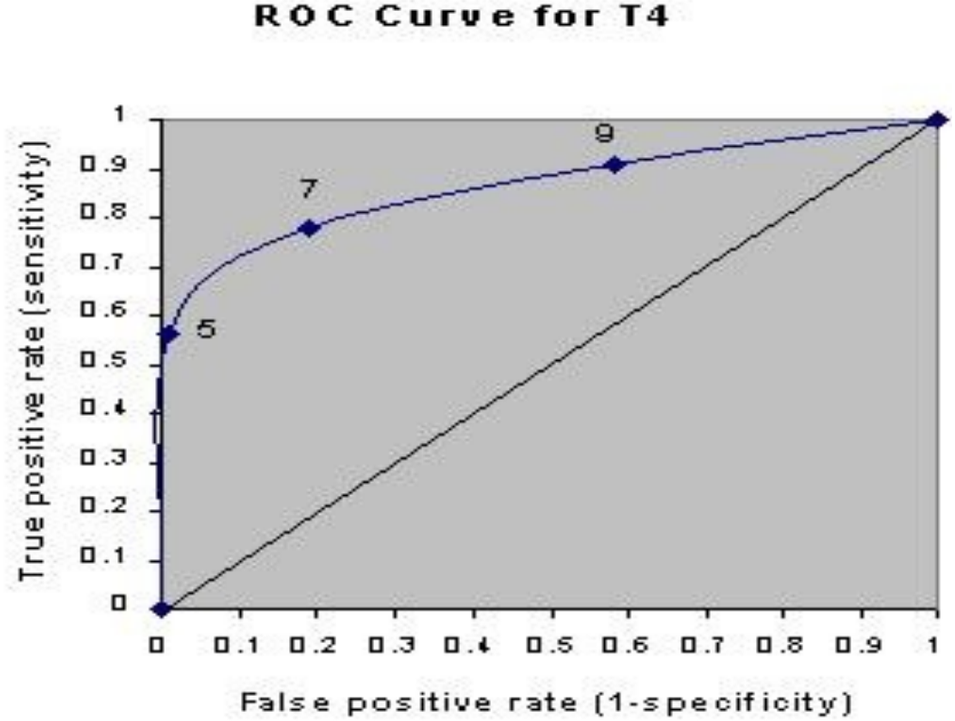


Original



Θ = 7

Θ = 9



RECOD

Histogram-based Skin Classifier

- More quantitative observations:
  - The performance of the skin classifier is surprisingly good considering the unconstrained nature of Web images;
  - The best classifier (size 32) can detect roughly 80% of skin pixels with a false positive rate of 8.5%, or 90% correct detections with 14.2% false positives;
  - Its equal error rate is 88%.

Comparison to GMM Classifier

- One advantage of mixture models is that they can be made to generalize well on small amounts of training data;

- One possible benefit of a large dataset is the ability to use density models such as histograms which are computationally simpler to learn and apply.

- Mixture models were trained for the dataset and compared their classification performance to the histogram models.

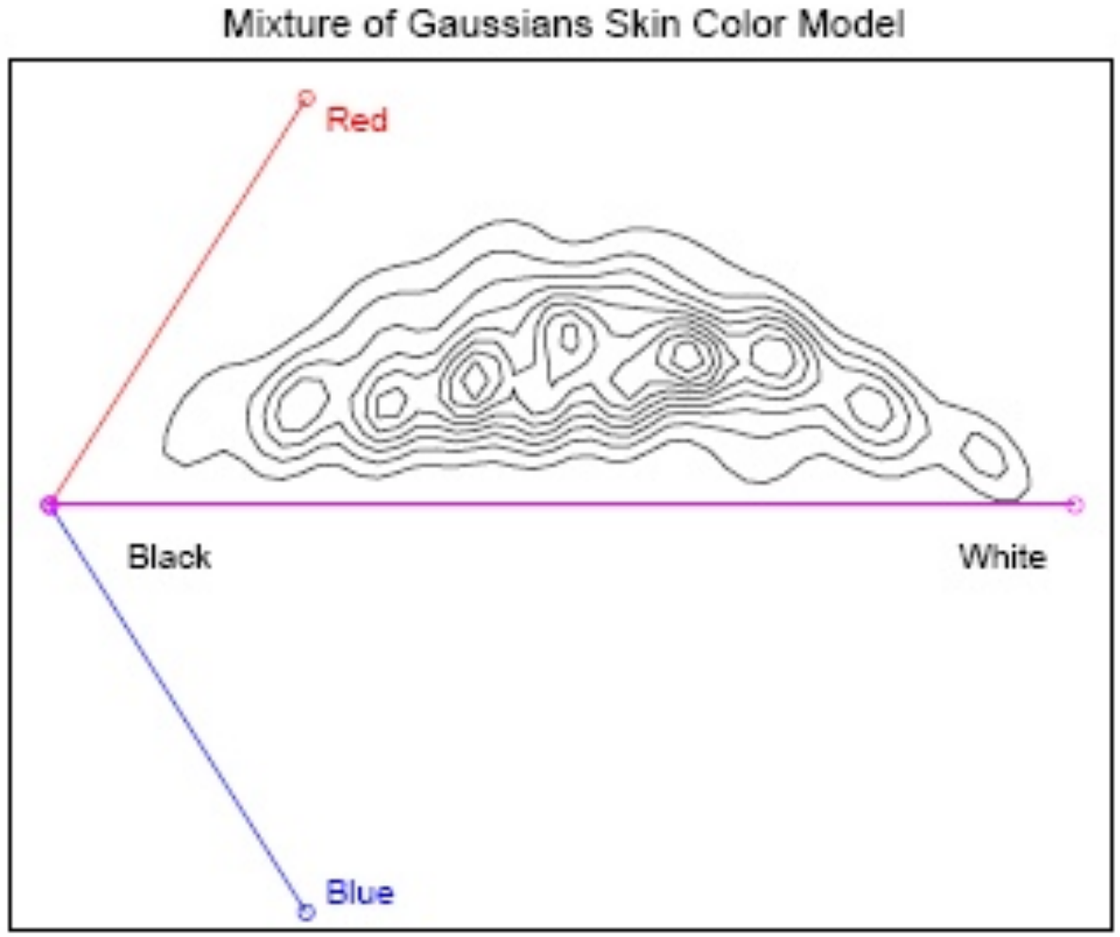
Comparison to GMM Classifier

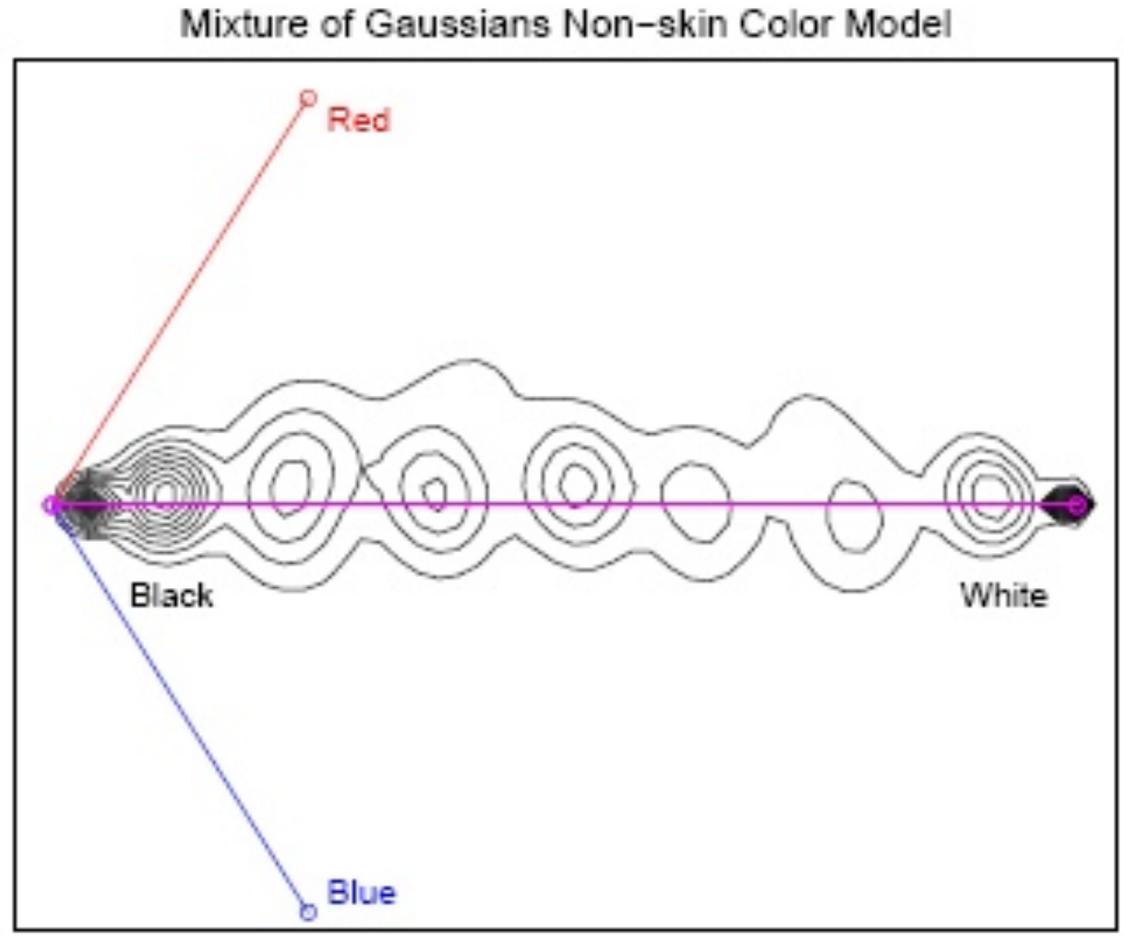
- two separate mixture models were trained for the skin and non-skin classes;

- It was used 16 Gaussians in each model;

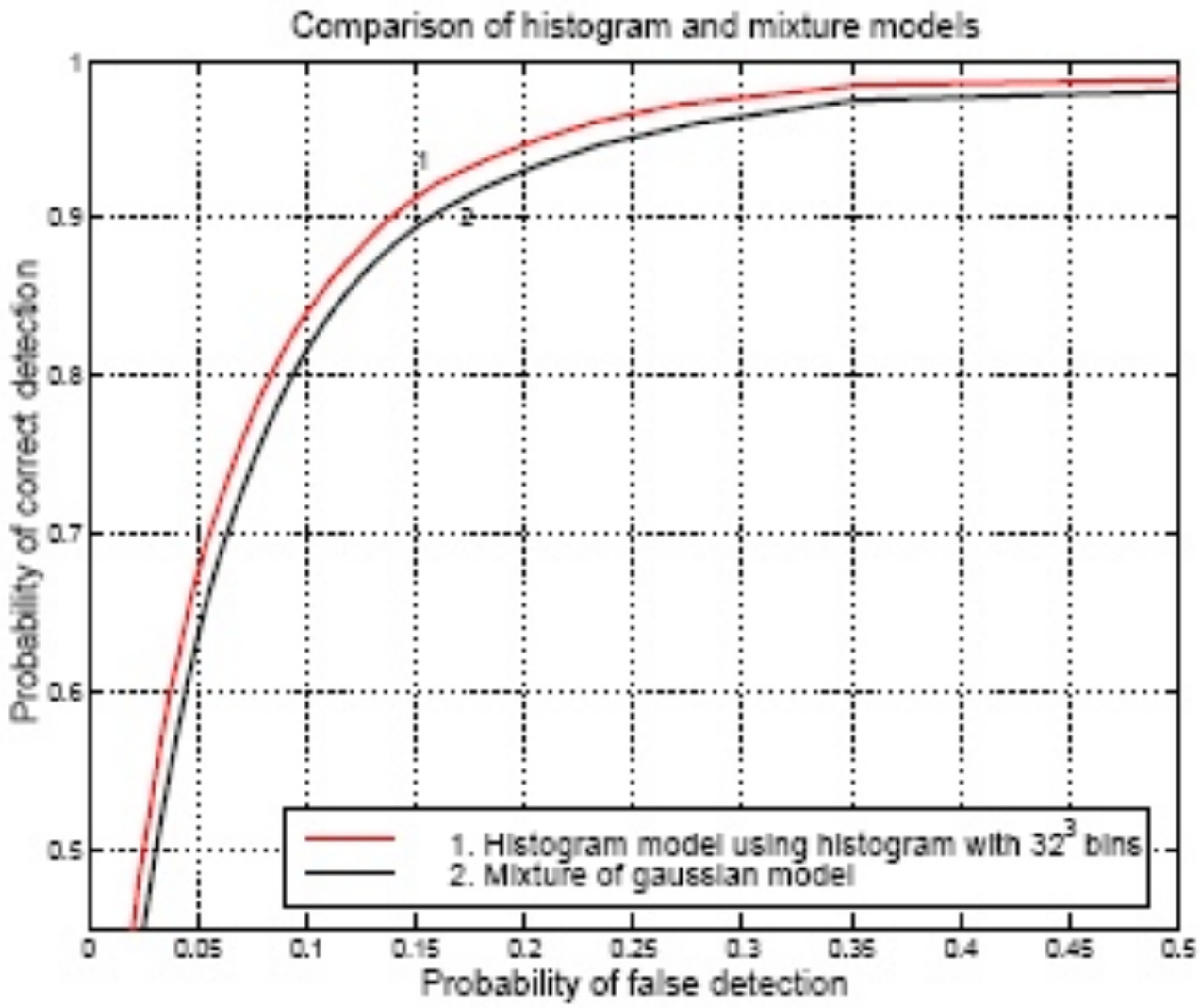
- The models were trained using a parallel implementation of the standard EM algorithm.

Comparison to GMM Classifier





Comparison to GMM Classifier



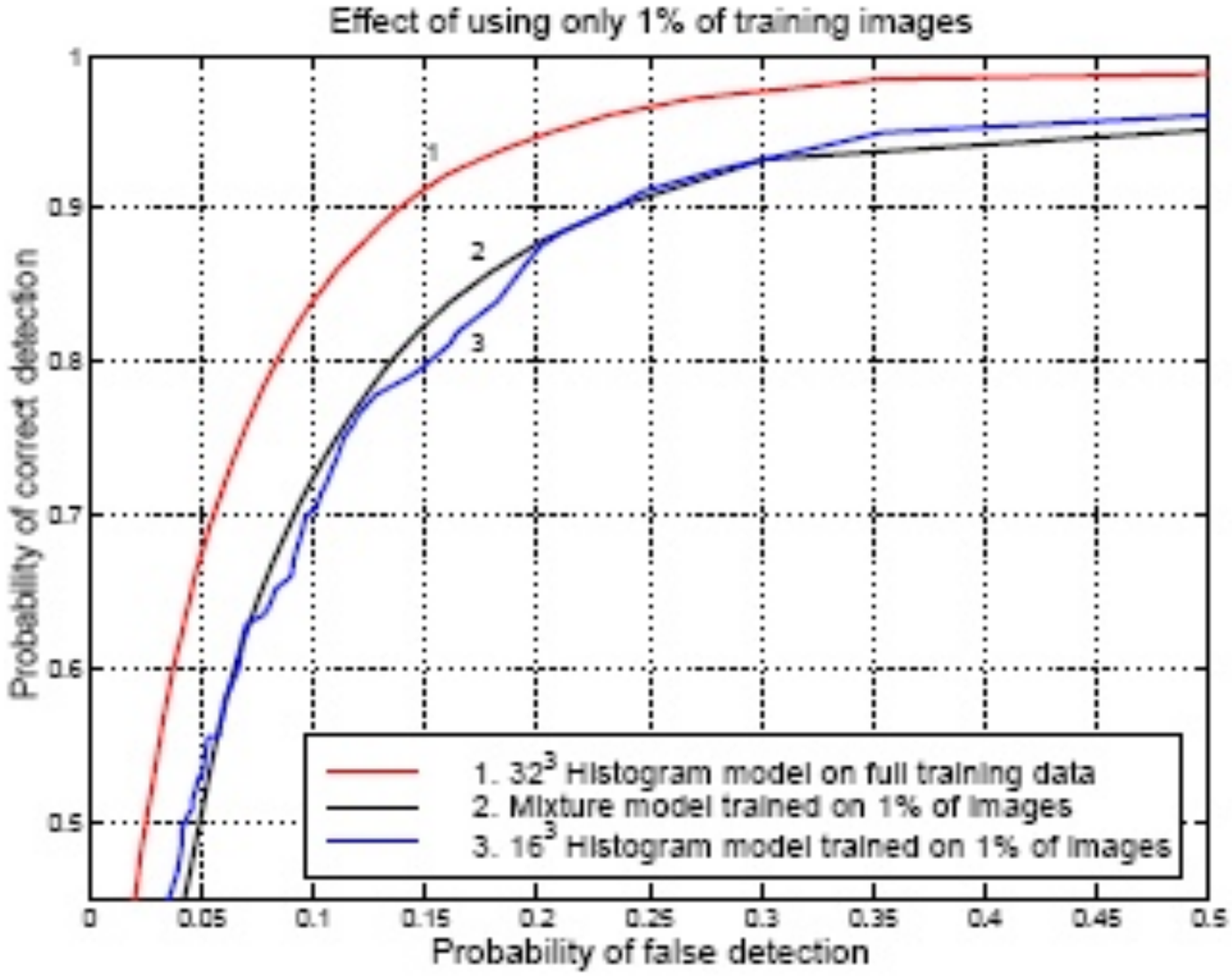
Comparison to GMM Classifier

- The mixture of Gaussian model is significantly more expensive to train than the histogram models;
- It took 24 hours to train both skin and non-skin mixture models using 10 Alpha workstations in parallel. In contrast, the histogram models could be constructed in a matter of minutes on a single workstation;
- The mixture model is also slower to use during classification since all of the Gaussians must be evaluated in computing the probability of a single color value;

Comparison to GMM Classifier

- In contrast, use of the histogram model results in a fast classifier since only two table lookups are required to compute the probability of skin.
- From the standpoint of storage space, however, the mixture model is a much more compact representation of the data.

Comparison to GMM Classifier



Applications

- Person Detector;
- Adult Image Detector;
- Incorporating Text Features into Adult Image Detection.

Applications - Person Detector

- The goal is to determine whether or not an input image contains one or more people;
- A classifier is trained on a simple feature vector computed from the output of the skin detector;

1. Percentage of pixels detected as skin;
2. Average probability of the skin pixels;
3. Size in pixels of the largest connected component of skin;
4. Number of connected components of skin;
5. Percent of colors with no entries in the skin and non-skin histograms.

Applications - Person Detector

- The goal is to determine whether or not an input image contains one or more people;
- A decision tree classifier is trained on a simple feature vector computed from the output of the skin detector;
  1. Percentage of pixels detected as skin;
  2. Average probability of the skin pixels;
  3. Size in pixels of the largest connected component of skin;
  4. Number of connected components of skin;
  5. Percent of colors with no entries in the skin and non-skin histograms.
- Results:
  - 83.2% (correctly classified person images)
  - 71.3% (correctly classified non-person images)
  - 77.5% (correctly classified images)

Applications - Adult Image Detector

- There is a strong correlation between images with large patches of skin and adult or pornographic images;

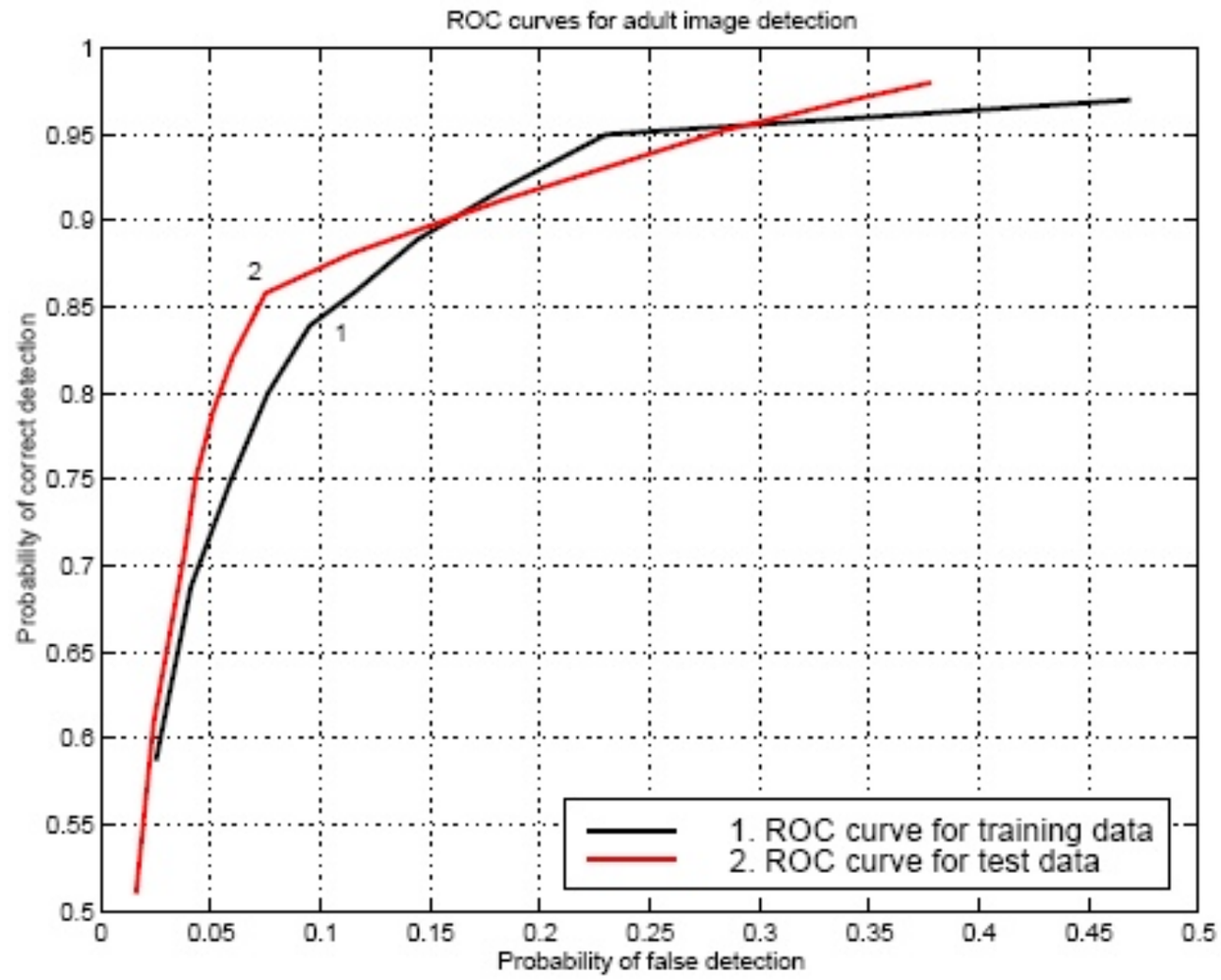
- All of these services currently operate by maintaining lists of objectionable URL's and newsgroups and require constant manual updating;

- An image-based scheme has the potential advantage of applying equally to all images without the need for updating.

Applications - Adult Image Detector

- The same approach as with person detection is used;
- A neural network classifier is used;
- The feature vector includes two additional elements corresponding to the height and width of the image;
- These two were added based on informal observations that adult images are often sized to frame a standing or reclining figure.

Applications - Adult Image Detector



Results:

- 85.8% correctly detected adult images
- 7.5% mistakenly classified non-adult images

Applications - Adult Image Detector

(a) Examples of images correctly classified by our detector. Both images were classified as adult images.

(b) Example of an image misclassified as adult by our detector.

Applications - Adult Image Detector

- Combining the adult image detector just described with a text-based classifier (obtained from AltaVista) which uses the text occurring on a Web page to determine if it is pornographic;
- The text-based detector uses the standard approach of matching words on the page against a manually-specified list of objectional words, and classifying based on the match frequencies;
- The text-based classifier on its own achieves 84.9% correct detections of adult images with 1.1% false positives.

Applications - Adult Image Detector

	% correctly detected adult images	*% false alarms*
Color-based Detector	85.8%	7.5%
Text-based Detector	84.9%	1.1%
Combined Detector	93.9%	8.0%

- The color-based detector is combined with the text-based detector by using an "OR" of the two classifiers, i.e. an image is labeled adult if either classifier labels it adult.

Comments

- Color distributions for skin and non-skin pixel classes learned from web images can be used as an accurate pixel-wise skin detector;
- The key is the use of a very large labeled dataset to capture the effects of the unconstrained imaging environment represented by web photos;
- Visualization studies show a surprising degree of separability in the skin and non-skin color distributions;
- They also reveal that the general distribution of color in web images is strongly biased by the presence of skin pixels.

Comments

- One possible advantage of using a large dataset is that simple learning rules may give good performance;
- A pixel-wise skin detector can be used to detect images containing naked people, which tend to produce large connected regions of skin;
- It is shown that a detection rate of 88% can be achieved with a false alarm rate of 11.3%, using a seven element feature vector and a neural network classifier;
- The results suggest that skin color is a very powerful cue for detecting people in unconstrained imagery.

Before

Continuing...

Bag-of-words models (BoW Models)

Li Fei-Fei (Stanford Univ.)

