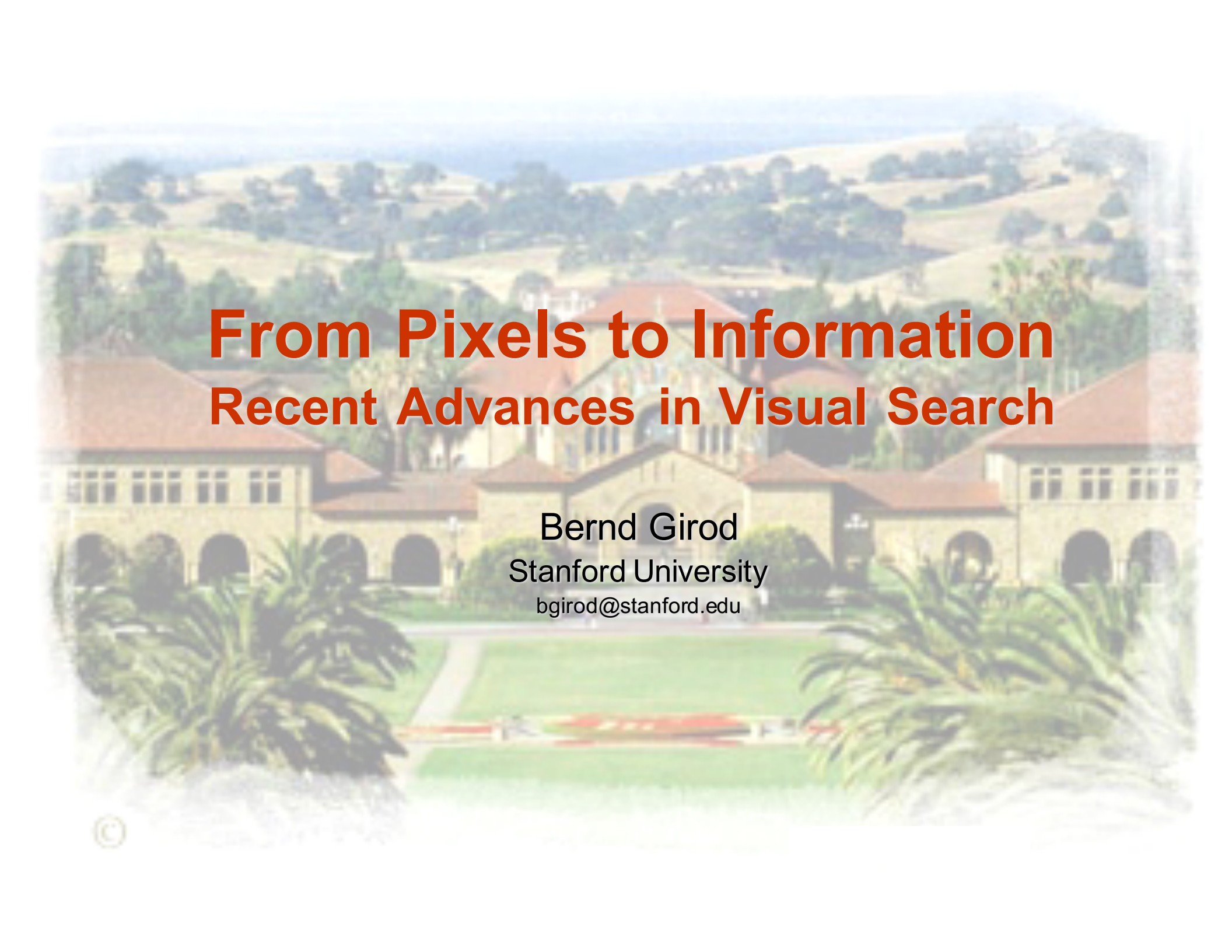


From Pixels to Information
Recent Advances in Visual Search

Bernd Girod
Stanford University
bgirod@stanford.edu