Related Work

- Early “bag of words” models: mostly texture recognition
  - Cula et al. 2001; Leung et al. 2001; Schmid 2001; Varma et al. 2002, 2003; Lazebnik et al. 2003
- Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
  - Hoffman 1999; Blei et al, 2004; Teh et al. 2004
- Object categorization
  - Dorko et al. 2004; Csurka et al. 2003; Sivic et al. 2005; Sudderth et al. 2005;
- Natural scene categorization
  - Fei-Fei et al. 2005

Object

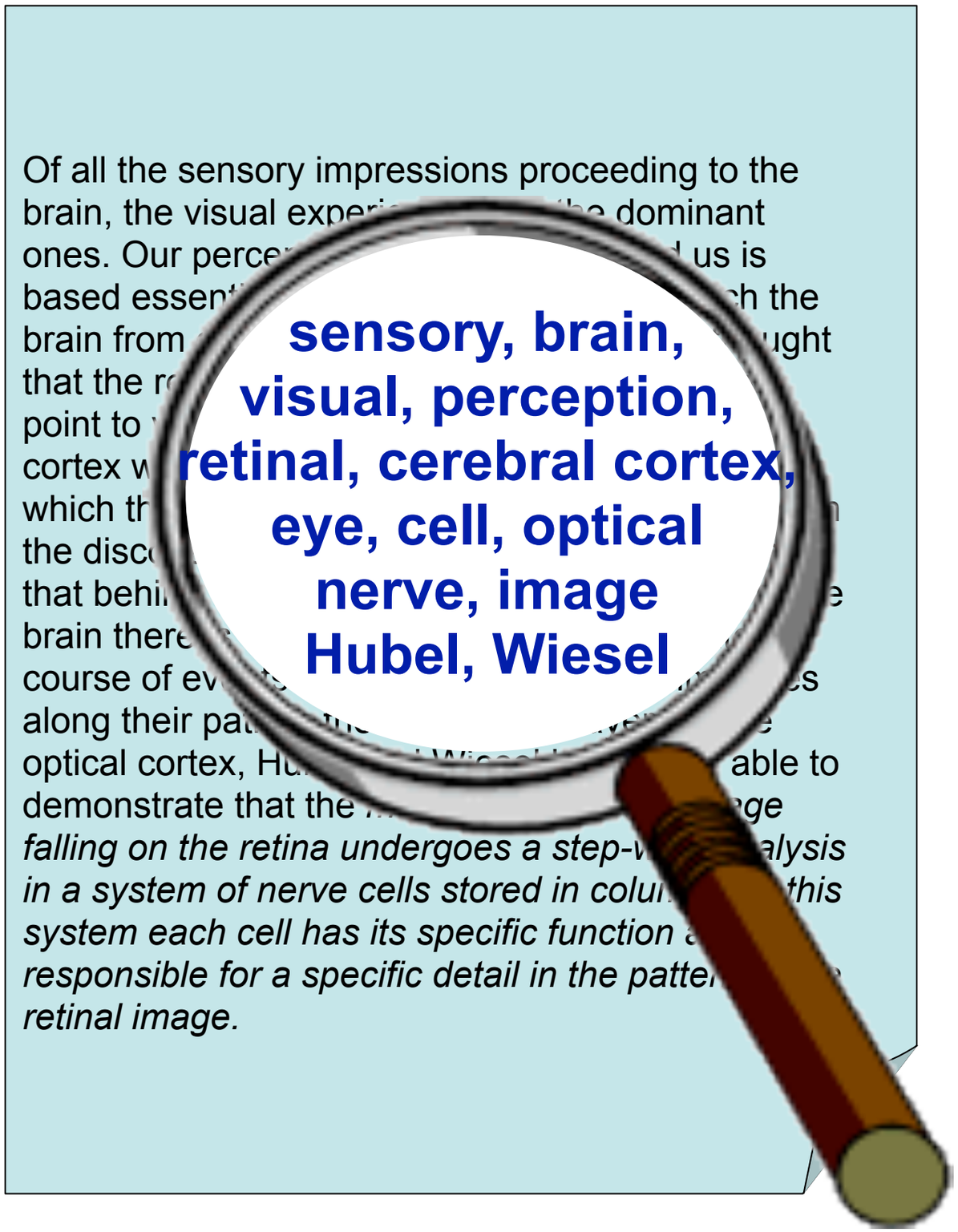
Object

Bag of ‘words’

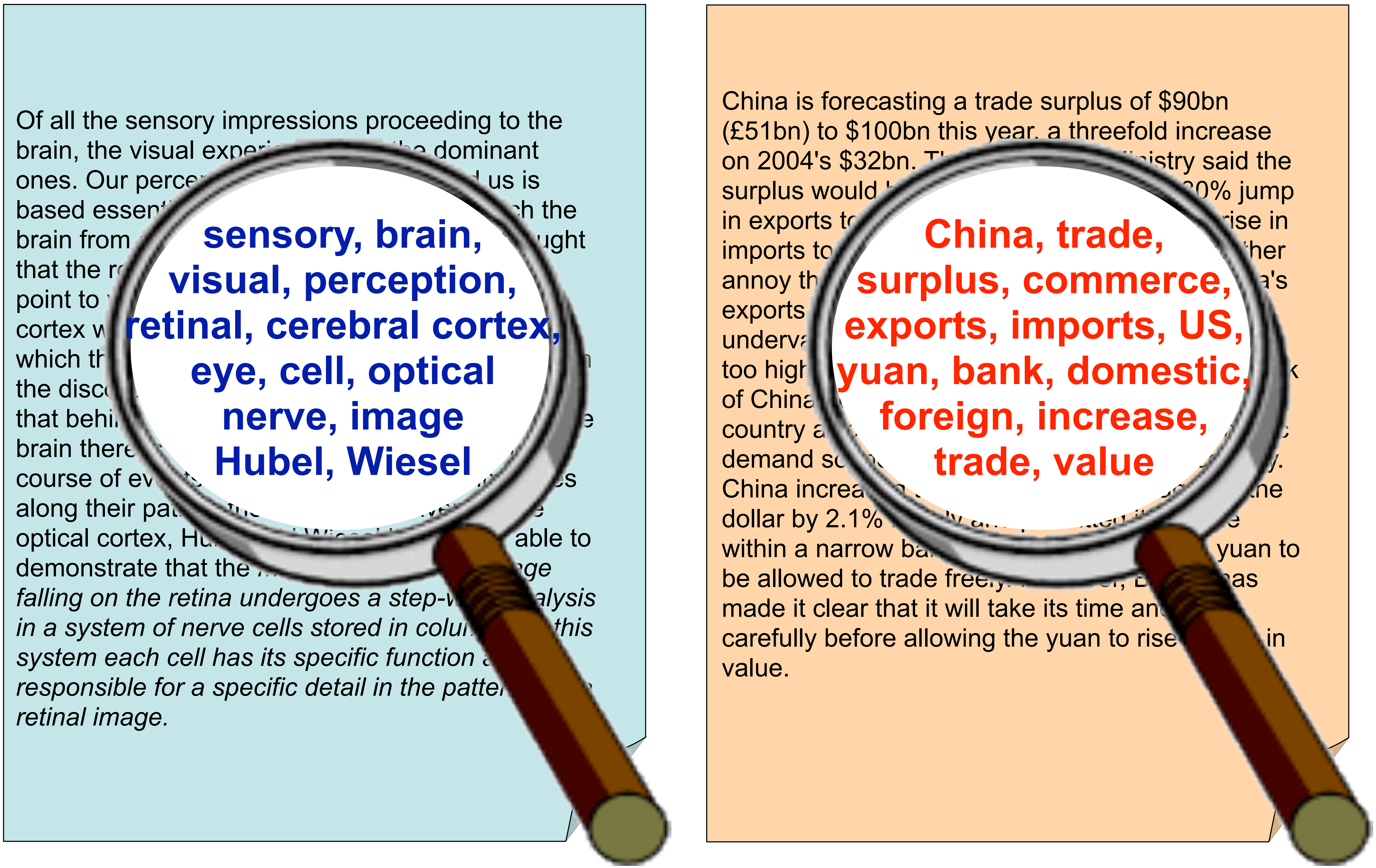
Analogy to documents

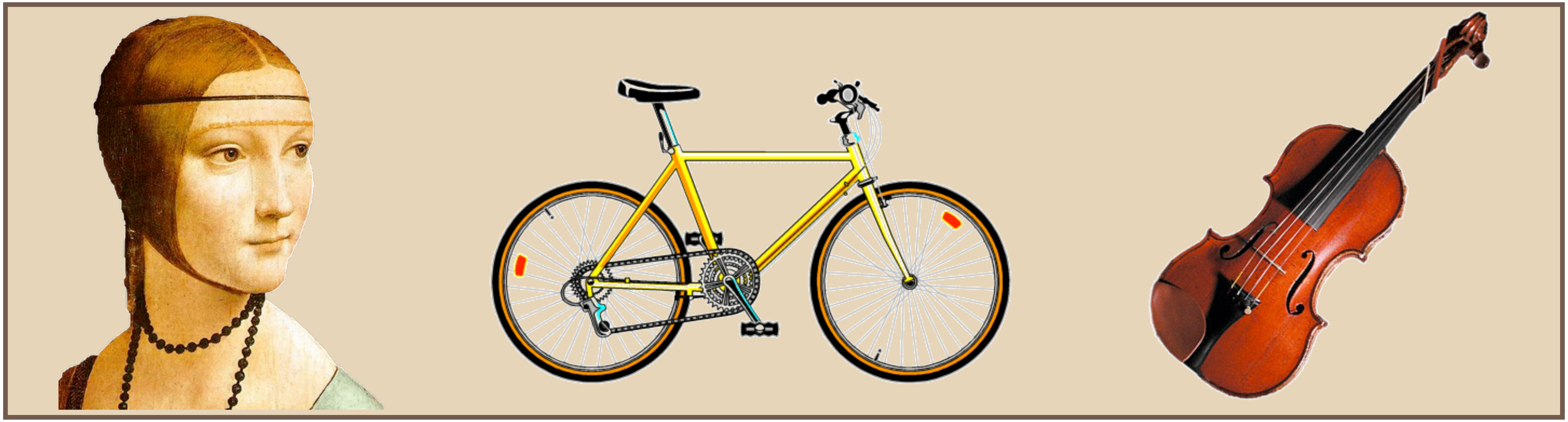
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex was a movie screen, so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the *message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.*

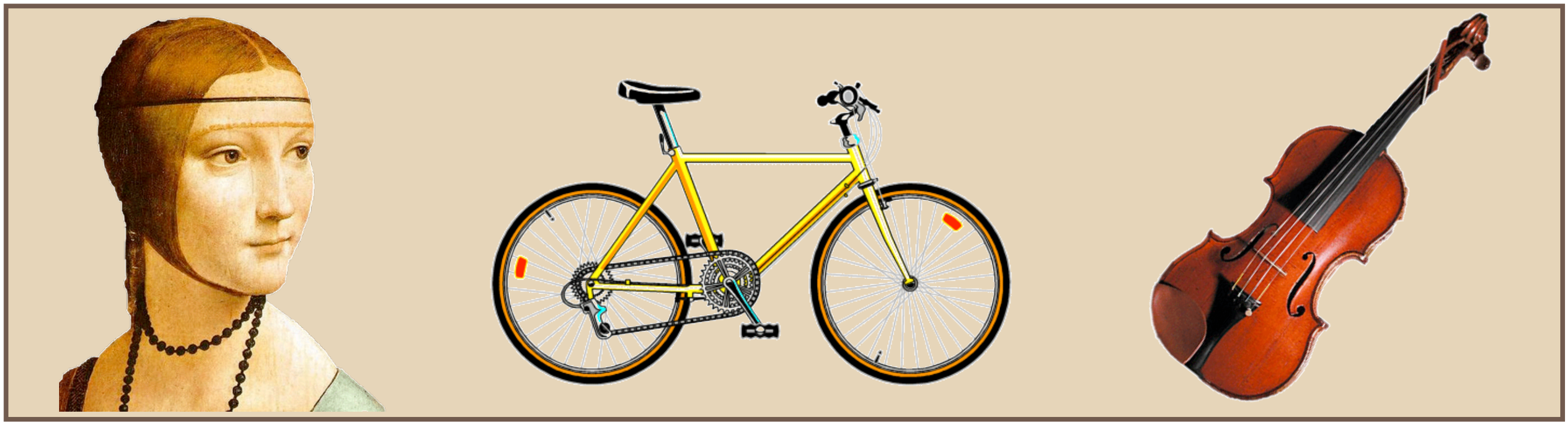
Analogy to documents

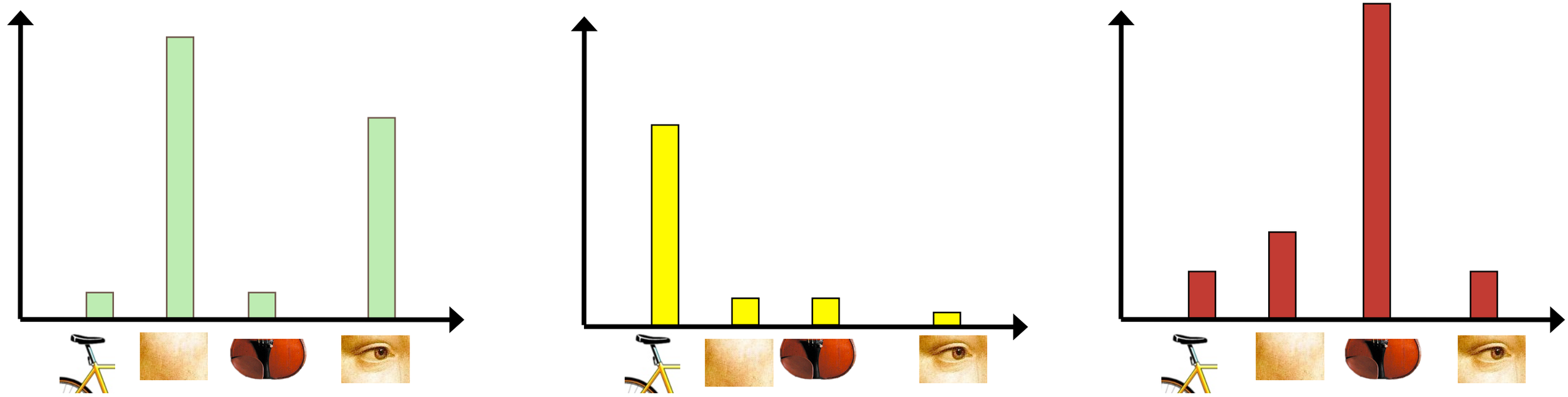


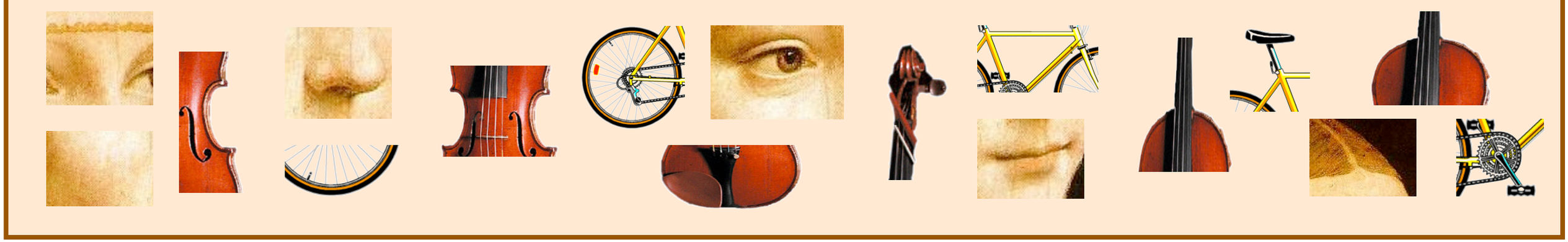
Analogy to documents



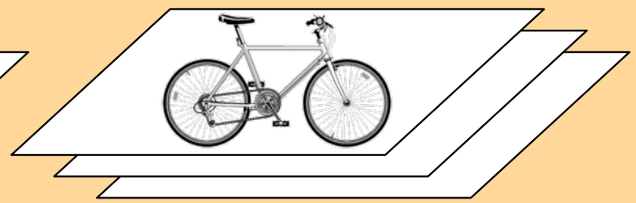




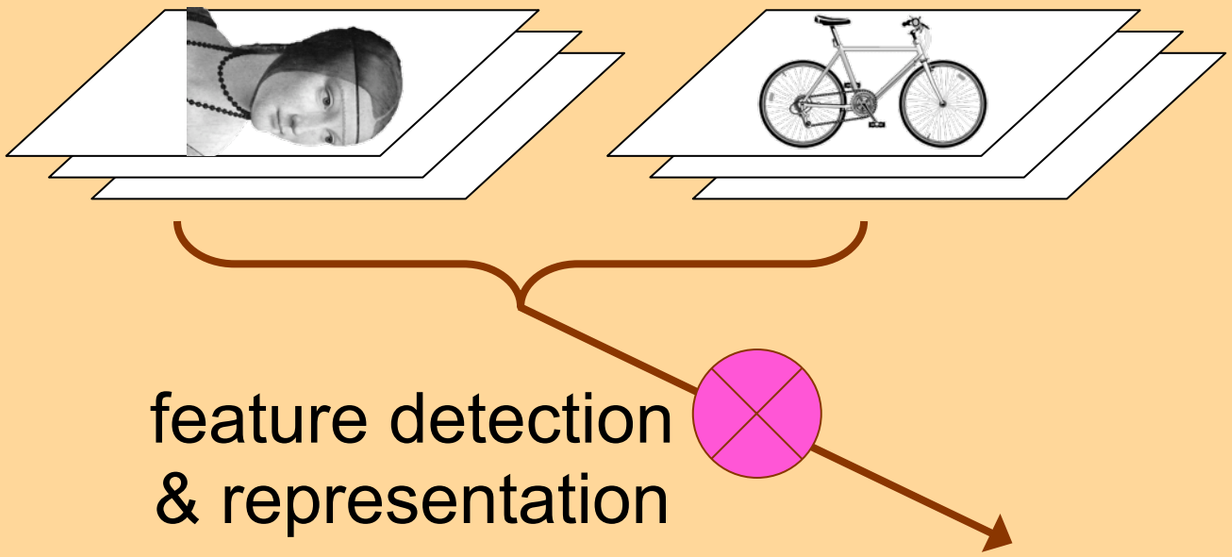




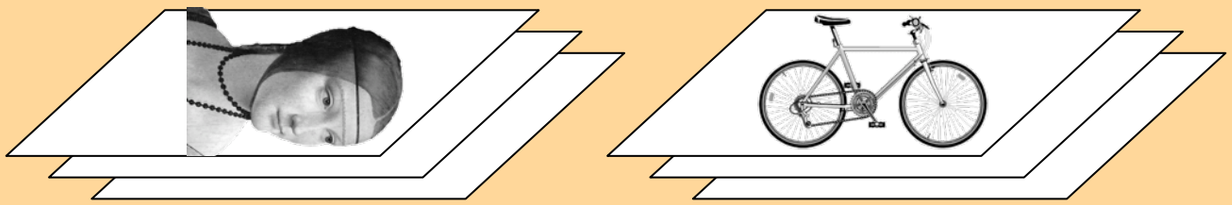
learning



learning

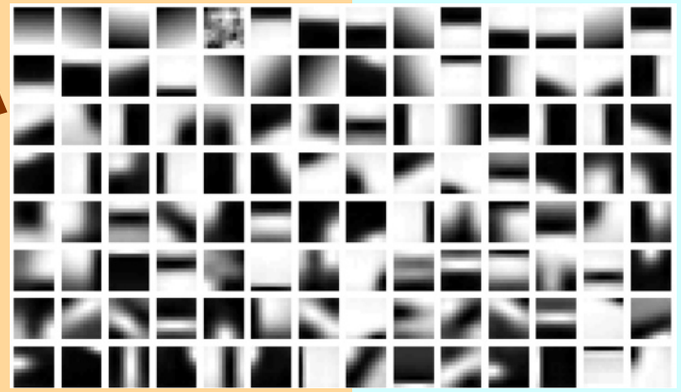


learning

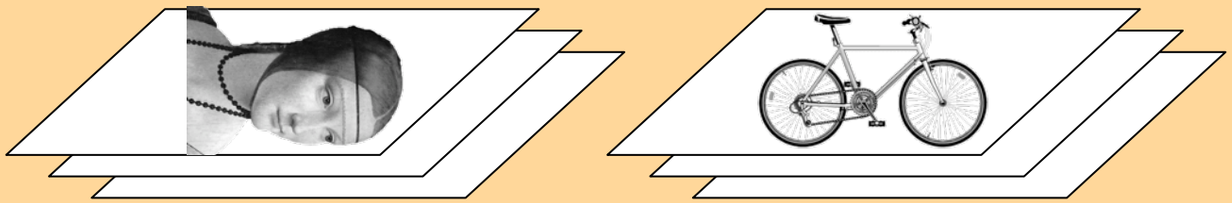


feature detection
& representation

codewords dictionary



learning



feature detection
& representation

codewords dictionary

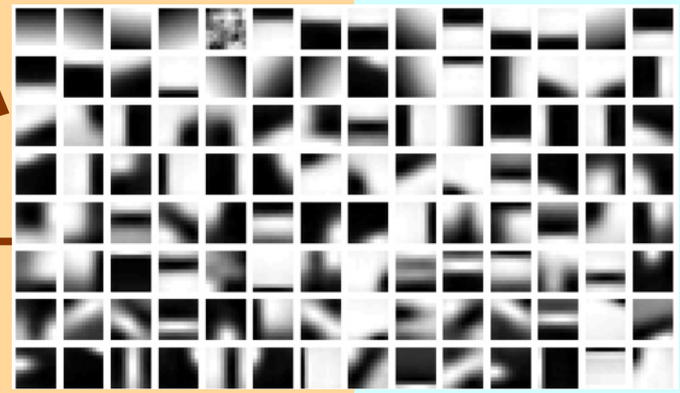
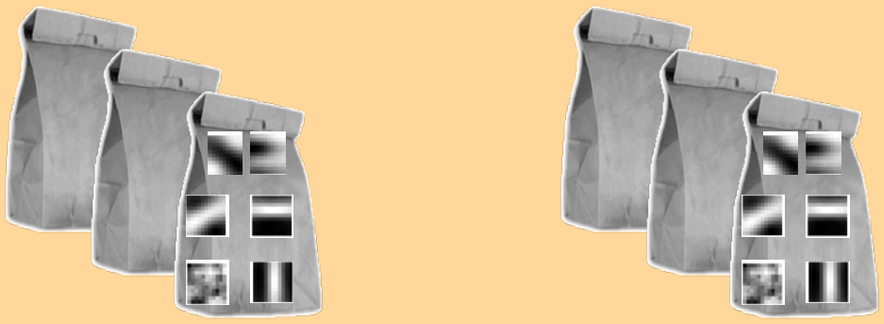
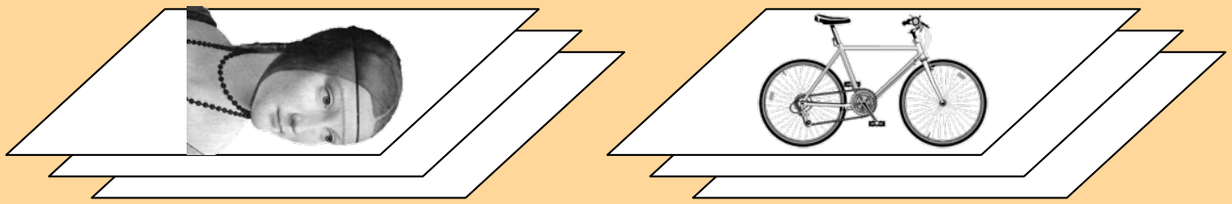


image representation



learning



feature detection
& representation

codewords dictionary

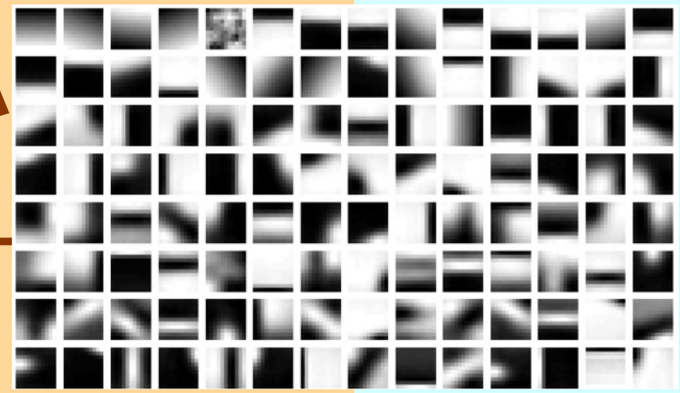
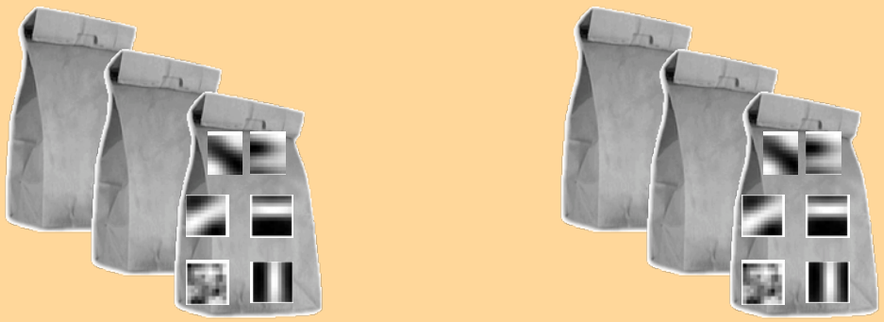
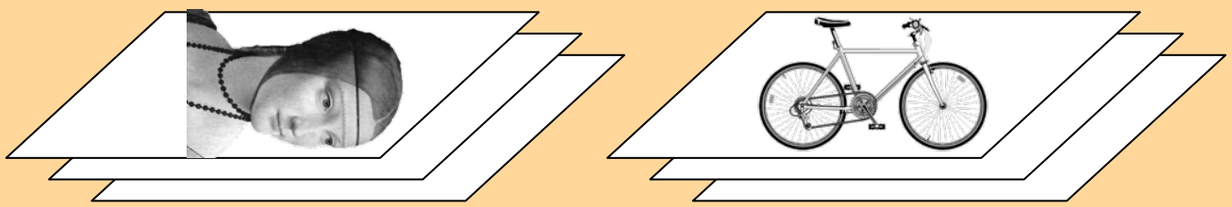


image representation



category models (and/or) classifiers

learning



feature detection
& representation

codewords dictionary

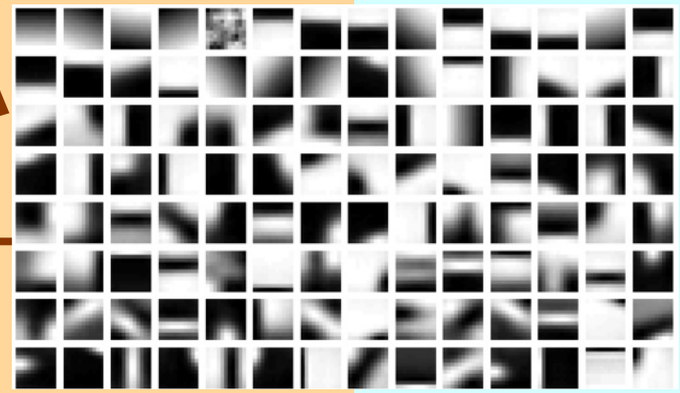
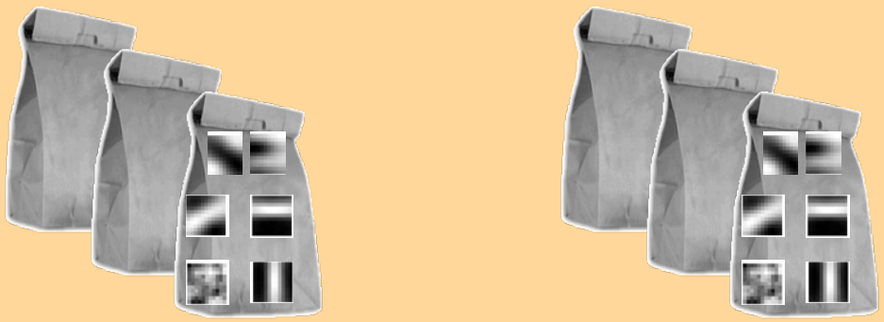


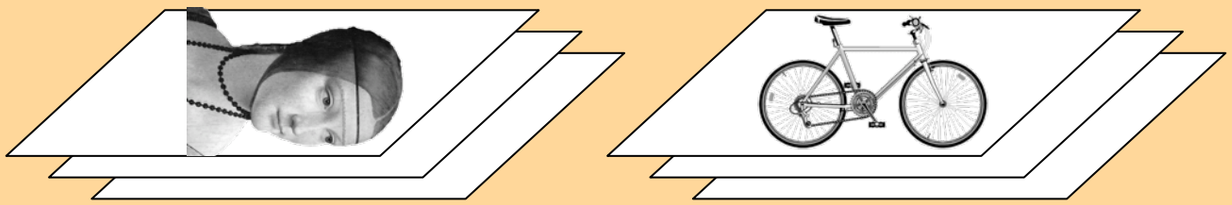
image representation



category models (and/or) classifiers

recognition

learning



feature detection
& representation

codewords dictionary

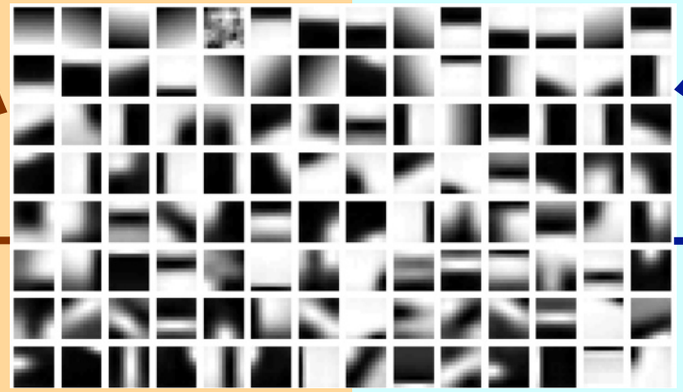
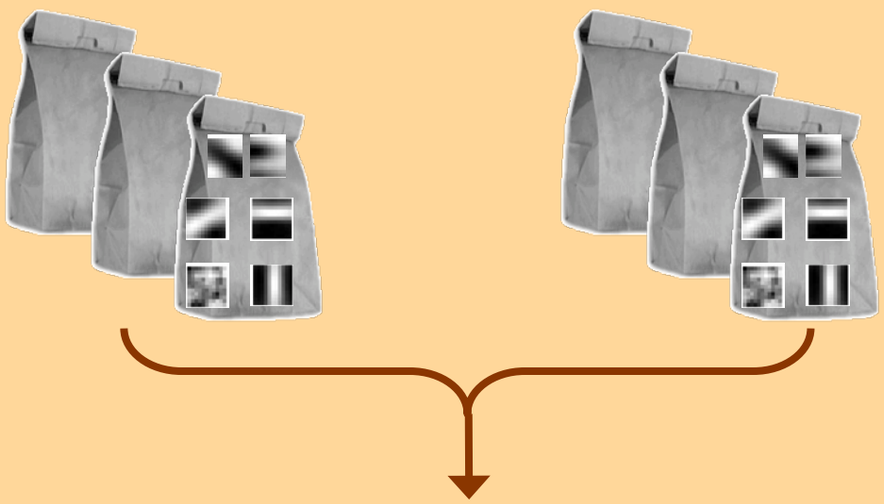
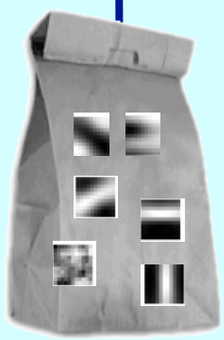


image representation

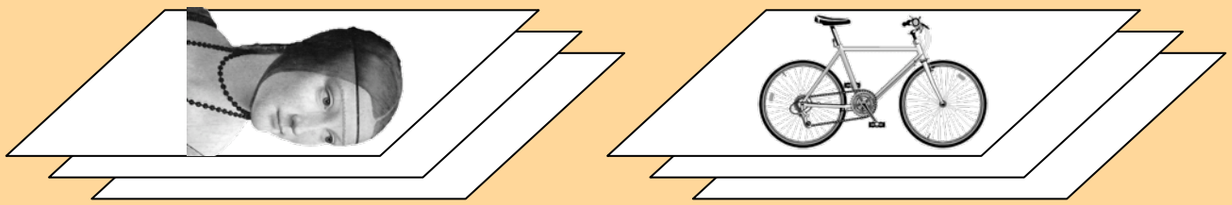


category models (and/or) classifiers

recognition



learning



feature detection
& representation

codewords dictionary

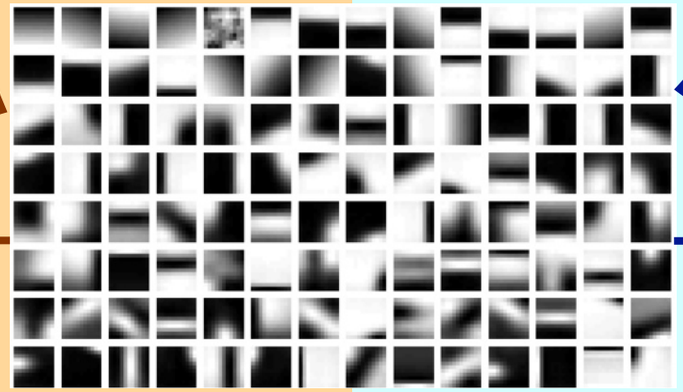
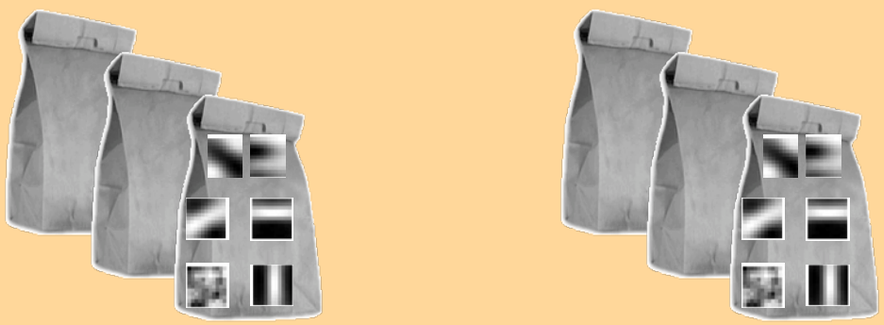
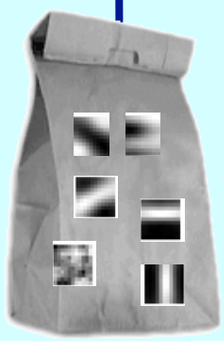


image representation



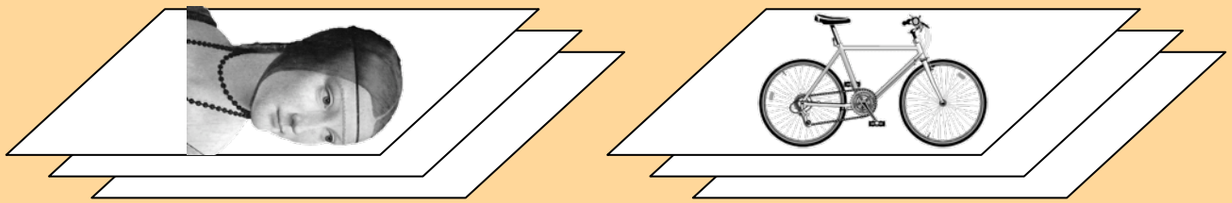
category models (and/or) classifiers

recognition



category decision

Representation



feature detection
& representation

codewords dictionary

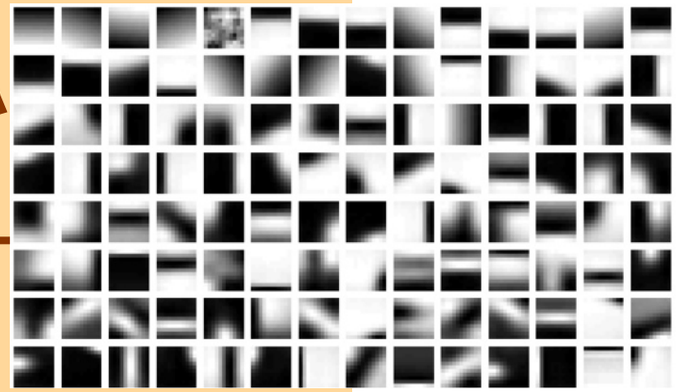
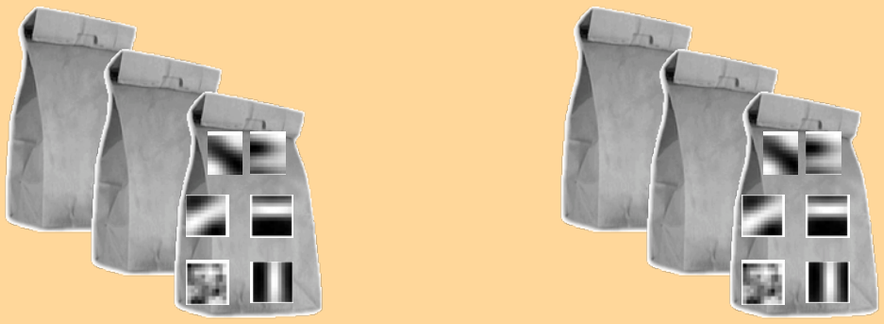
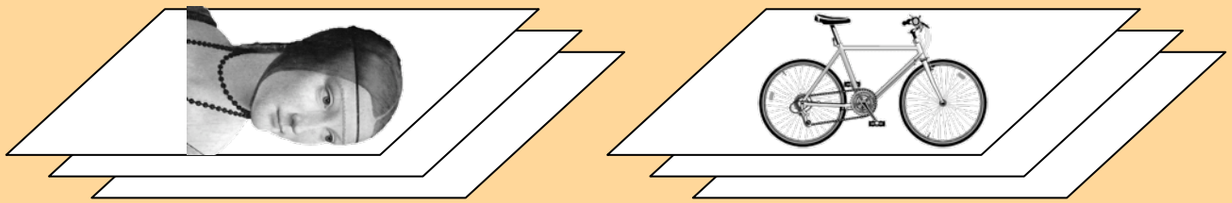


image representation



Representation



1. feature detection & representation

2. codewords dictionary

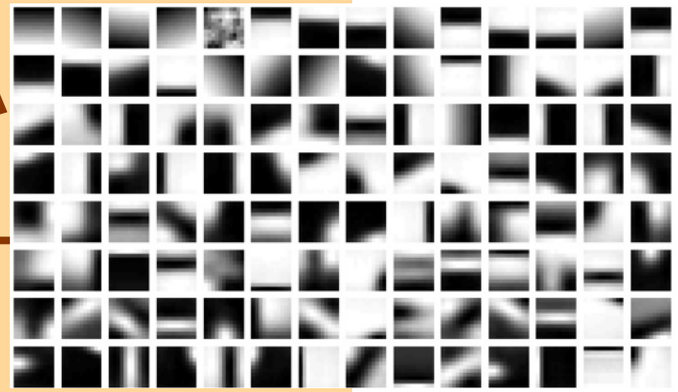
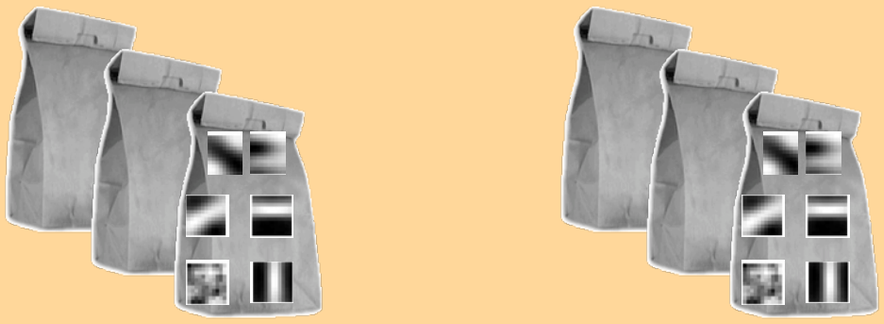


image representation

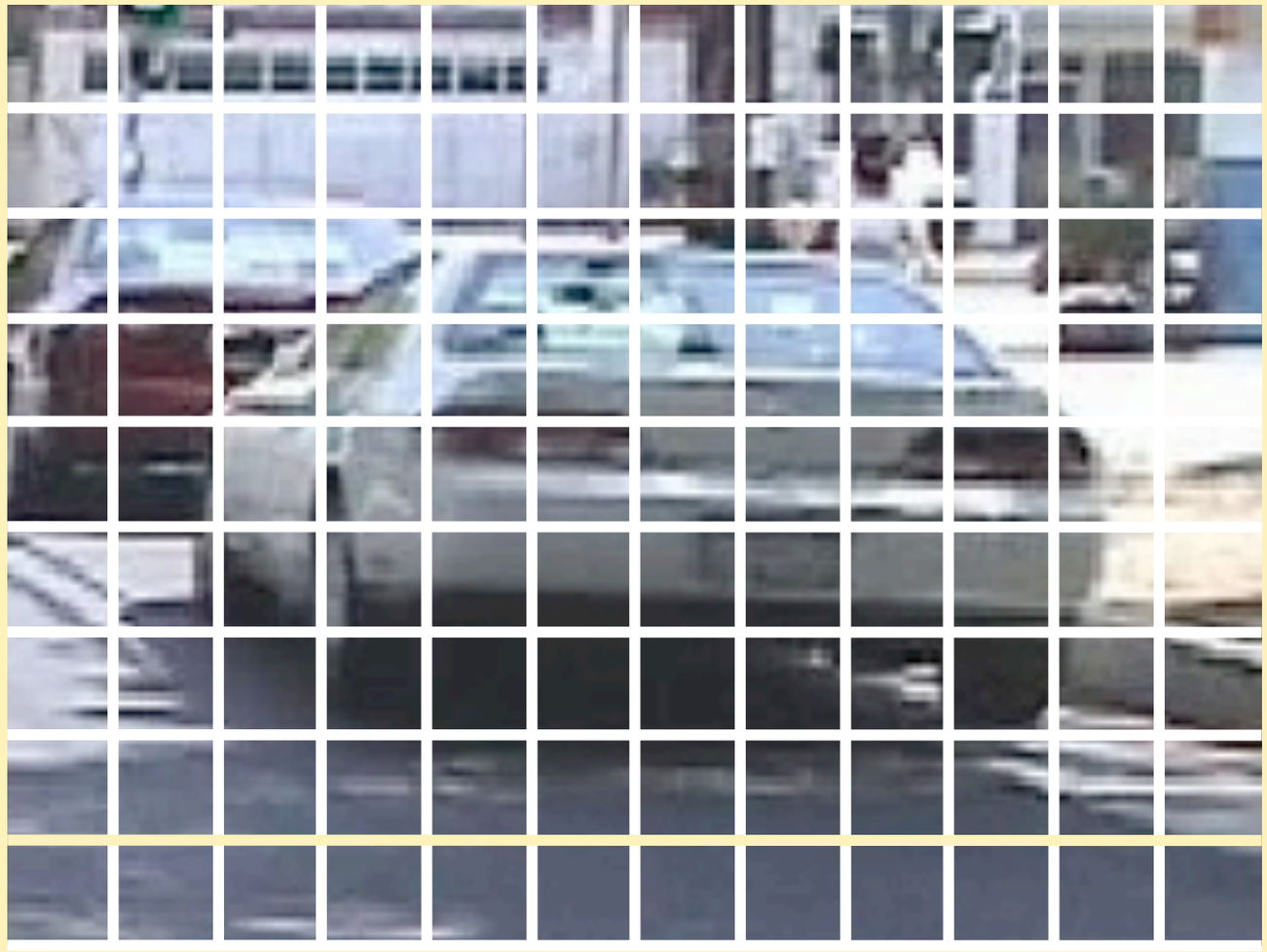
3.



1.Feature detection and representation

1.Feature detection and representation

- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005

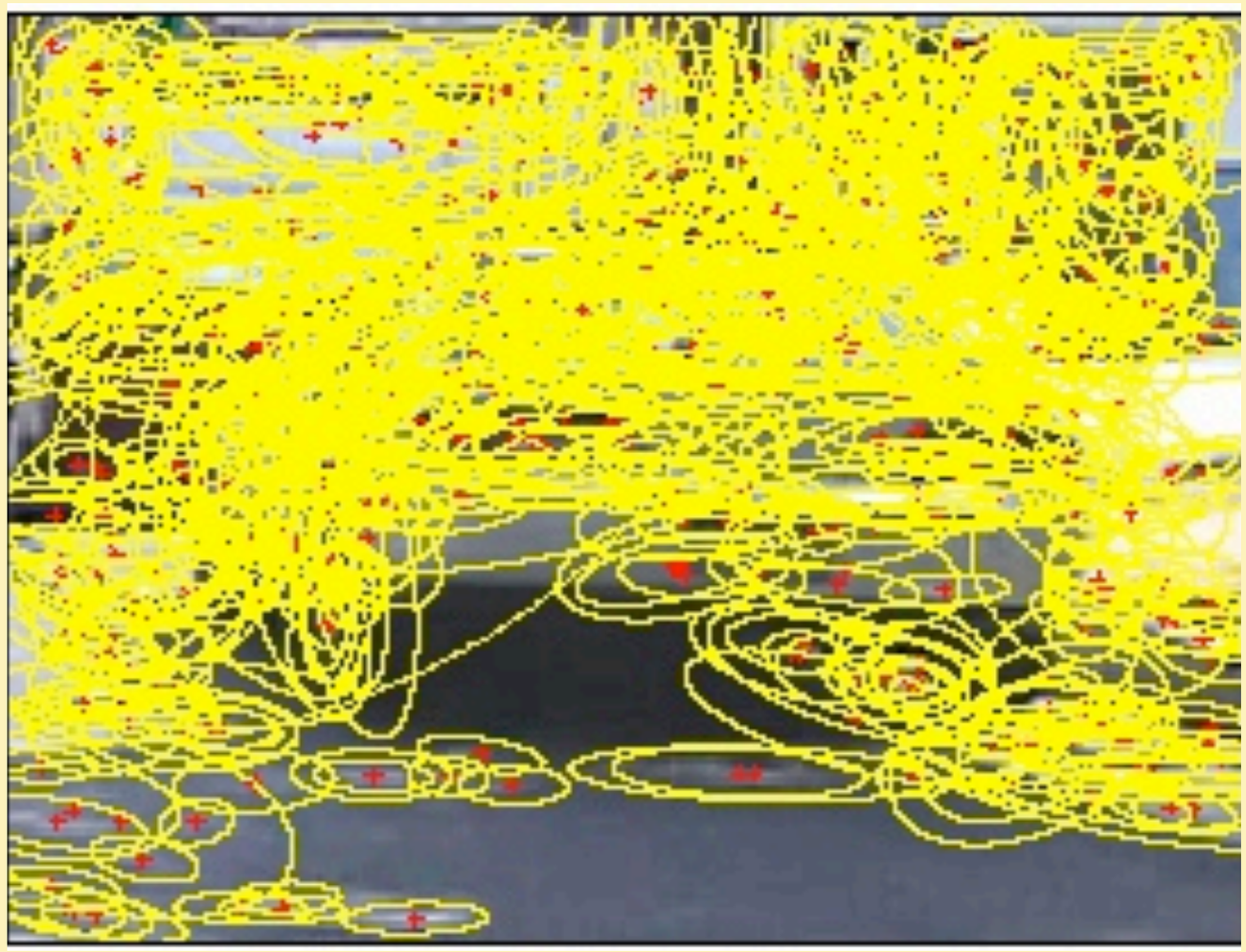


1.Feature detection and representation

- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei et al. 2005
  - Sivic et al. 2005

1.Feature detection and representation

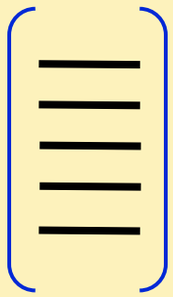
- Regular grid
  - –Vogel et al. 2003
  - –Fei-Fei et al. 2005
- Interest point detector
  - –Csurka et al. 2004
  - –Fei-Fei et al. 2005
  - –Sivic et al. 2005



1.Feature detection and representation

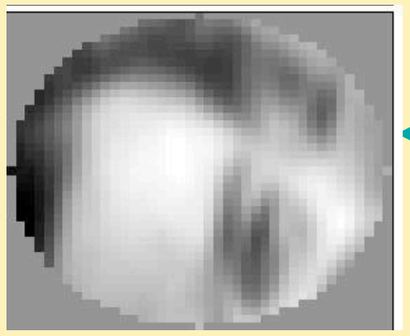
- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei et al. 2005
  - Sivic et al. 2005
- Other methods
  - Random sampling (Ullman et al. 2002)
  - Segmentation based patches (Barnard et al. 2003)

1.Feature detection and representation

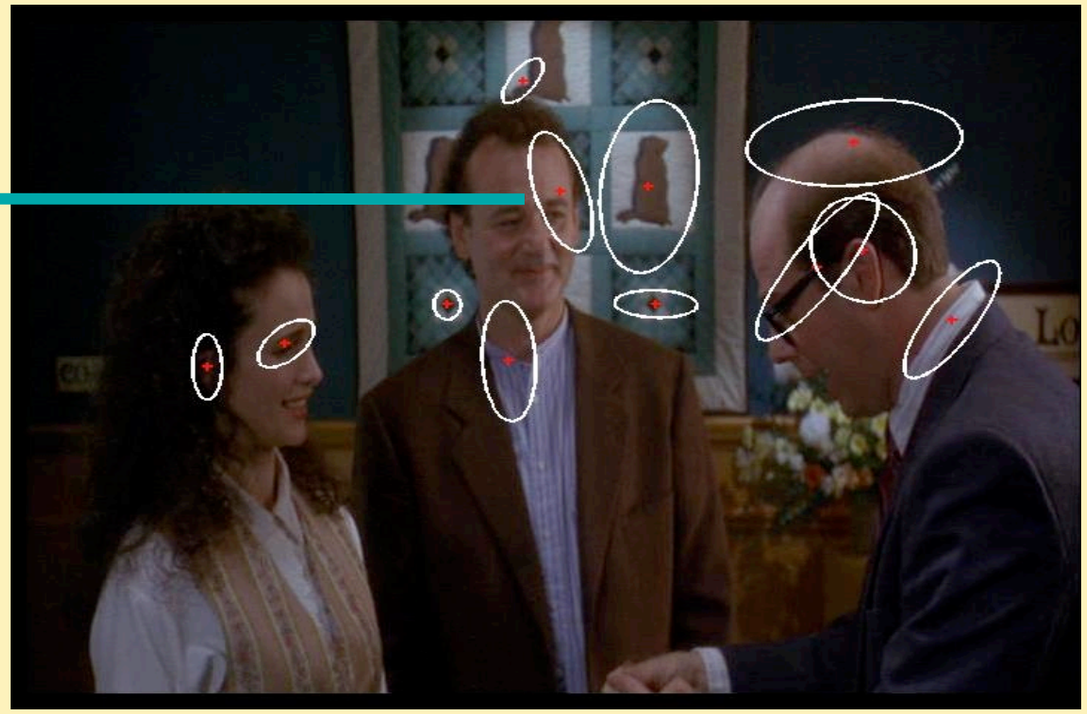


Compute SIFT descriptor

[Lowe'99]



Normalize patch



Detect patches

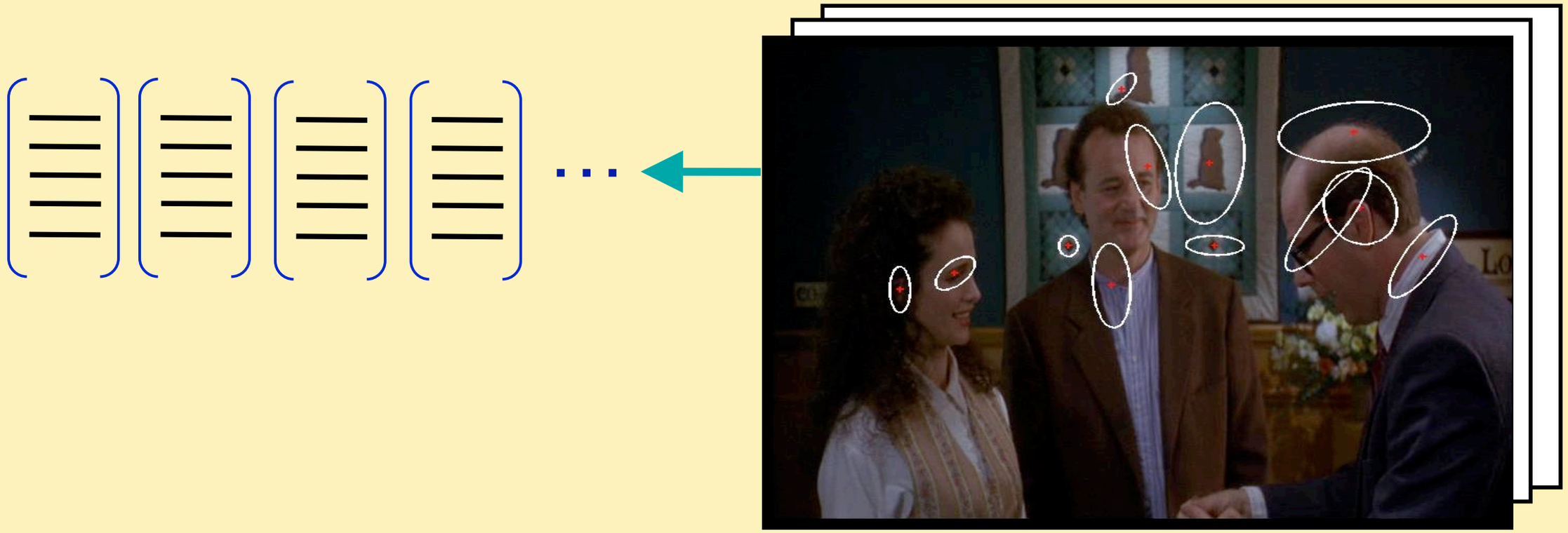
[Mikojaczyk and Schmid '02]

[Matas et al. '02]

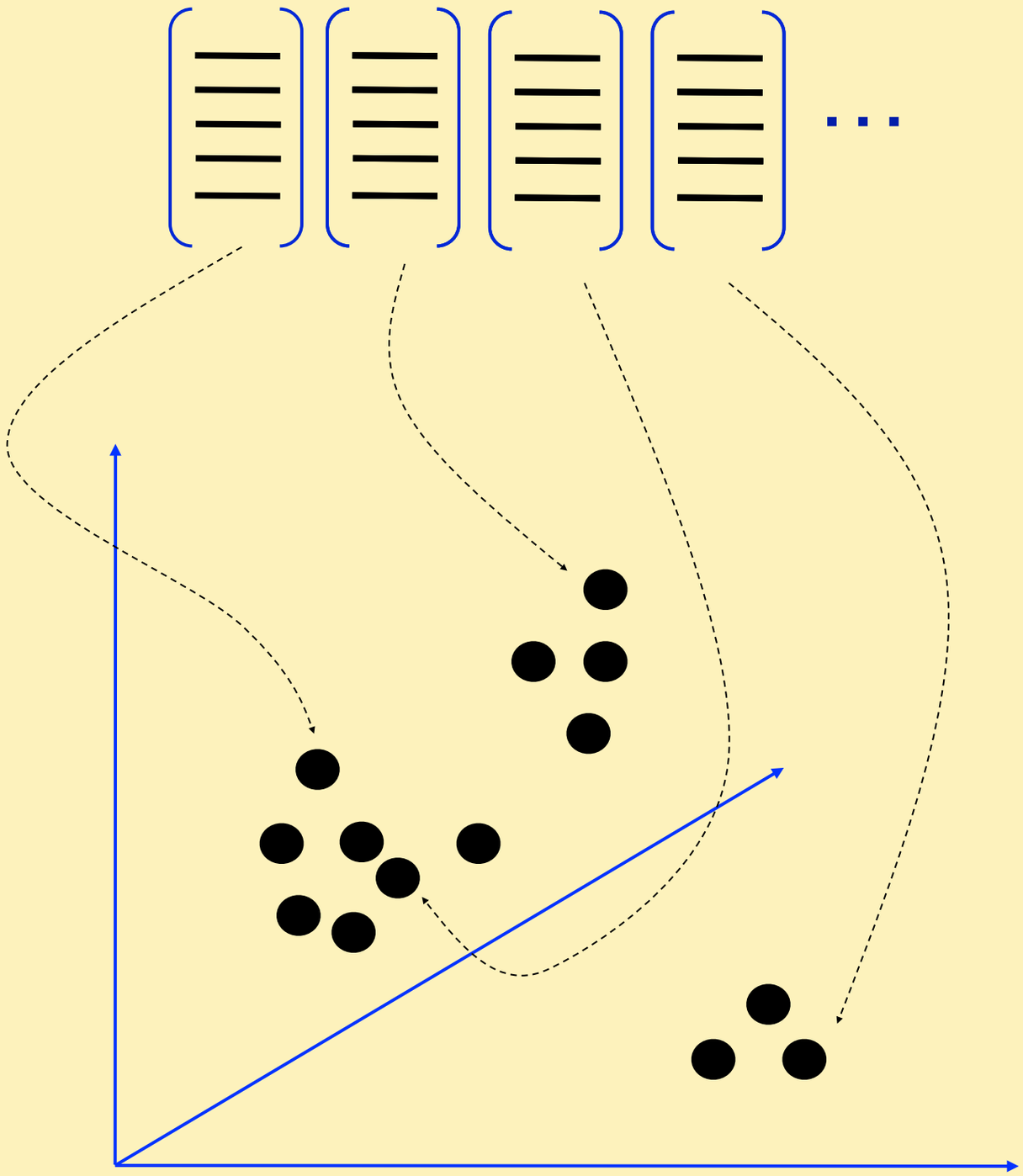
[Sivic et al. '03]

Slide credit: Josef Sivic

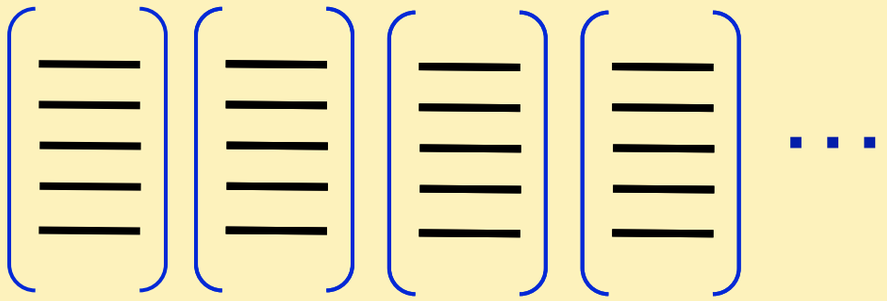
1.Feature detection and representation

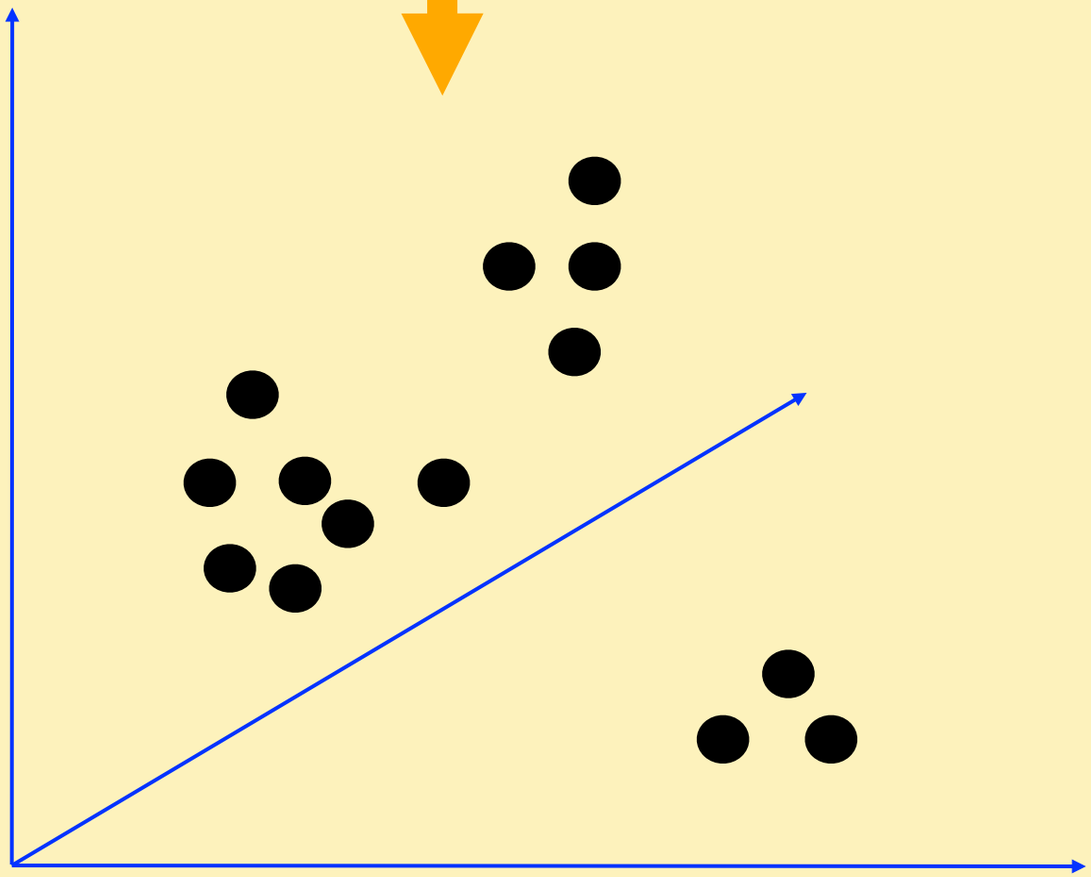


2. Codewords dictionary formation

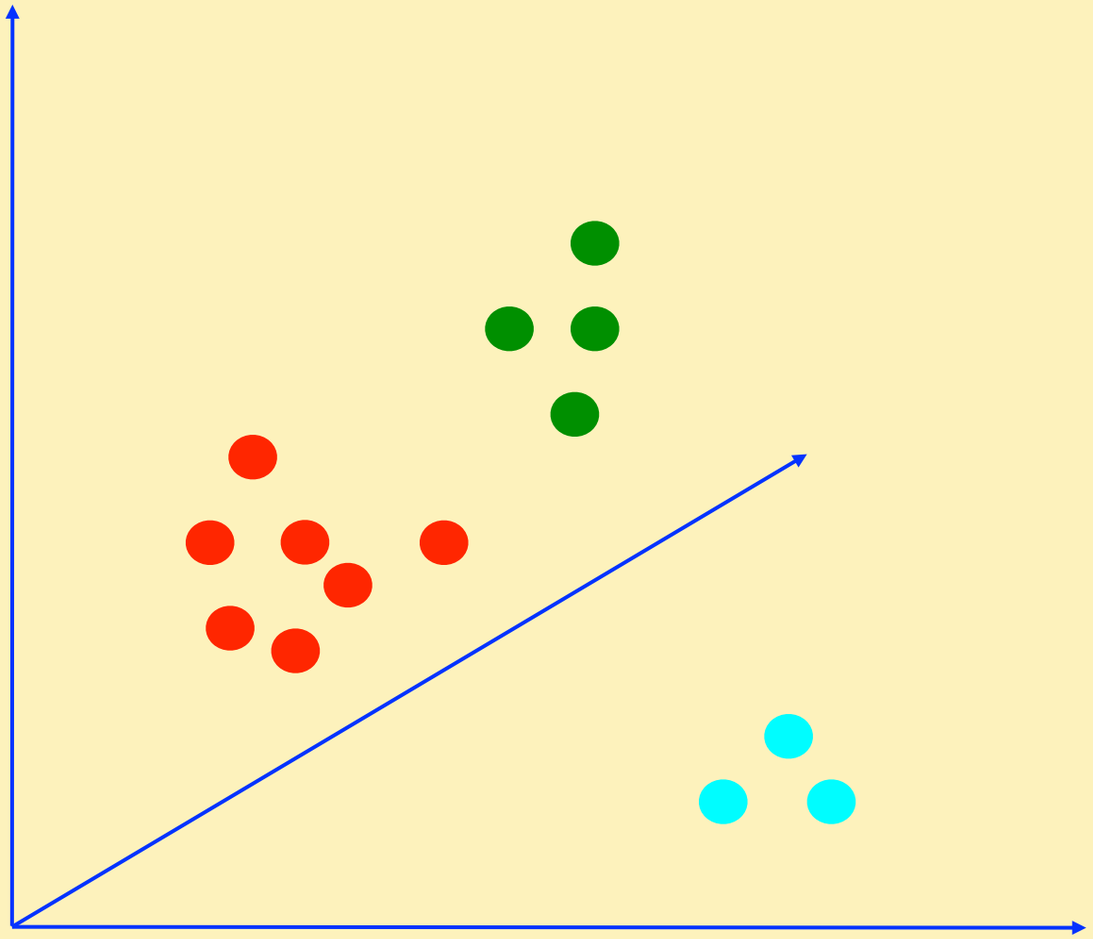


2. Codewords dictionary formation



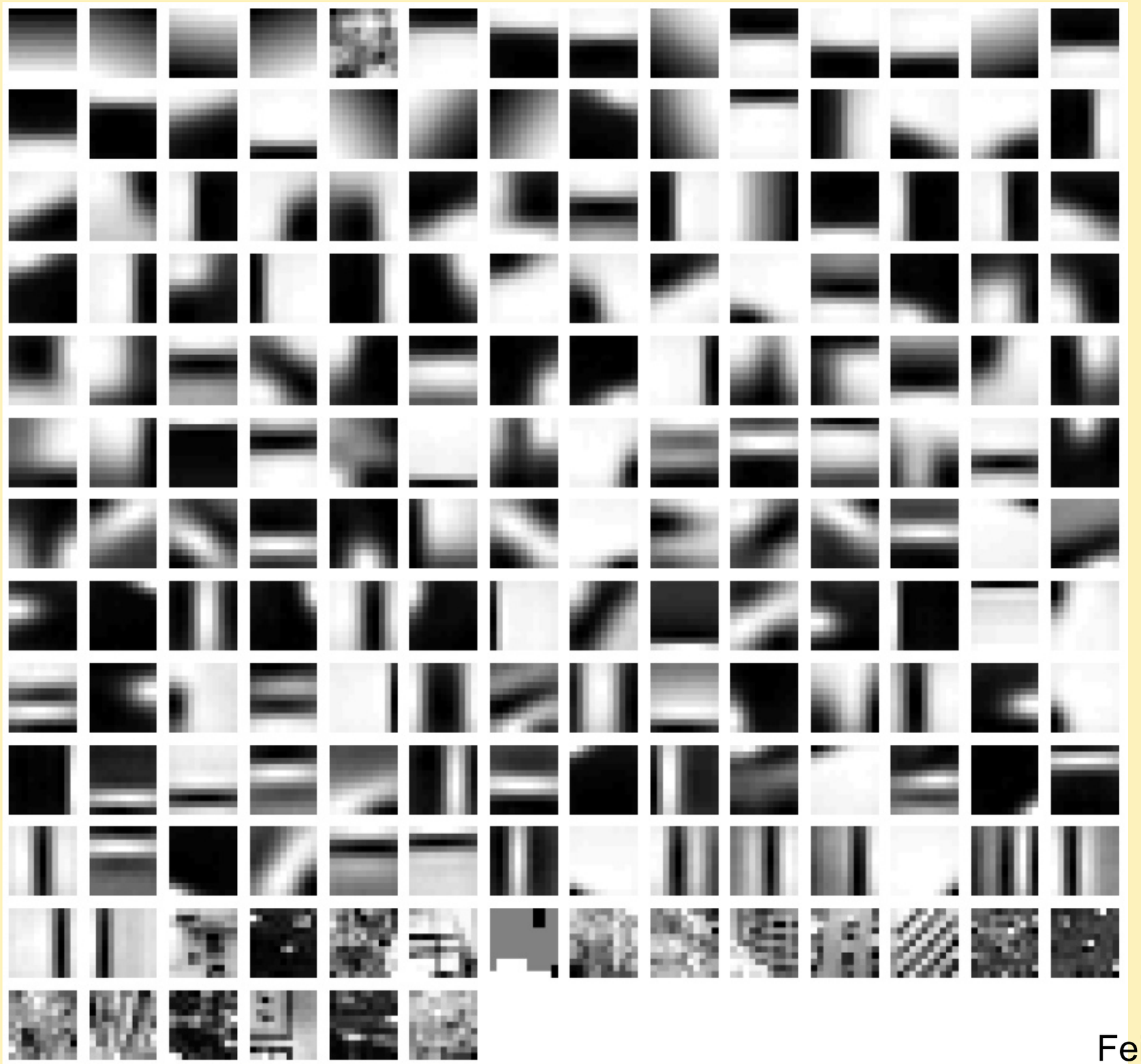


Vector quantization



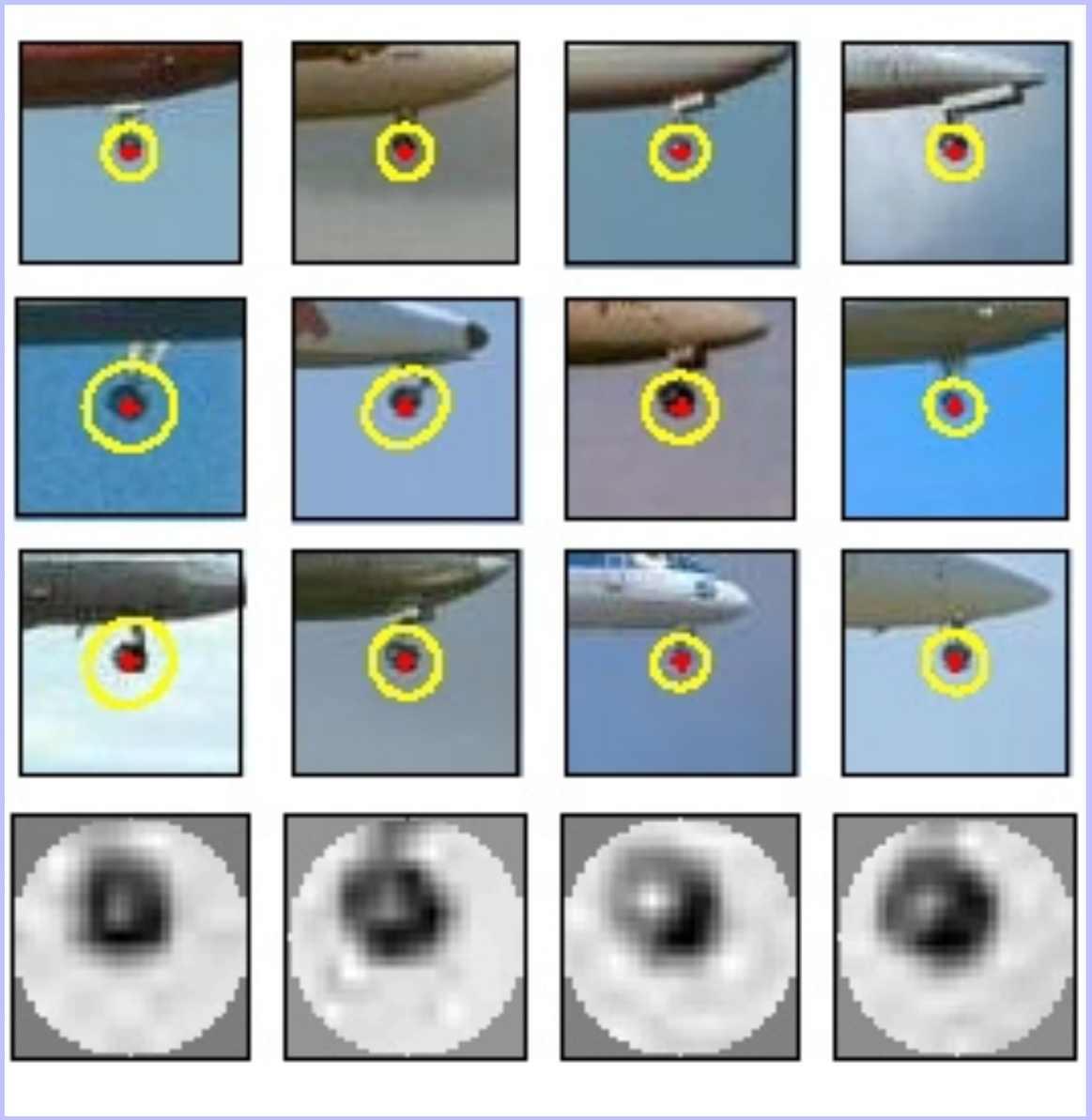
Slide credit: Josef Sivic

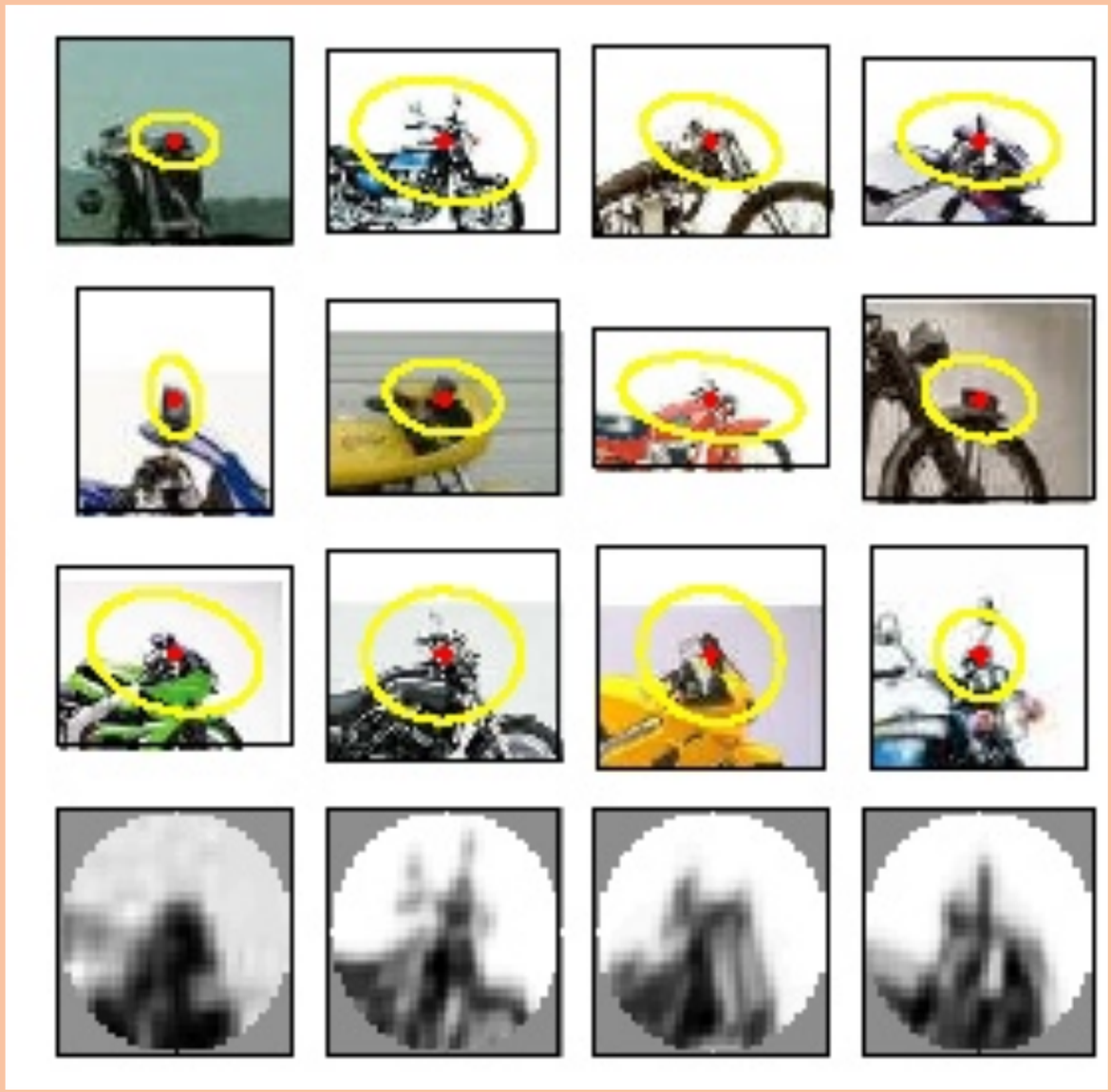
2. Codewords dictionary formation



Fei-Fei et al. 2005

Image patch examples of codewords

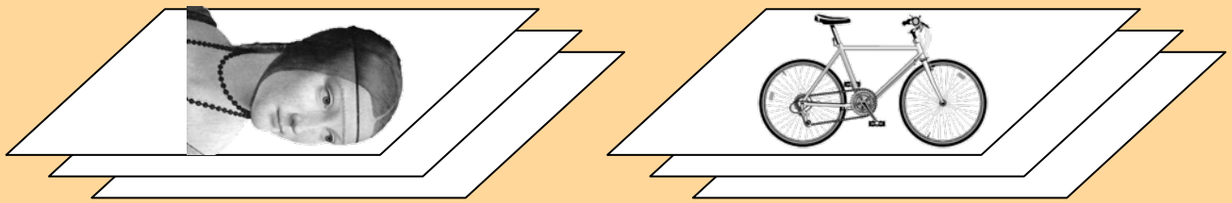




Sivic et al. 2005

3. Image representation

Representation



1. feature detection & representation

2. codewords dictionary

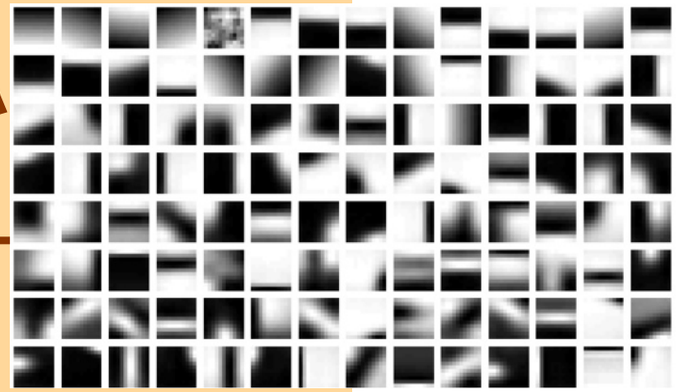
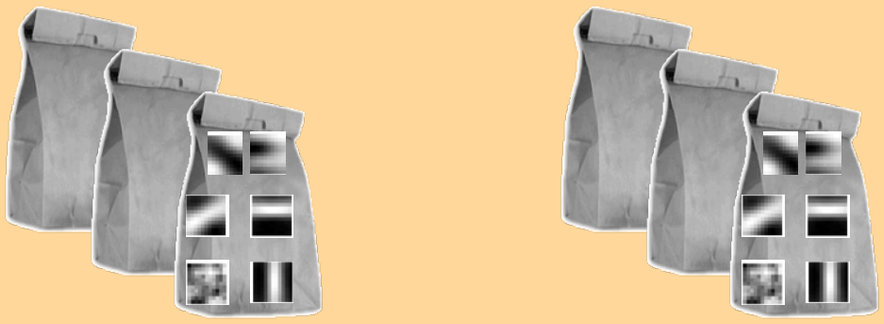


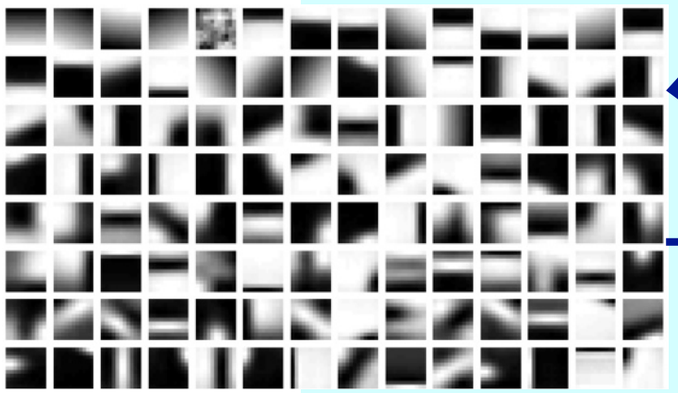
image representation

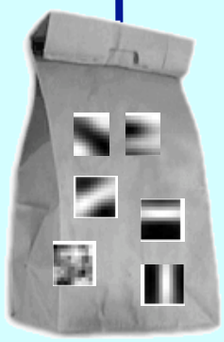
3.



Learning and Recognition

codewords dictionary





category models (and/or) classifiers

category decision

Nude Detection in Video using Bag-of-Visual-Features

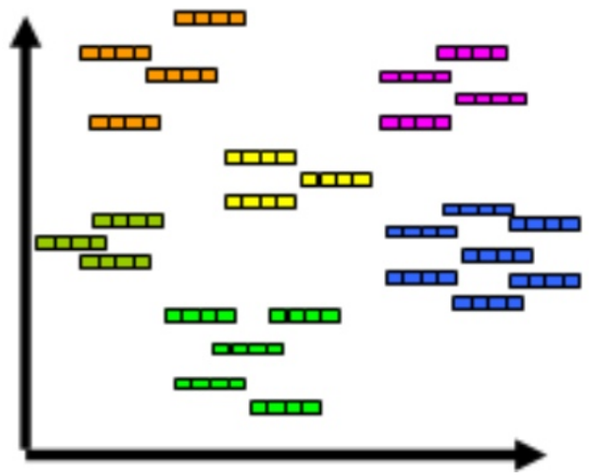
Ana Paula B. Lopes*†, Sandra E. F. de Avila*, Anderson N. A. Peixoto*,
Rodrigo S. Oliveira*, Marcelo de M. Coelho*‡ and Arnaldo de A. Araújo*

*Computer Science Department, Federal University of Minas Gerais – UFMG
31270–010, Belo Horizonte, MG, Brazil
†Exact and Technological Sciences Department, State University of Santa Cruz – UESC
45662–000, Ilhéus, BA, Brazil
‡Preparatory School of Air Cadets – EPCAR
36205–900, Barbacena, MG, Brazil

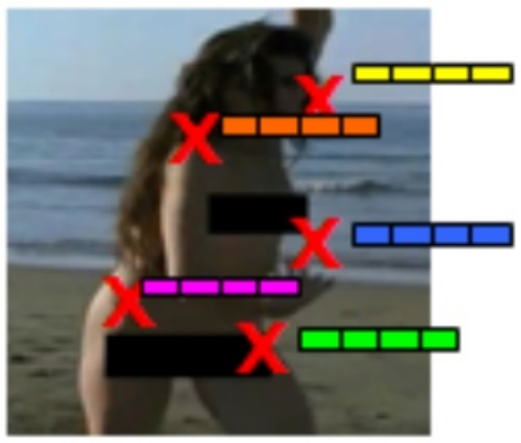
Representing Images as BoVF

1) Point selection

2) Point description



3) Vocabulary discovery

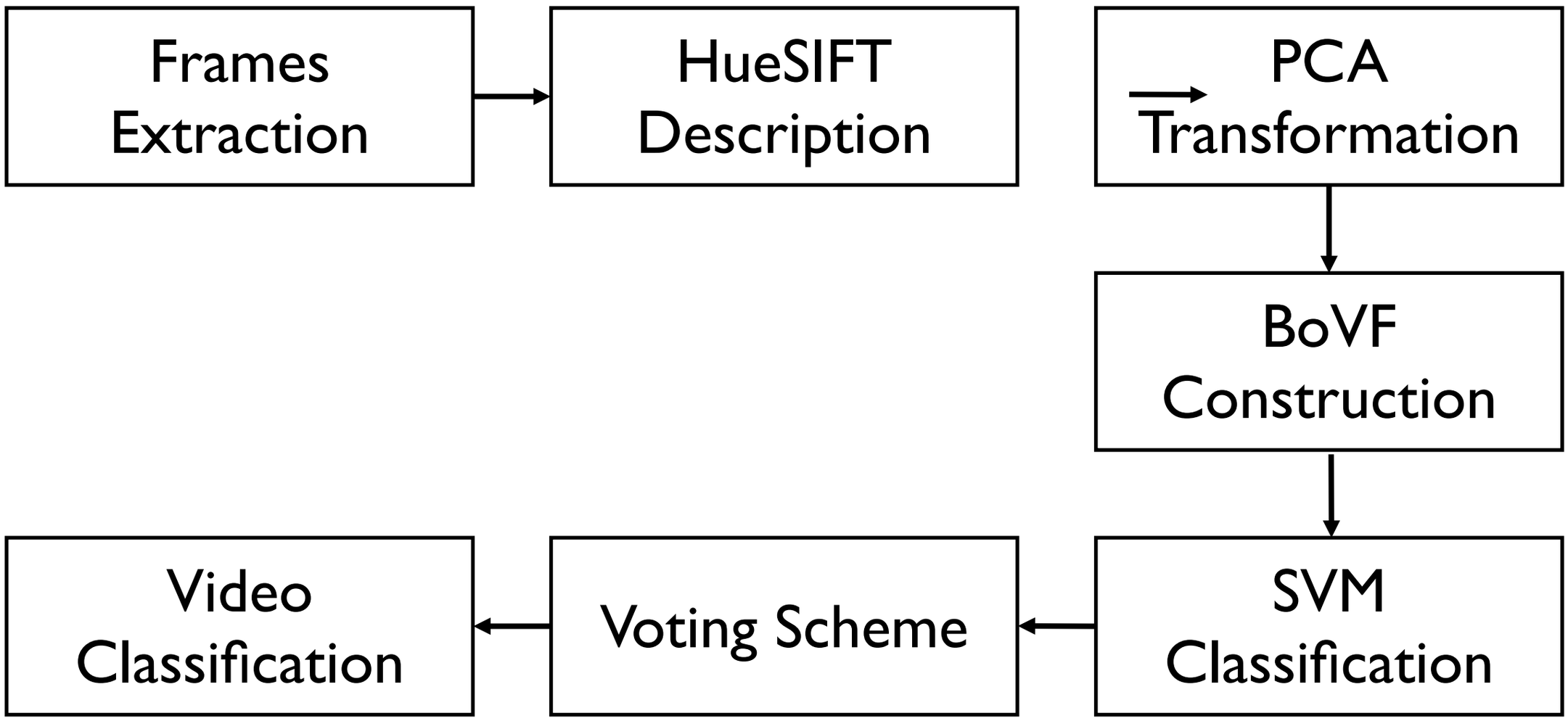


4) Cluster association

5) Histogram computation

Detecting Nudity from Videos

Lopes et al. 2009 - Approach



Experimental Results

- Database
  - 179 segments
  - Nude sequences
  - From 43 to 308 frames long
  - Non-nude sequences
  - From 50 to 278 frames long
- http://www.dcc.ufmg.br/nudeDetection

Experimental Results

- Experimental Setup
  - Frames selection - Sample rates – 1/15 and 1/30
  - BoVF creation
    - 10,000 random HueSIFT points
    - Vocabulary size: 60, 120 and 180
  - Linear SVM classifier
    - $10^{-5} \leq C \leq 10^{5}$
    - 30 5-folds cross-validation runs

Experimental Results

Table I: Comparing recognition rates for keyframe and voting based classification.

Voc. Size	Keyframe (%)	Voting (%)	Increase
60	76.4 ± 0.2	77.1 ± 0.4	0.7
120	80.2 ± 0.3	80.9 ± 0.4	0.7
180	83.9 ± 0.2	88.4 ± 0.6	4.5

(a) 1/30 frames

Voc. Size	Keyframe (%)	Voting (%)	Increase
60	79.1 ± 0.1	80.5 ± 0.4	1.4
120	83.7 ± 0.2	87.3 ± 0.4	3.6
180	85.9 ± 0.1	93.2 ± 0.4	7.3

(b) 1/15 frames

Experimental Results

Table II: False-negative rates for keyframe and voting based classification.

Voc. Size	Keyframe (%)	Voting (%)	Decrease
60	12.2 ± 0.2	10.4 ± 0.3	1.8
120	11.0 ± 0.2	9.1 ± 0.2	1.9
180	8.0 ± 0.2	4.2 ± 0.3	3.3

(a) 1/30 frames

Voc. Size	Keyframe (%)	Voting (%)	Decrease
60	10.7 ± 0.1	10.7 ± 0.3	0.0
120	10.0 ± 0.1	8.5 ± 0.2	1.5
180	7.5 ± 0.1	4.2 ± 0.2	3.3

(b) 1/15 frames

General Comments

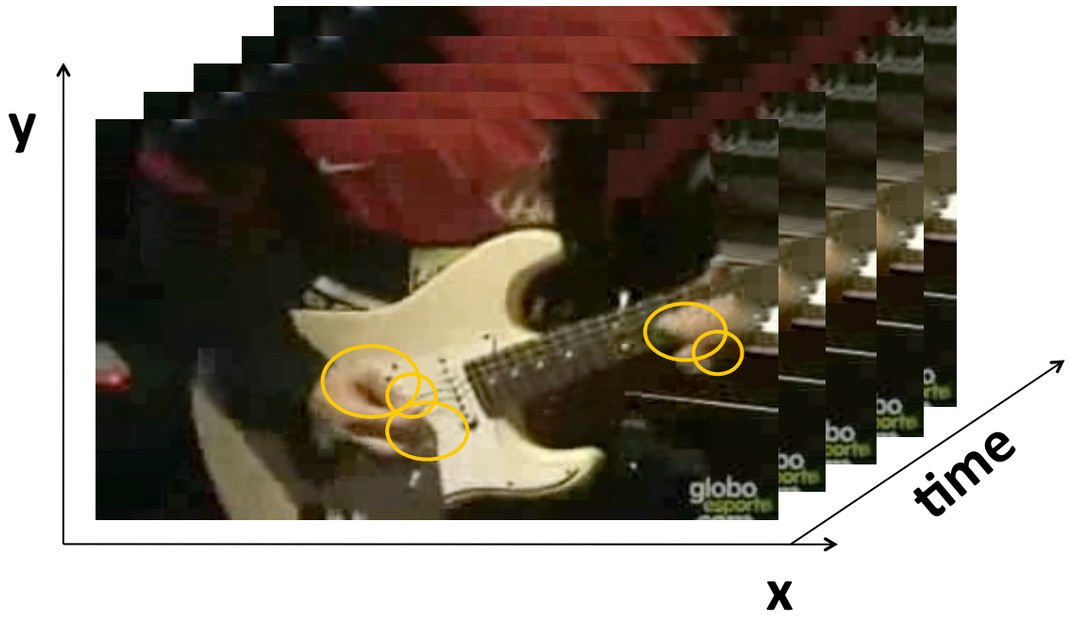
- 93.2% of correct classification
- Identified misclassification causes:
  - Background colors near to skin tones
  - Presence of large skin areas
  - Illumination variations

An Evaluation on Color Invariant Based Spatiotemporal Features for Action Recognition

Fillipe Dias Moreira de Souza, DCC/UFMG
Prof. Dr. Guillermo Cámara Chávez, DECOMP/UFOP (co-advisor)
Prof. Dr. Arnaldo de Albuquerque Araújo, DCC/UFMG (advisor)

Theory on STIP

Extension of the 2D Harris corner detector to 3D space;



Invariant to multiple scales.

Theory on STIP

Given a sequence of images
f(x,y,t) = f(.),

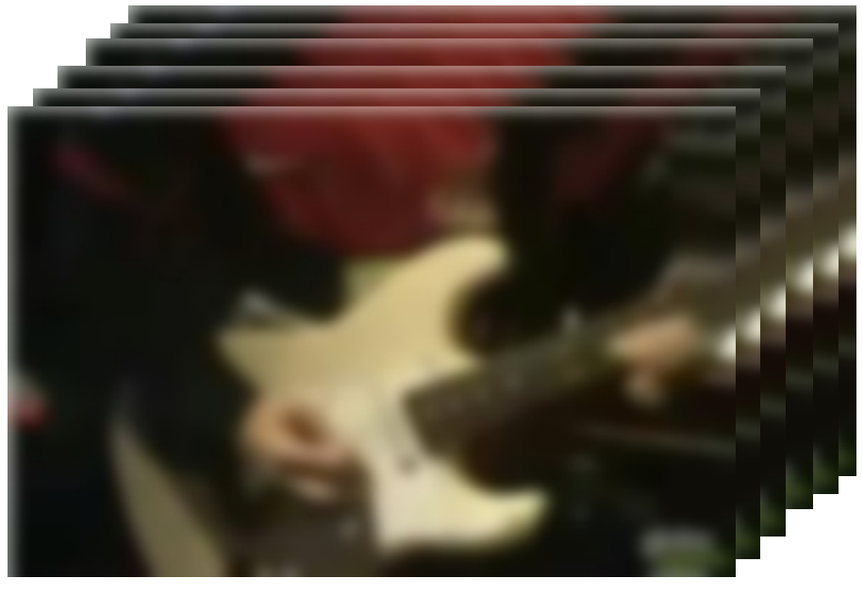
Theory on STIP

Given a sequence of images

$$f(x,y,t) = f(.),$$

its scale-space representation is denoted by

$$L(.;\sigma_l^2) = g(.;\sigma_l^2,\tau_l^2) * f(.)$$



Theory on STIP

Multiscale 3D Harris matrix

$$\mu = g(.; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

Theory on STIP

Gaussian kernel

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}}$$
$$\times \exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/\tau_l^2),$$

2D Harris matrix

Multiscale 3D Harris matrix

$$\mu = g(.; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

Theory on STIP

Gaussian kernel

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}}$$
$$\times \exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/\tau_l^2),$$

2D Harris matrix

Multiscale 3D Harris matrix

$$\mu = g(.; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

Saliency measurement

$$H = \det(\mu) - k.\mathrm{trace}^3(\mu)$$
$$= \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3$$

Theory on STIP

(simulating STIP processing)

Input video to STIP

Theory on STIP

(simulating STIP processing)

STIP detects spatiotemporal interest points (IPs)

Input video to STIP

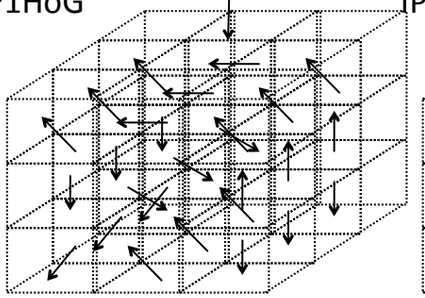
Theory on STIP

(simulating STIP processing)

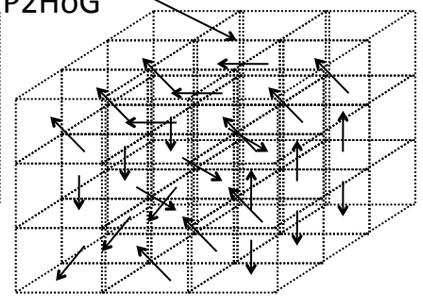
Input video to STIP

STIP detects spatiotemporal interest points (IPs)

IP1HoG

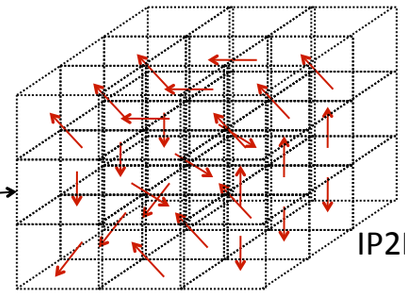


IP2HoG

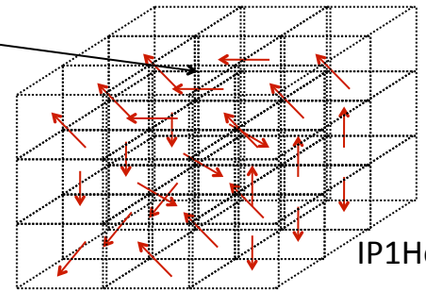


STIP describes IPs in terms of histograms of oriented gradients

STIP describes IPs in terms of histograms of optical flow



IP2HoF



IP1HoF

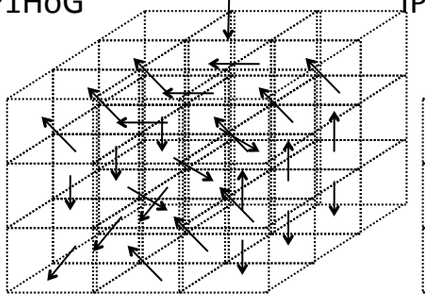
Theory on STIP

(simulating STIP processing)

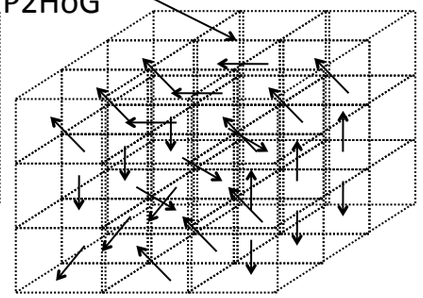
Input video to STIP

STIP detects spatiotemporal interest points (IPs)

IP1HoG

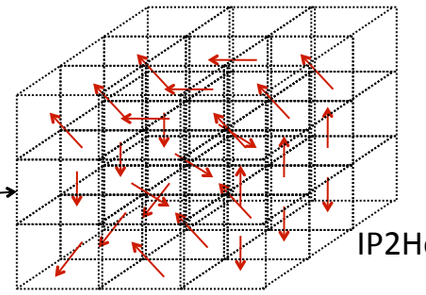


IP2HoG

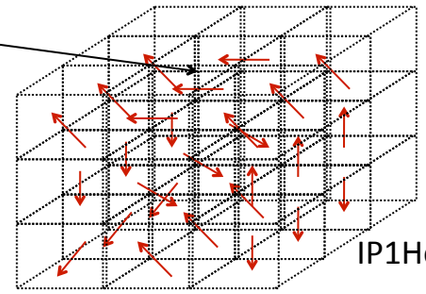


STIP describes IPs in terms of histograms of oriented gradients

STIP describes IPs in terms of histograms of optical flow

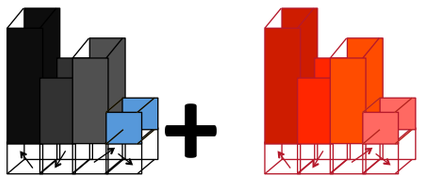


IP2HoF

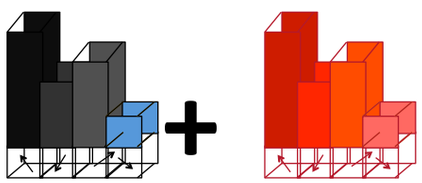


IP1HoF

Concatenating descriptors



IP1HoG + IP1HoF



IP2HoG + IP2HoF

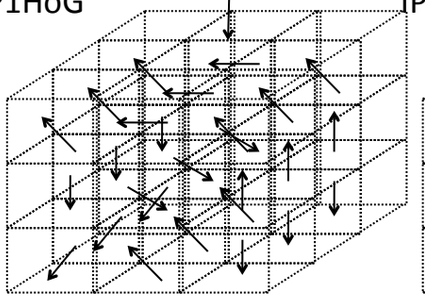
Theory on STIP

(simulating STIP processing)

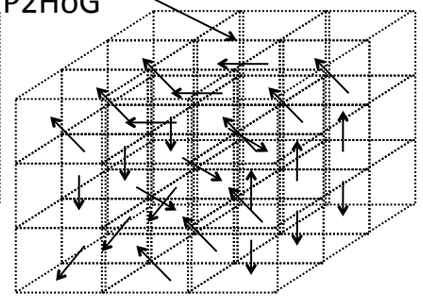
Input video to STIP

STIP detects spatiotemporal interest points (IPs)

IP1HoG

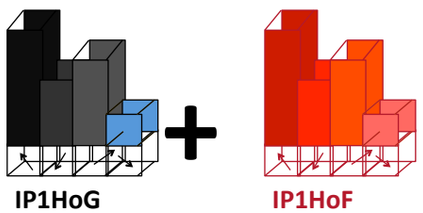


IP2HoG

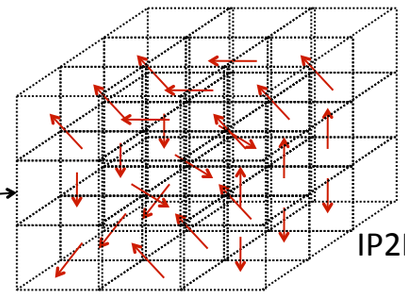


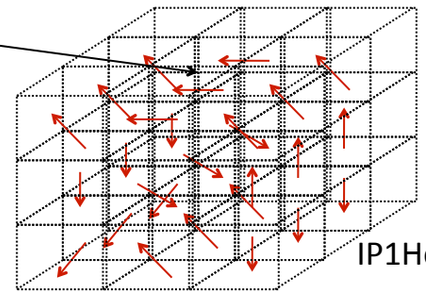
STIP describes IPs in terms of histograms of oriented gradients

Concatenating descriptors

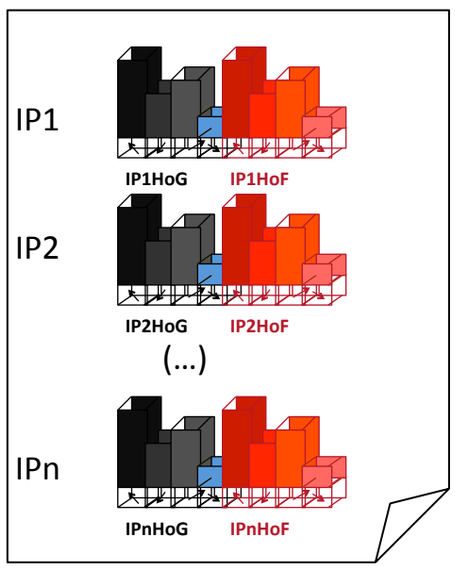


STIP describes IPs in terms of histograms of optical flow





The set of features representing the video content (STIP output)

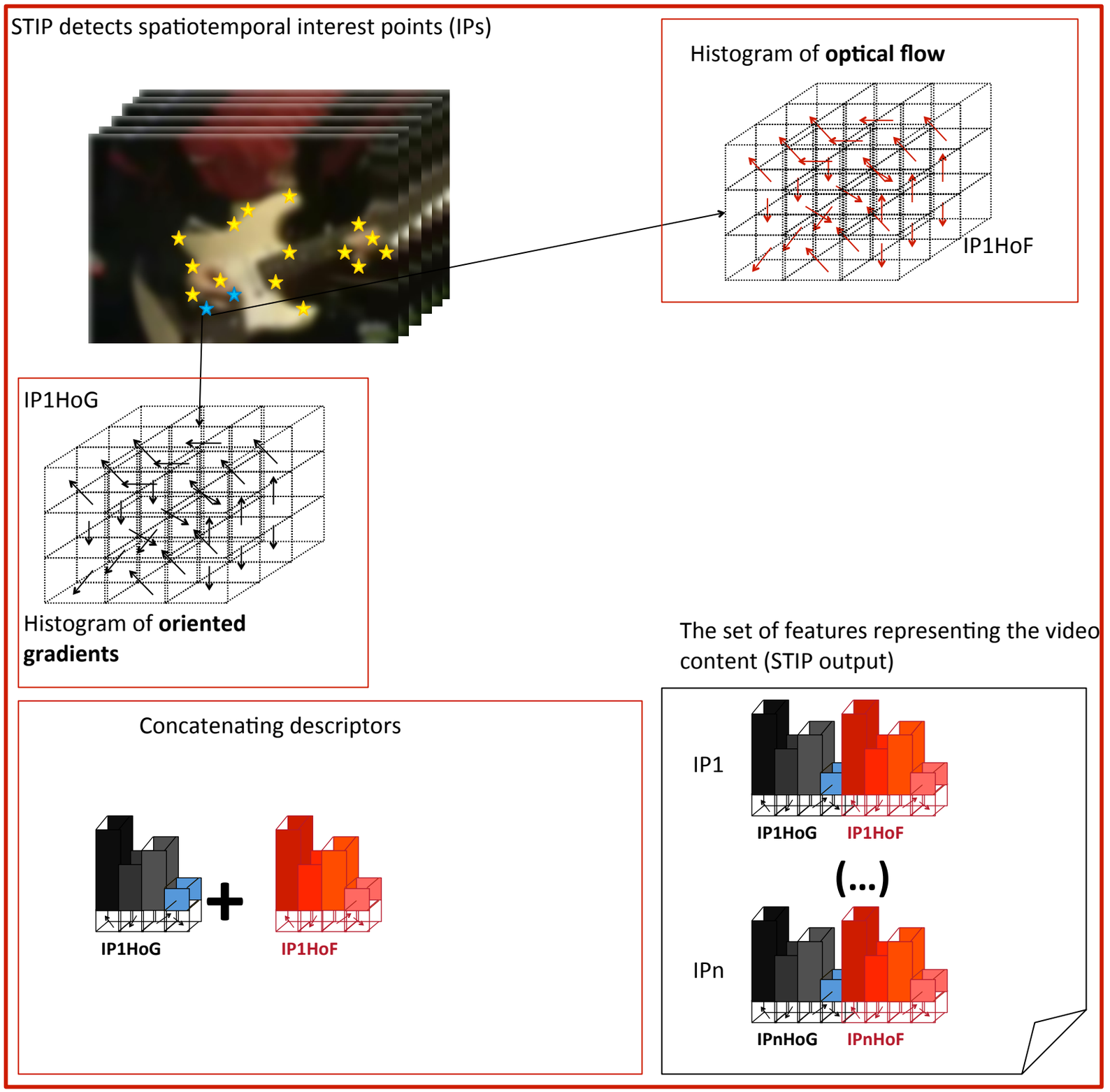


Color-based STIPs

HueSTIP

STIP also uses color histograms to describe local spatiotemporal features

Hue histogram has **36 bins**

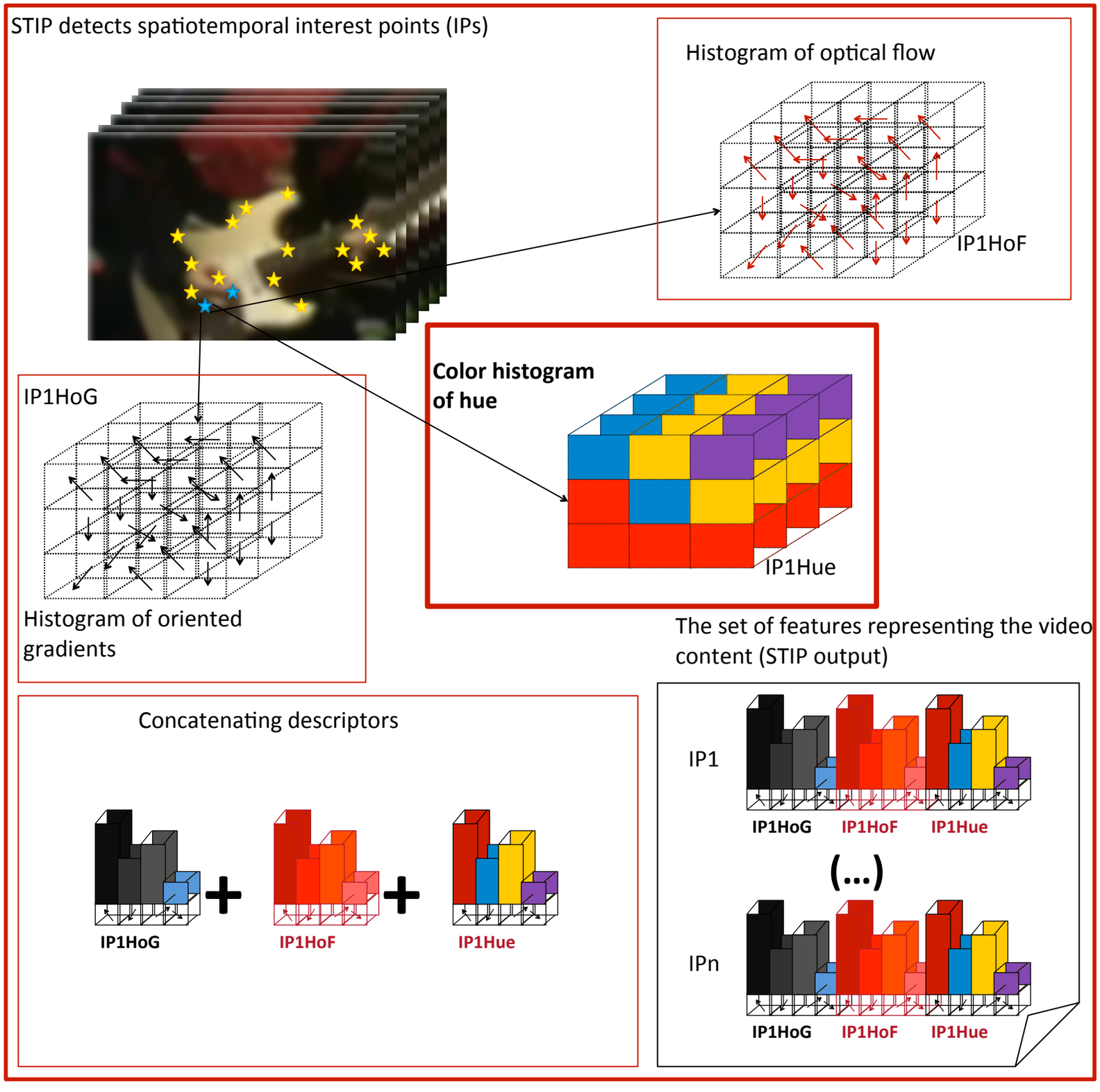


Color-based STIPs

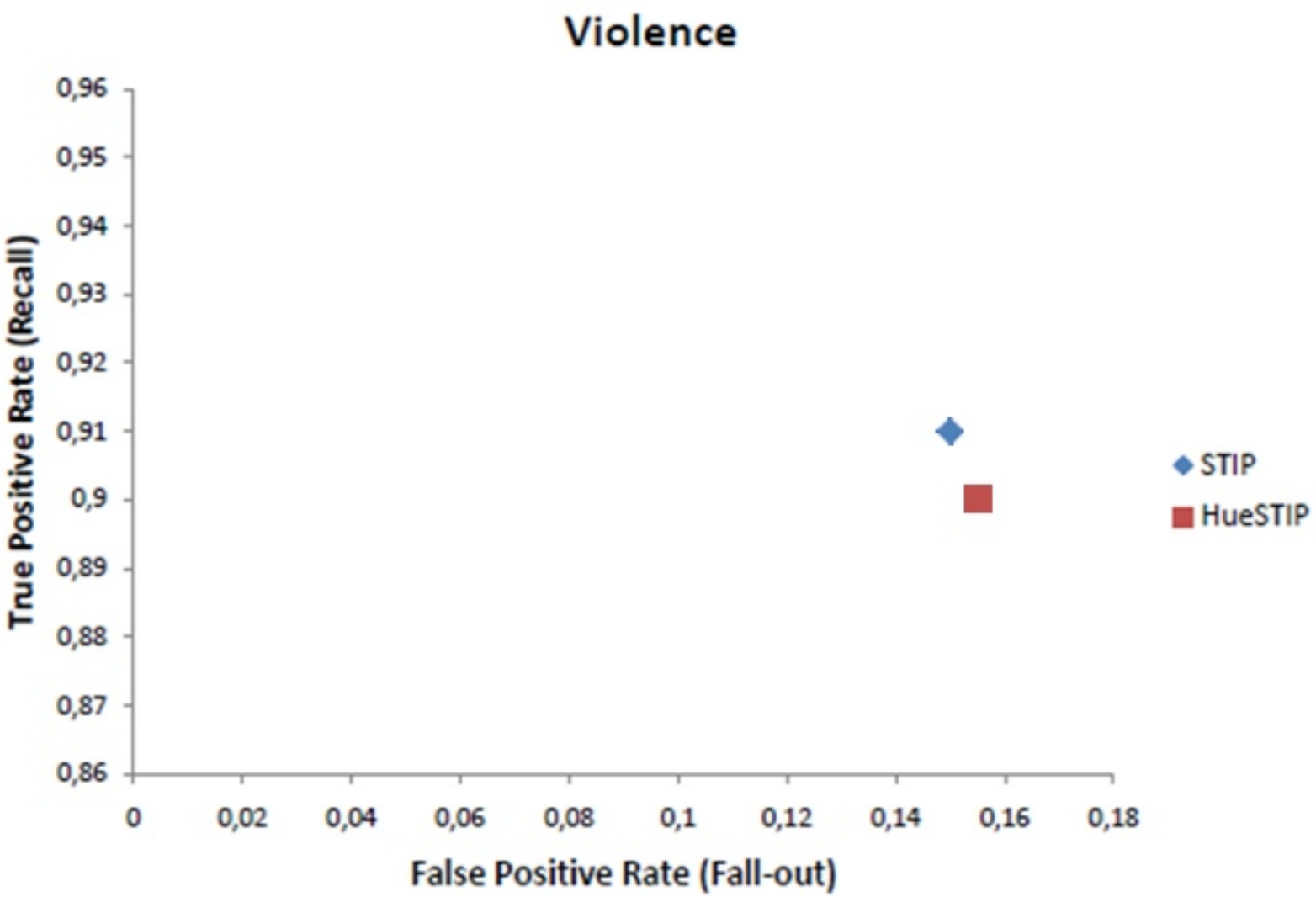
HueSTIP

STIP also uses color histograms to describe local spatiotemporal features

Hue histogram has **36 bins**



Experimental Results



STIP		
Class	*Violence*	*NonViolence*
Violence	**.91**	.09
NonViolence	.15	**.85**

HUESTIP		
Class	*Violence*	*NonViolence*
Violence	**.90**	.10
NonViolence	.155	**.845**

Experimental Results

Fights taking place in outdoor environments



Surveillance on traffic

Professional fights

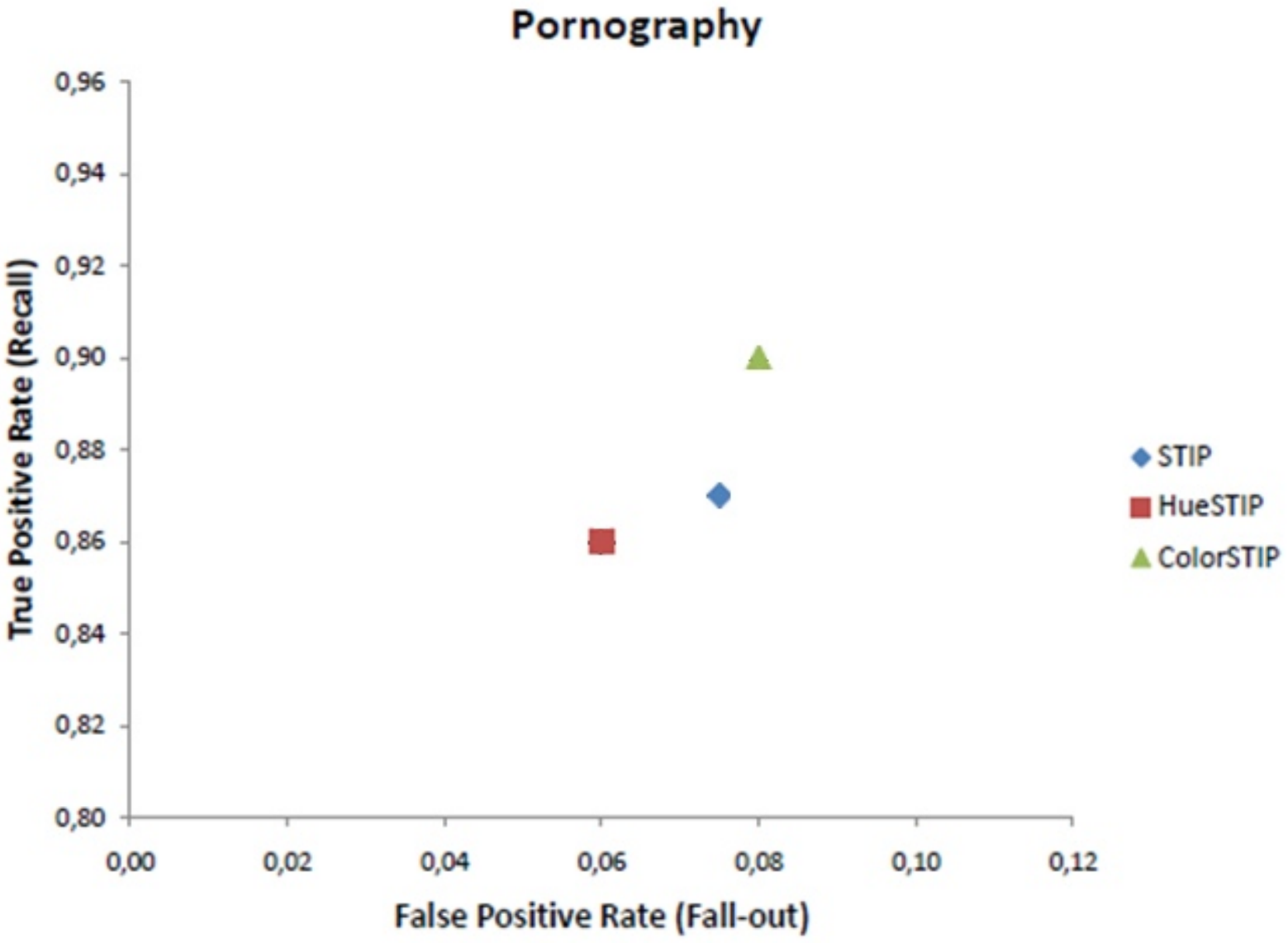

Fights in crowd

Fight between members of gangs

Fights in sports

Fight in school

Experimental Results

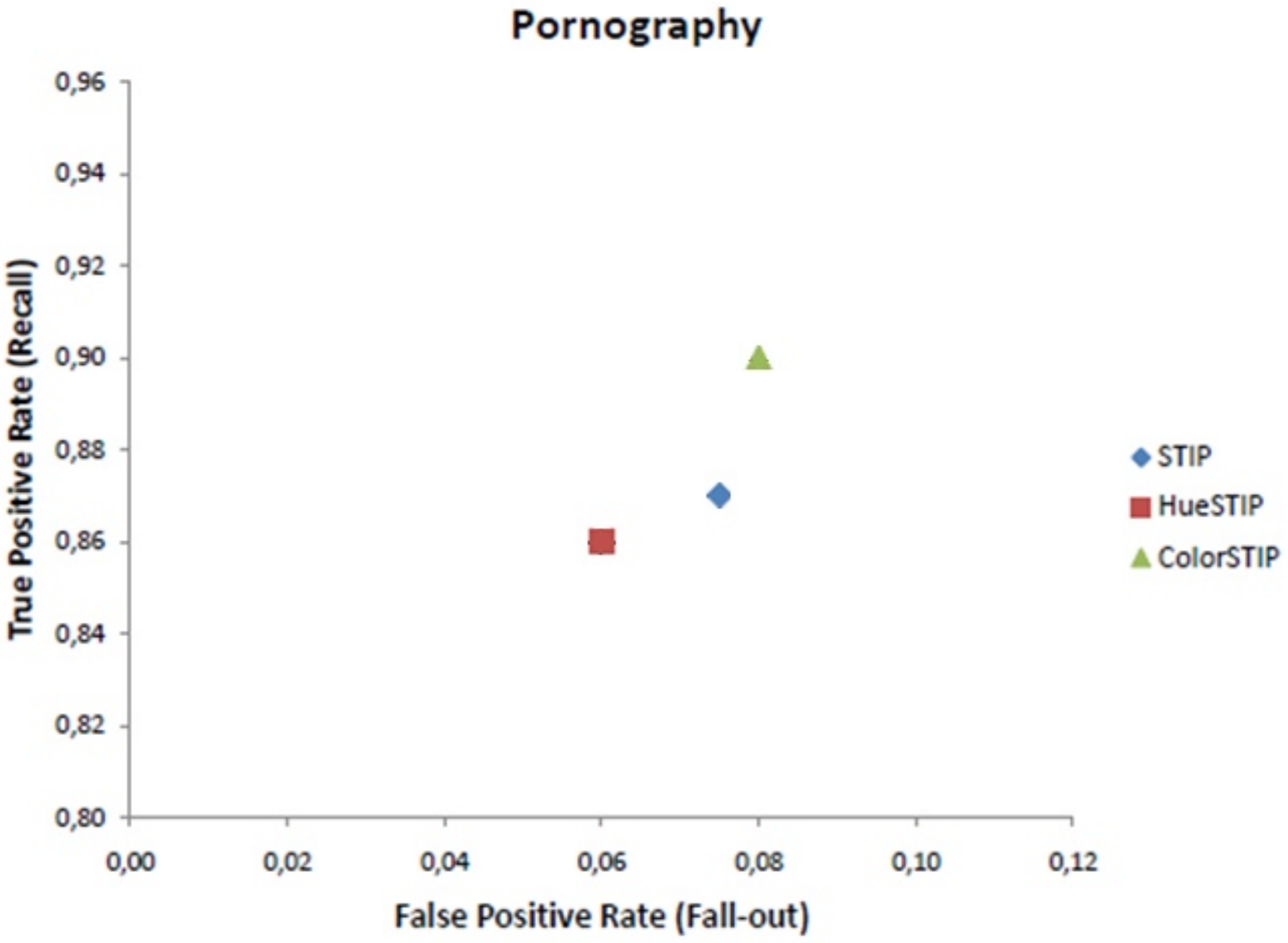


STIP		
Class	*Porn*	*NonPorn*
Porn	**.87**	.13
NonPorn	.075	**.925**

HUESTIP		
Class	*Porn*	*NonPorn*
Porn	**.86**	.14
NonPorn	.06	**.94**

COLORSTIP		
Class	*Porn*	*NonPorn*
Porn	**.90**	.10
NonPorn	.08	**.92**

Experimental Results



STIP		
Class	*Porn*	*NonPorn*
Porn	**.87**	.13
NonPorn	.075	**.925**

HUESTIP		
Class	*Porn*	*NonPorn*
Porn	**.86**	.14
NonPorn	.06	**.94**

COLORSTIP		
Class	*Porn*	*NonPorn*
Porn	**.90**	.10
NonPorn	.08	**.92**

Experimental Results

Violence Detection in Videos Using Spatio-Temporal Features

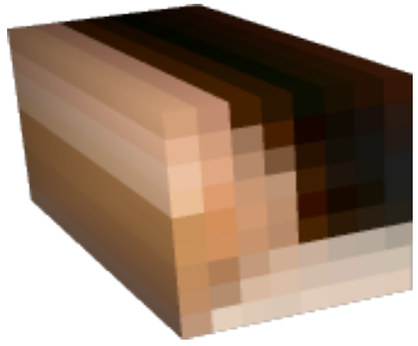
Fillipe D. M. de Souza[1], Guillermo C. Chávez[2]
Eduardo Valle Jr.[3] and Arnaldo de A. Araújo[1]

1 – NPDI/DCC/ICEx/UFMG
2 – ICEB/UFOP
3 – IC/UNICAMP

Introduction

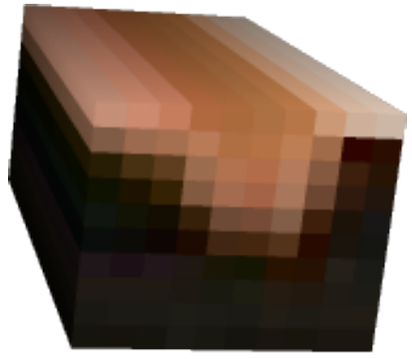
What's new in this work?

- Data representation
  - Use of local spatio-temporal features as representative patterns for the interest actions



Left Arm Moving

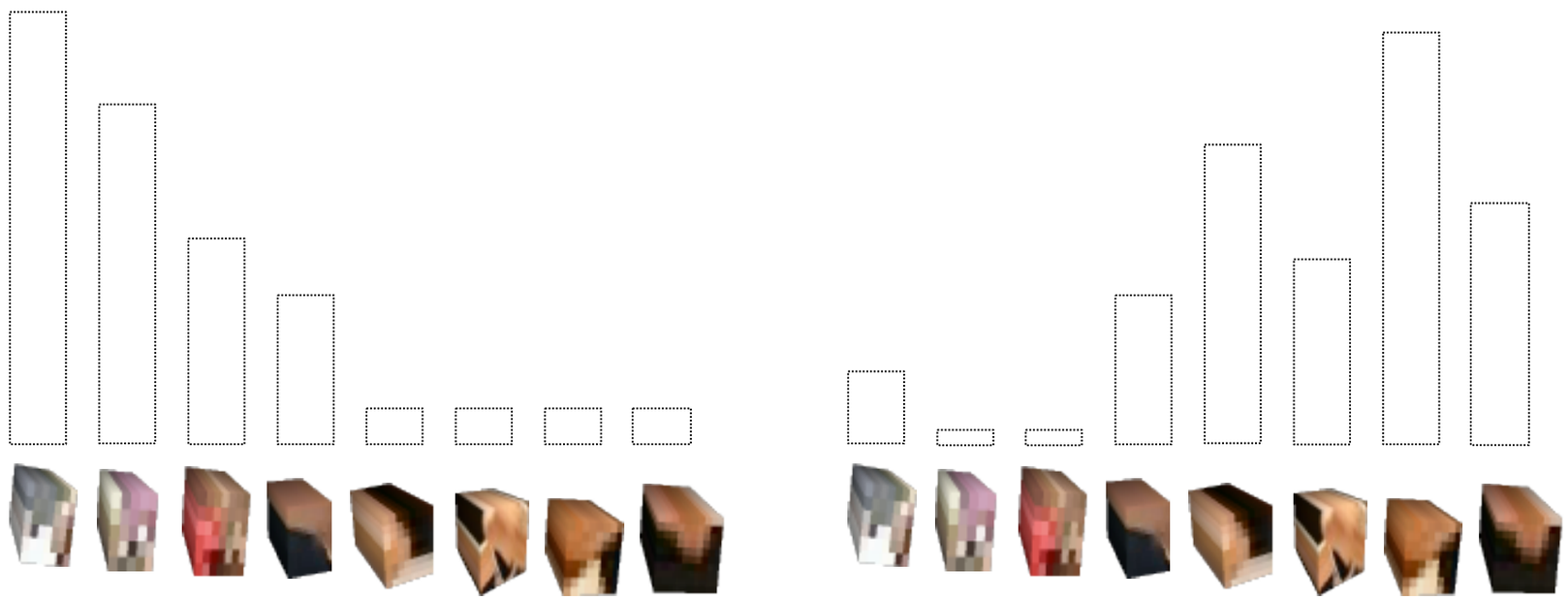
Knee Lifting



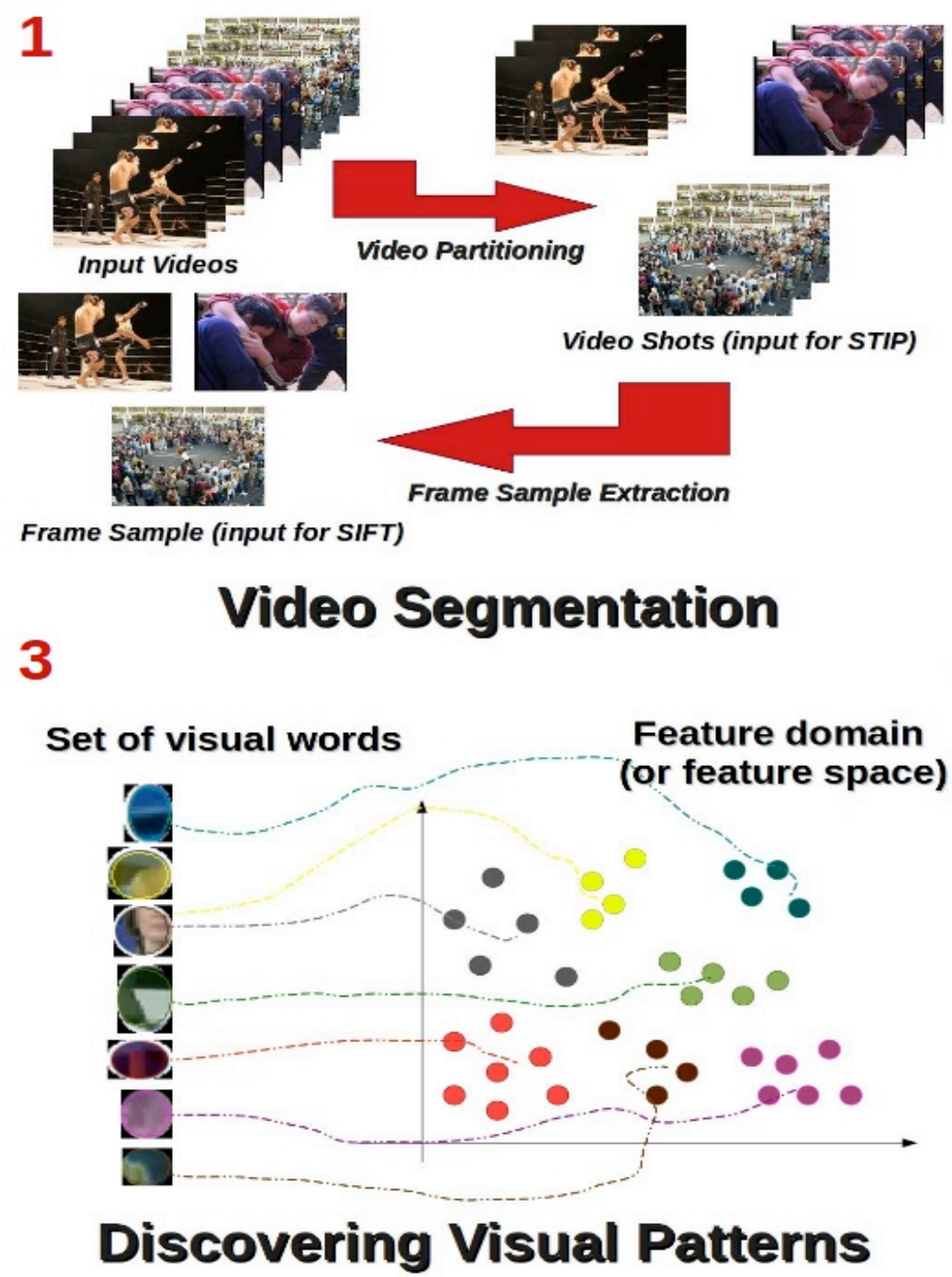
Elbow Up Moving

What's new in this work?

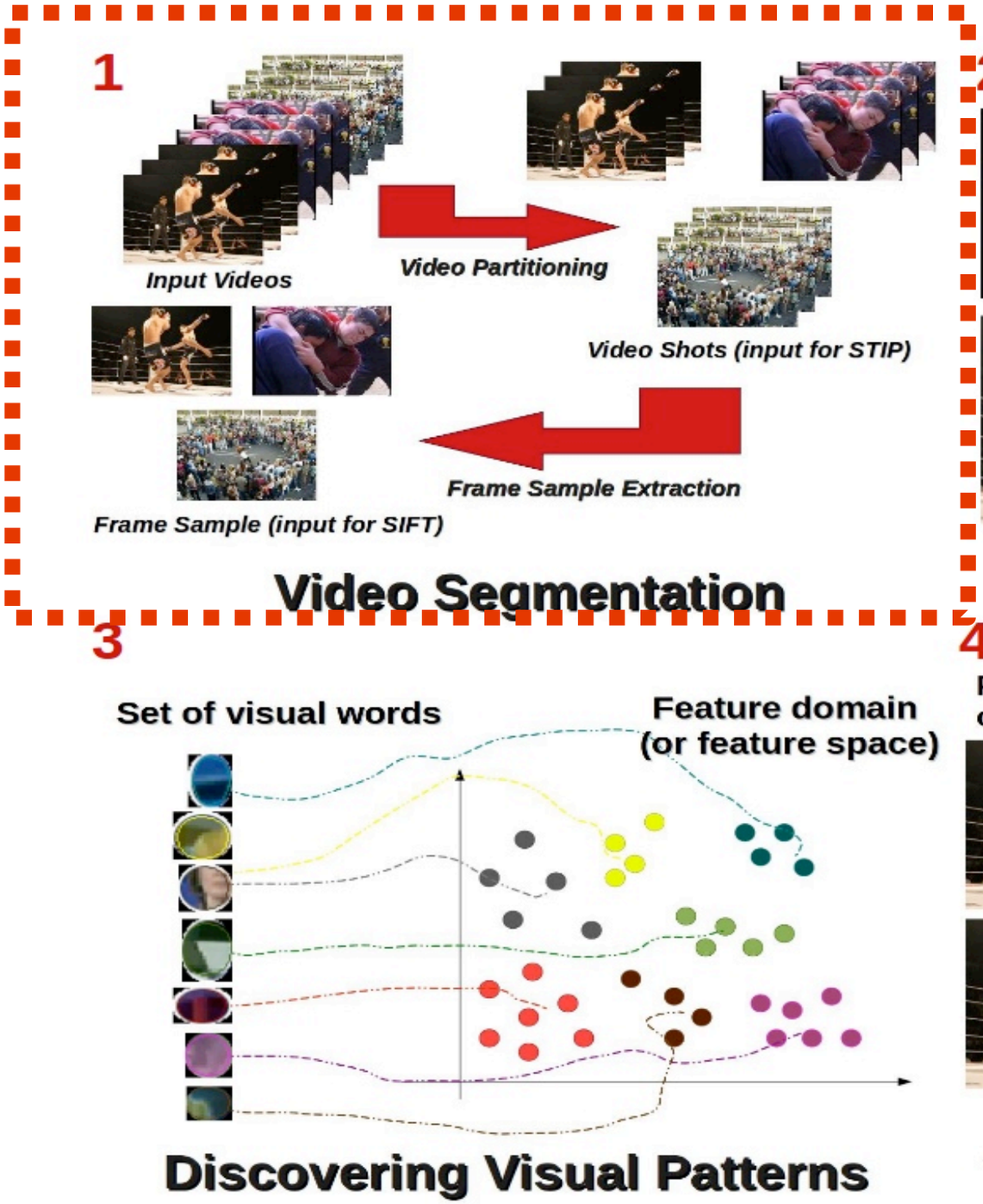
- Data representation:
  - Use of local spatio-temporal features as representative patterns for the interest actions;
  - Encoded as bags of visual words.



The Method Overview

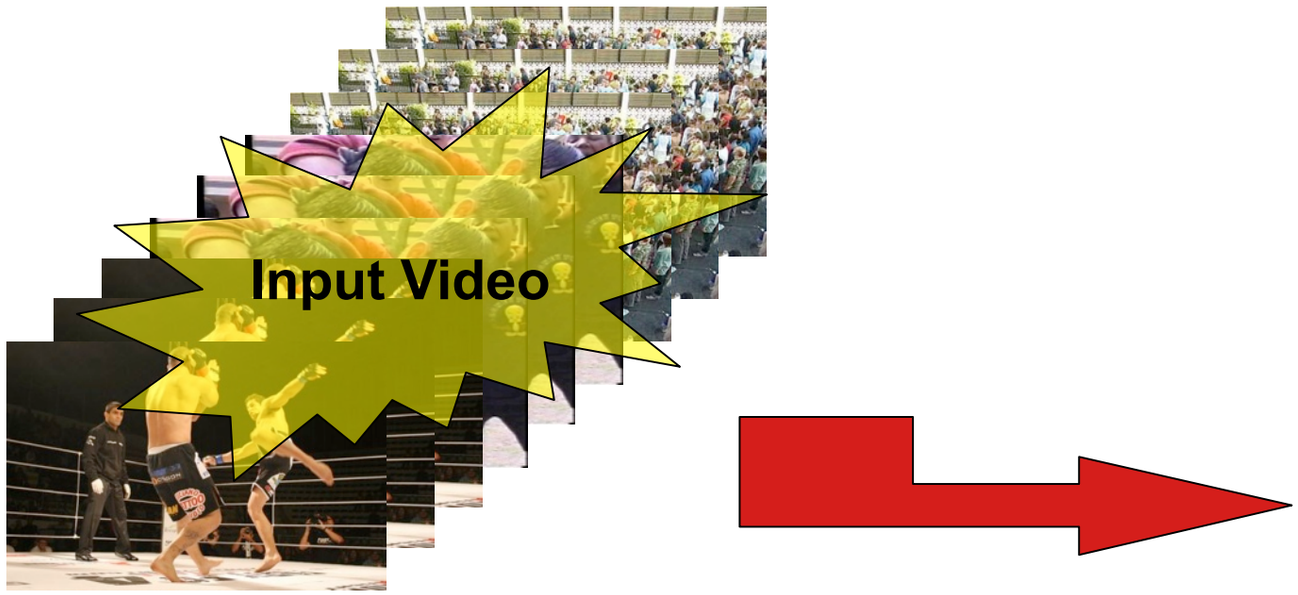


The Method Overview



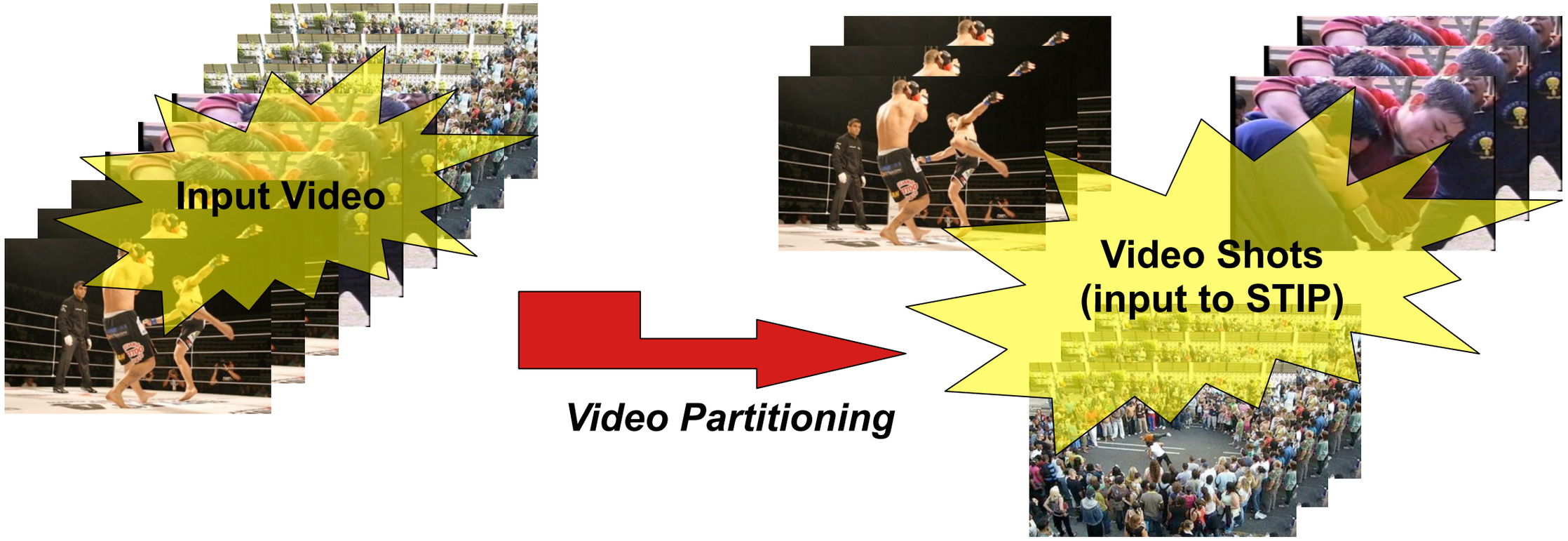
Preliminary Step

- Temporal Video Segmentation



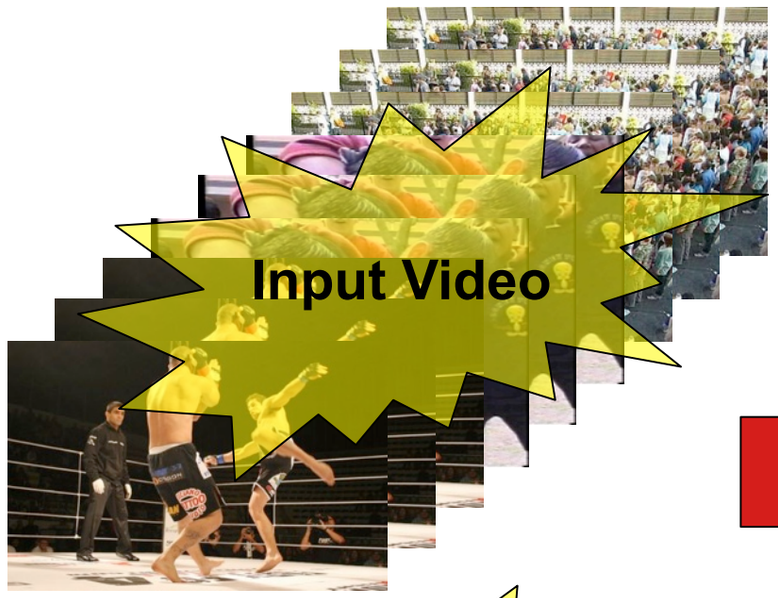
Preliminary Step

- Temporal Video Segmentation

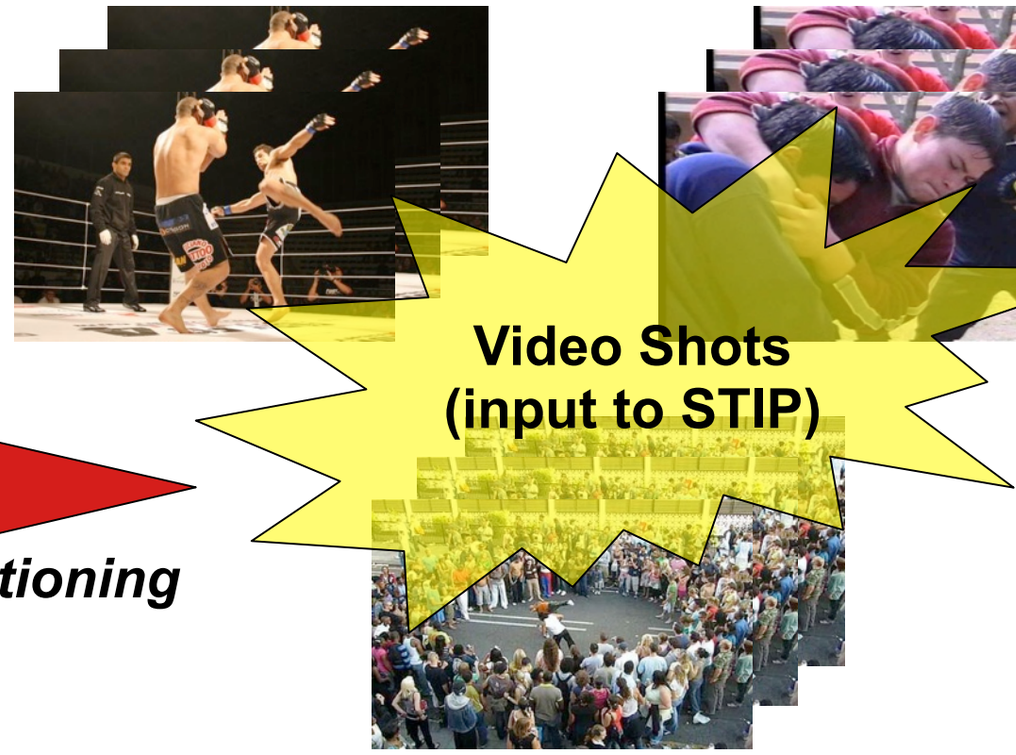


Preliminary Step

- Temporal Video Segmentation

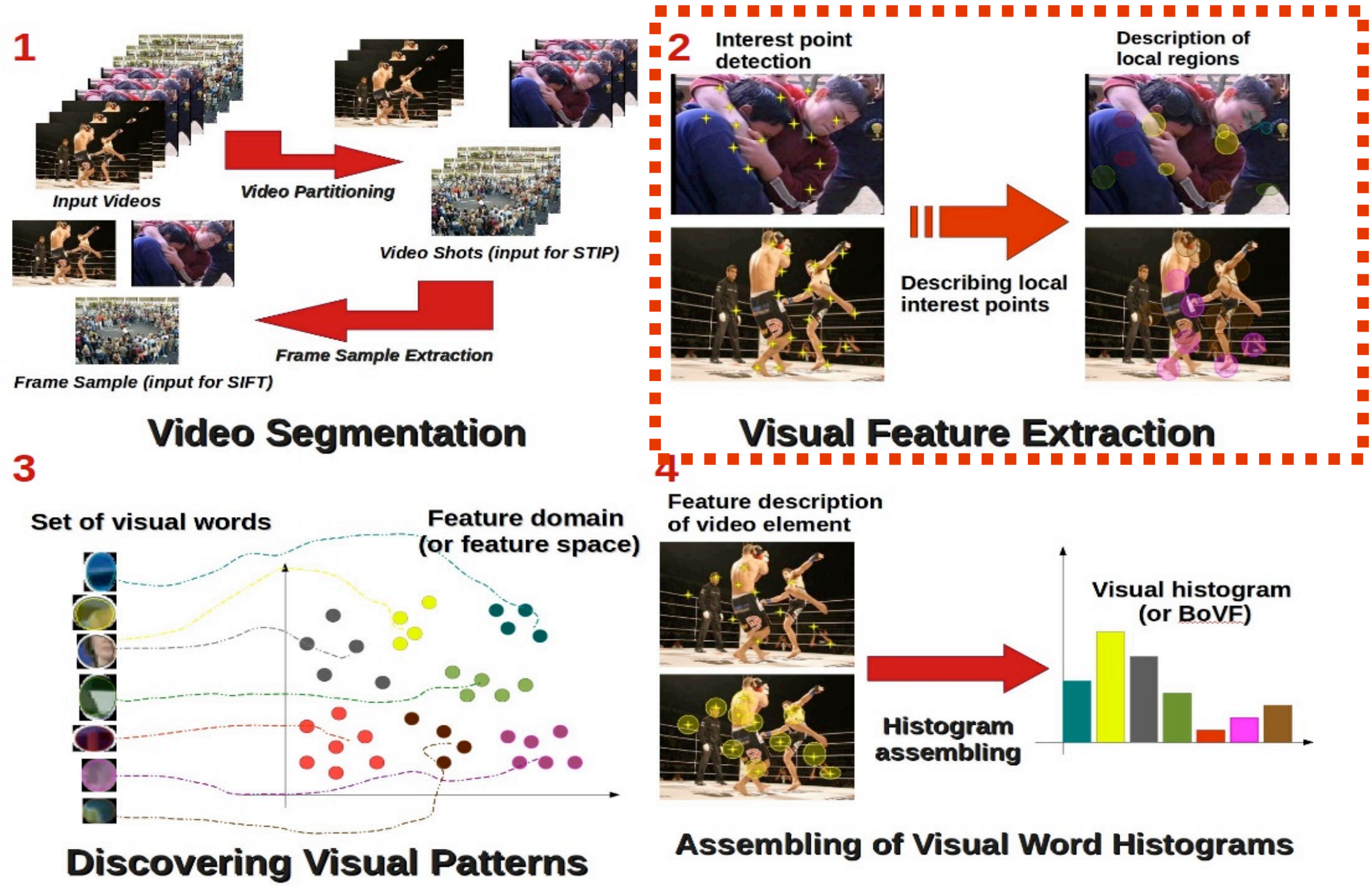


Video Partitioning



Frame Sample Extraction

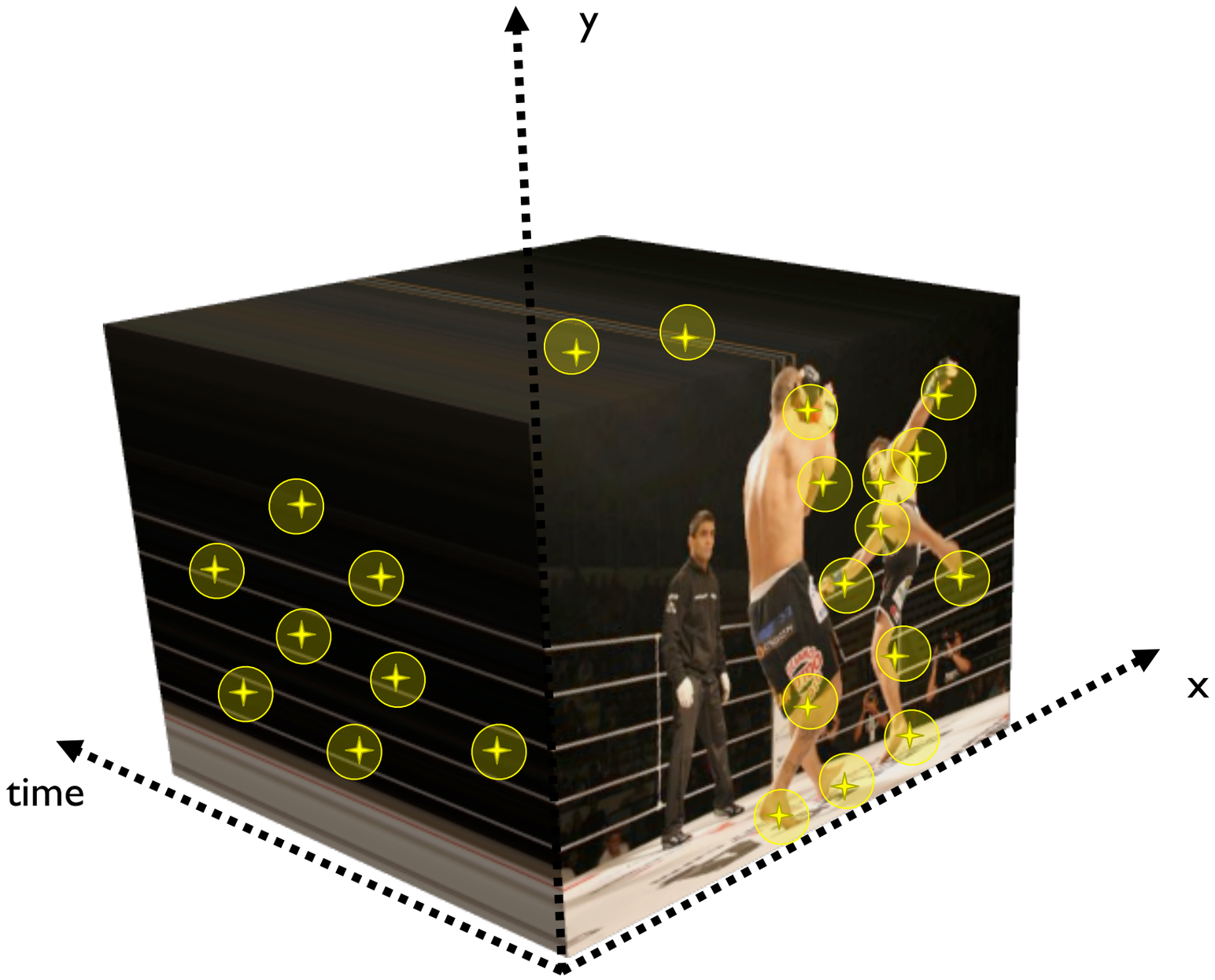
The Method Overview



RECOD

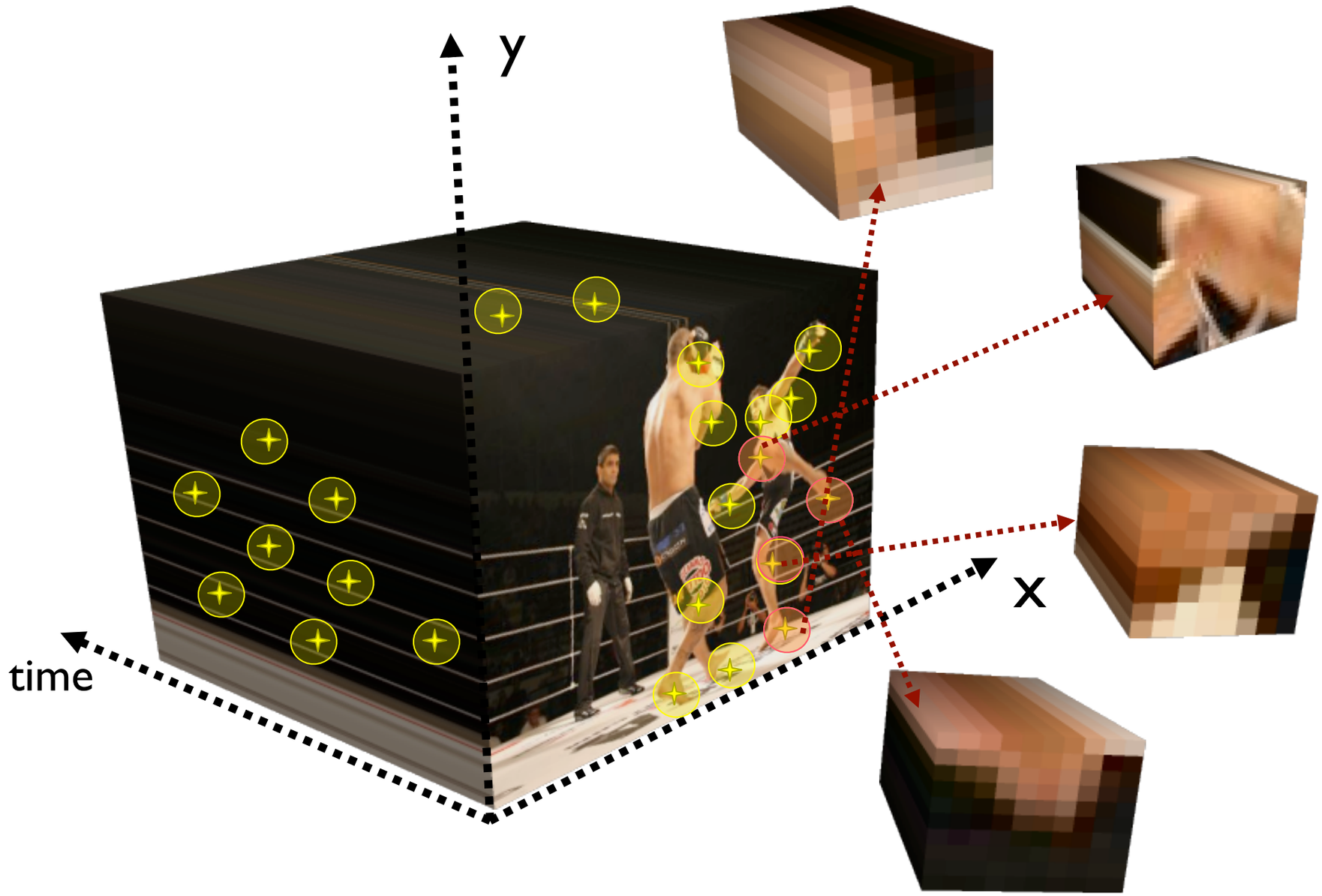
Local interest point detection in space-time

- Spatio-temporal domain features (Laptev 2005)



Local interest point detection in space-time

- Spatio-temporal domain features (Laptev 2005)



Data representation by visual words

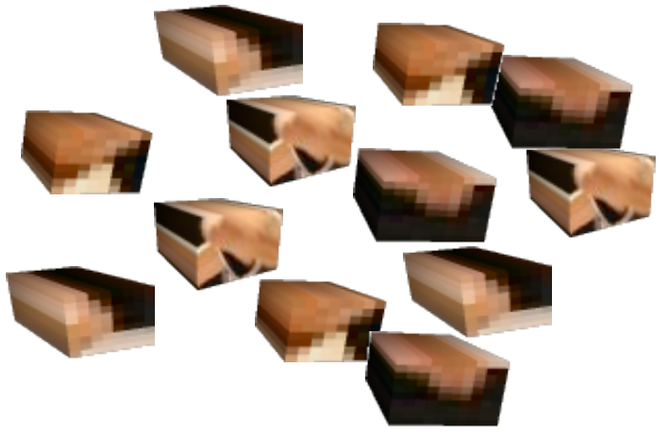
Video showing an agitated crowd

Video showing two people fighting

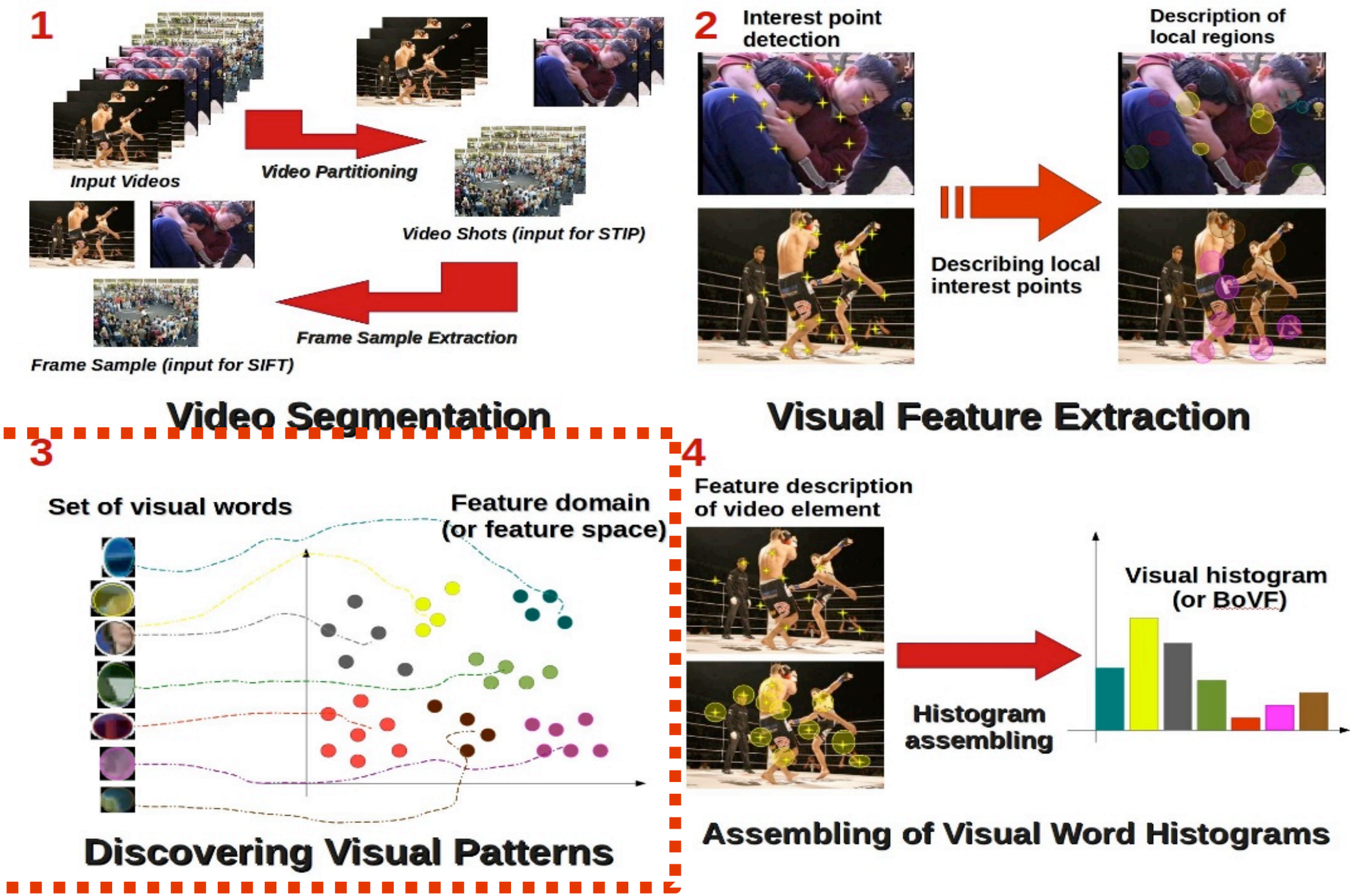
Data representation by visual words

Video showing an agitated crowd

Video showing two people fighting



The Method Overview



RECOD

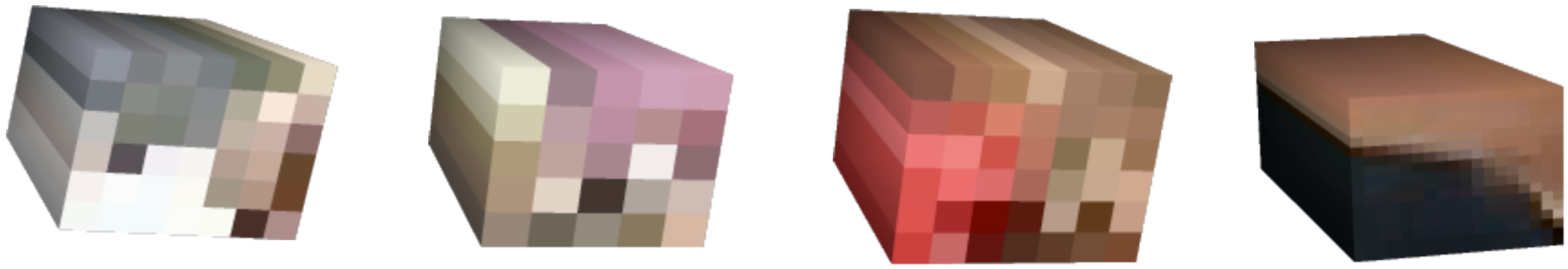
Feature Clustering

- Discovering visual patterns (so-called visual words)

Data representation by visual words

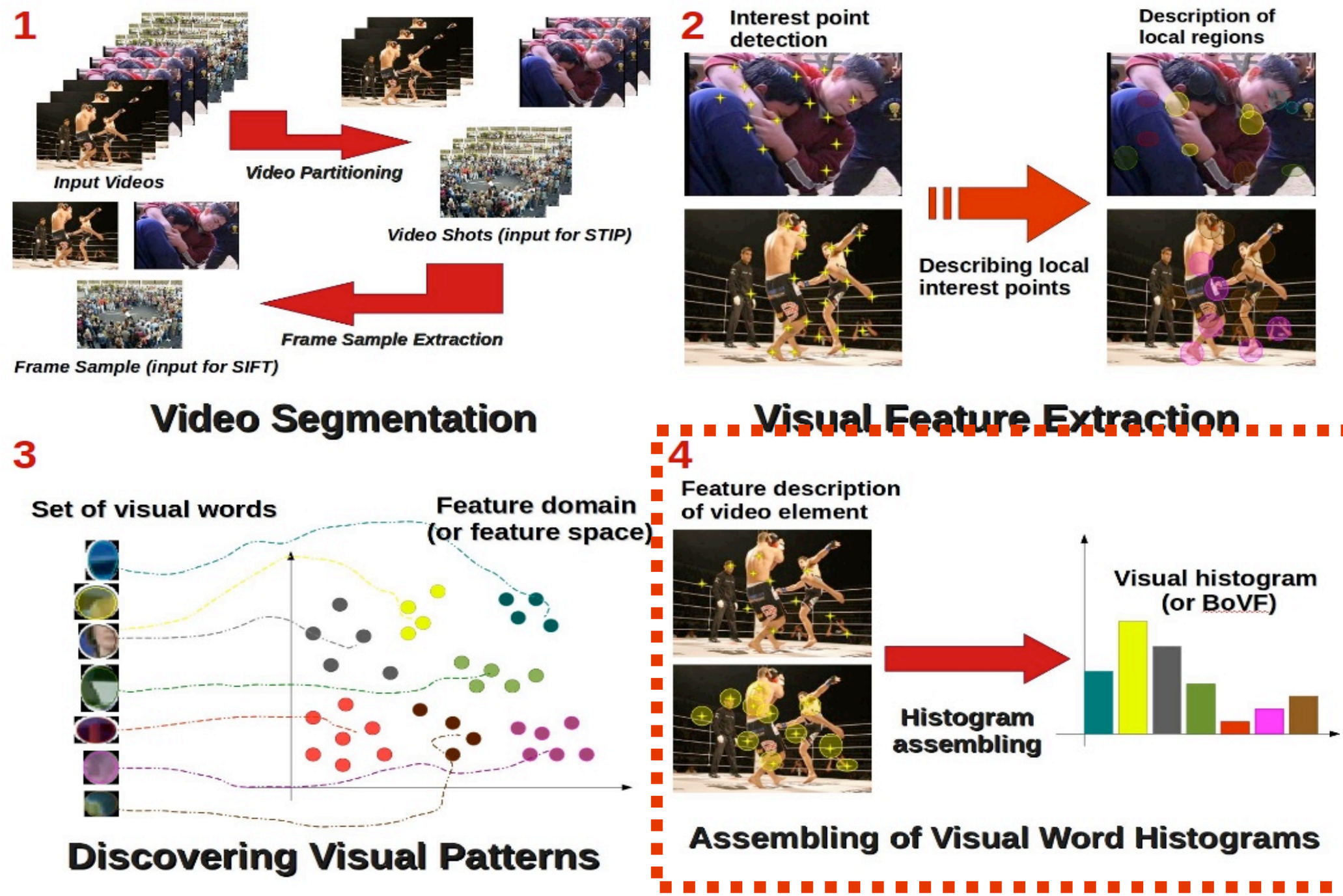
- Visual codebook formed by 8 visual 'words':

4 spatio-temporal descriptions (visual words) of non-violent content



4 spatio-temporal descriptions (visual words) of violent content

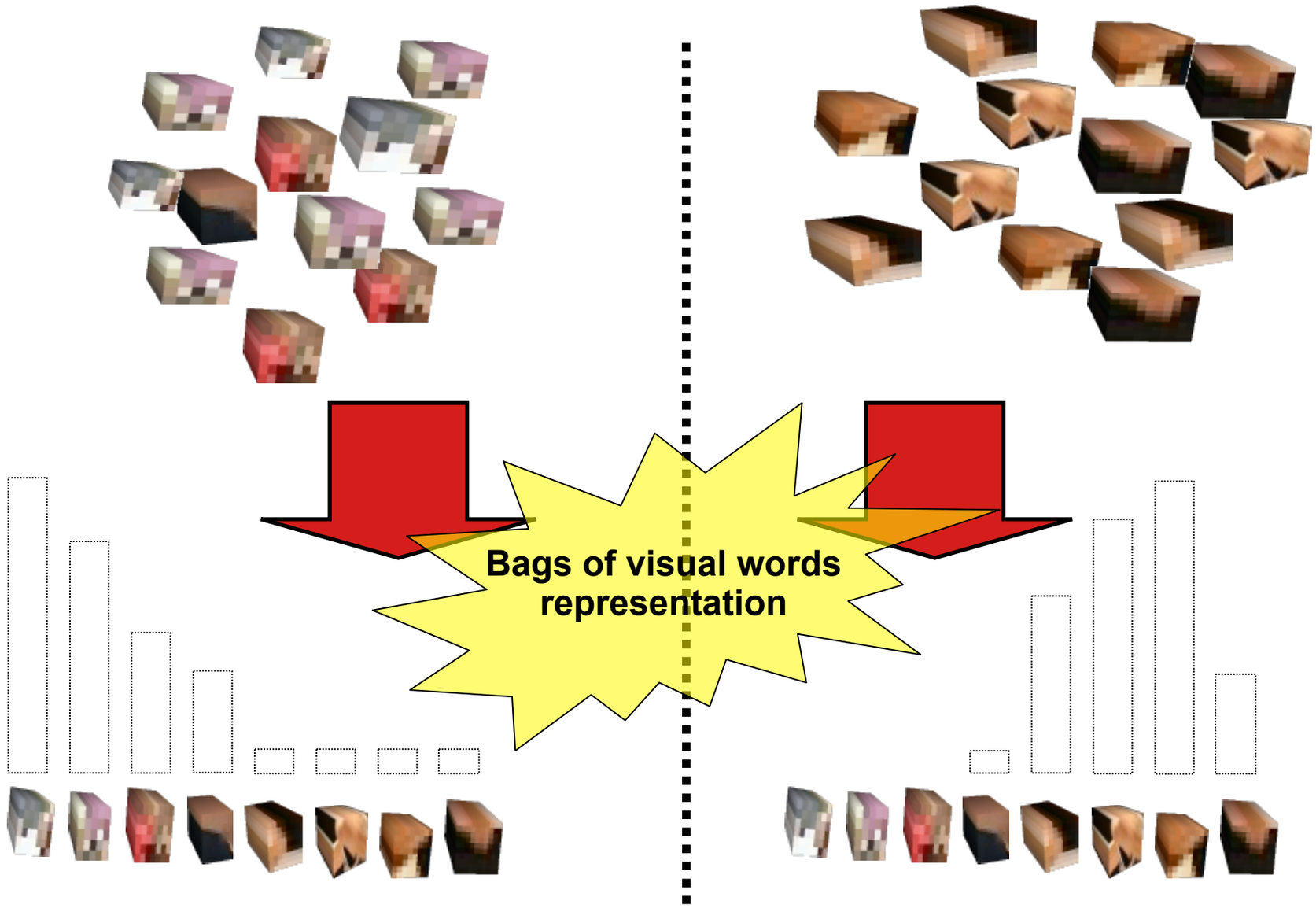
The Method Overview



RECOD

Data representation by visual words

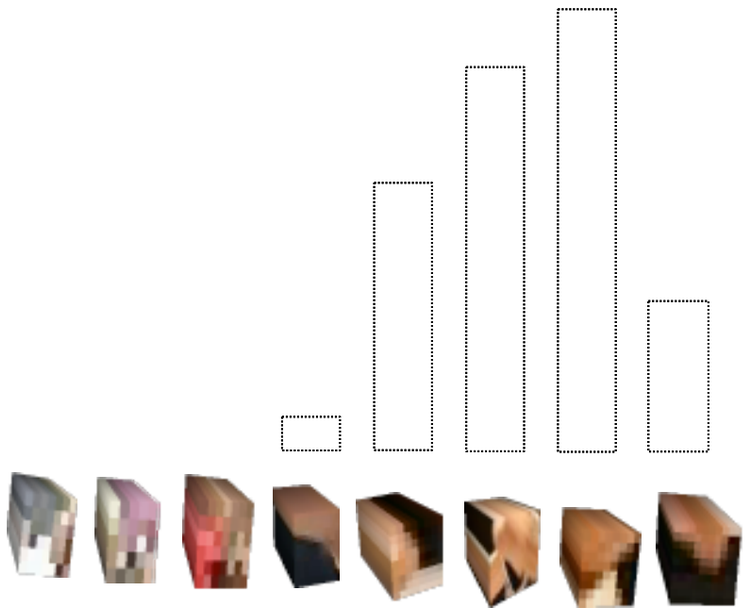
Sets of spatio-temporal features computed for each video



Data representation by visual words

Video showing an agitated crowd

Video showing two people fighting



Learning a classifier to detect the interest event

- Support Vector Machine (SVM, libSVM, Ref.)
- Binary problem (only 2 classes);

Learning a classifier to detect the interest event

- Support Vector Machine (SVM, libSVM, Ref.)
  - Linear kernel (simple, fast, good results);
  - LibSVM default parameters;
  - Binary problem (only 2 classes);
  - Good capability of generalization;
  - **Input:** *annotated bags of visual words*;
  - **Output:** *classification model*.

Non-violent Dataset Sample

Violent Dataset Sample

Experimental Results

Codebook 100	***SIFT***	
(%)	Violent	Non-Violent
Violent	**80.09**	**19.91**
Non-Violent	**14.65**	**85.35**

Codebook 100	***STIP***	
(%)	Violent	Non-Violent
Violent	**99.54**	**0.46**
Non-Violent	**0.0**	**100.0**

Comments

- We proposed a method with:
  - Local spatio-temporal features
  - Visual words representation
  - Supervised learning with SVM
- Comparison between spatio-temporal and spatial features

Conclusions

Conclusions

- We are just starting to solve this problem
- Many challenges still in existence
- Efficiency is also an important issue


PhotoDNA™
The Next Chapter in Protecting Children Online

References

References

- Valle, E., Avila, S., da Luz Jr., A., de Souza, F., Coelho, M., and de A. Araú jo, A. (2012). Detecting undesirable content in video social networks (to appear). In Brazilian Symposium on Information and Computer System Security.

- Zuo, H., Hu, W., and Wu, O. (2010). Patch-based skin color detection and its appli- cation to pornography image filtering. In International Conference on World Wide Web (WWW), pages 1227--1228.

References

- Deselaers, T., Pimenidis, L., and Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. In International Conference on Pattern Recognition (ICPR), pages 1–4.

- Endeshaw, T., Garcia, J., and Jakobsson, A. (2008). Fast classification of indecent video by low complexity repetitive motion detection. In IEEE Applied Imagery Pattern Recognition Workshop, pages 1--7.

- Fleck, M., Forsyth, D. A., and Bregler, C. (1996). Finding naked people. In European Conference on Computer Vision (ECCV), pages 593--602.

- Forsyth, D. A. and Fleck, M. M. (1996). Identifying nude pictures. In IEEE Workshop on Applications of Computer Vision (WACV), pages 103--108.

- Forsyth, D. A. and Fleck, M. M. (1997). Body plans. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 678--683.

References

- Forsyth, D. A. and Fleck, M. M. (1999). Automatic detection of human nudes. Inter- national Journal on Computer Vision (IJCV), 32(1):63--77.

- Jansohn, C., Ulges, A., and Breuel, T. M. (2009). Detecting pornographic video con- tent by combining image features with motion information. In ACM International Conference on Multimedia, pages 601--604.

- Jones, M. J. and Rehg, J. M. (2002). Statistical color models with application to skin detection. International Journal of Computer Vision (IJCV), 46(1):81--96.

- Lopes, A., Avila, S., Peixoto, A., Oliveira, R., Coelho, M., and de A. Araú jo, A. (2009a). Nude detection in video using bag-of-visual-features. In Brazilian Sympo- sium on Computer Graphics and Image (SIBGRAPI), pages 224--231.

- Lopes, A., Avila, S., Peixoto, A., Oliveira, R., and de A. Araú jo, A. (2009b). A bag- of-features approach based on hue-sift descriptor for nude detection. In European Signal Processing Conference (EUSIPCO), pages 1552--1556.

References

- Rowley, H. A., Jing, Y., and Baluja, S. (2006). Large scale image-based adult-content filtering. In International Conference on Computer Vision Theory and Applications (VISAPP), pages 290–296.
- Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to ob- ject matching in videos. In International Conference on Computer Vision (ICCV), volume 2.
- Steel, C. (2012). The mask-sift cascading classifier for pornography detection. In World Congress on Internet Security (WorldCIS), pages 139--142.
- Tong, X., Duan, L., Xu, C., Tian, Q., Hanqing, L., Wang, J., , and Jin, J. (2005). Peri- odicity detection of local motion. In IEEE International Conference on Multimedia and Expo (ICME), pages 650--653.
- Ulges, A. and Stahl, A. (2011). Automatic detection of child pornography using color visual words. In International Conference on Multimedia Retrieval (ICME), pages 1–6.

Obrigado!

Thank you!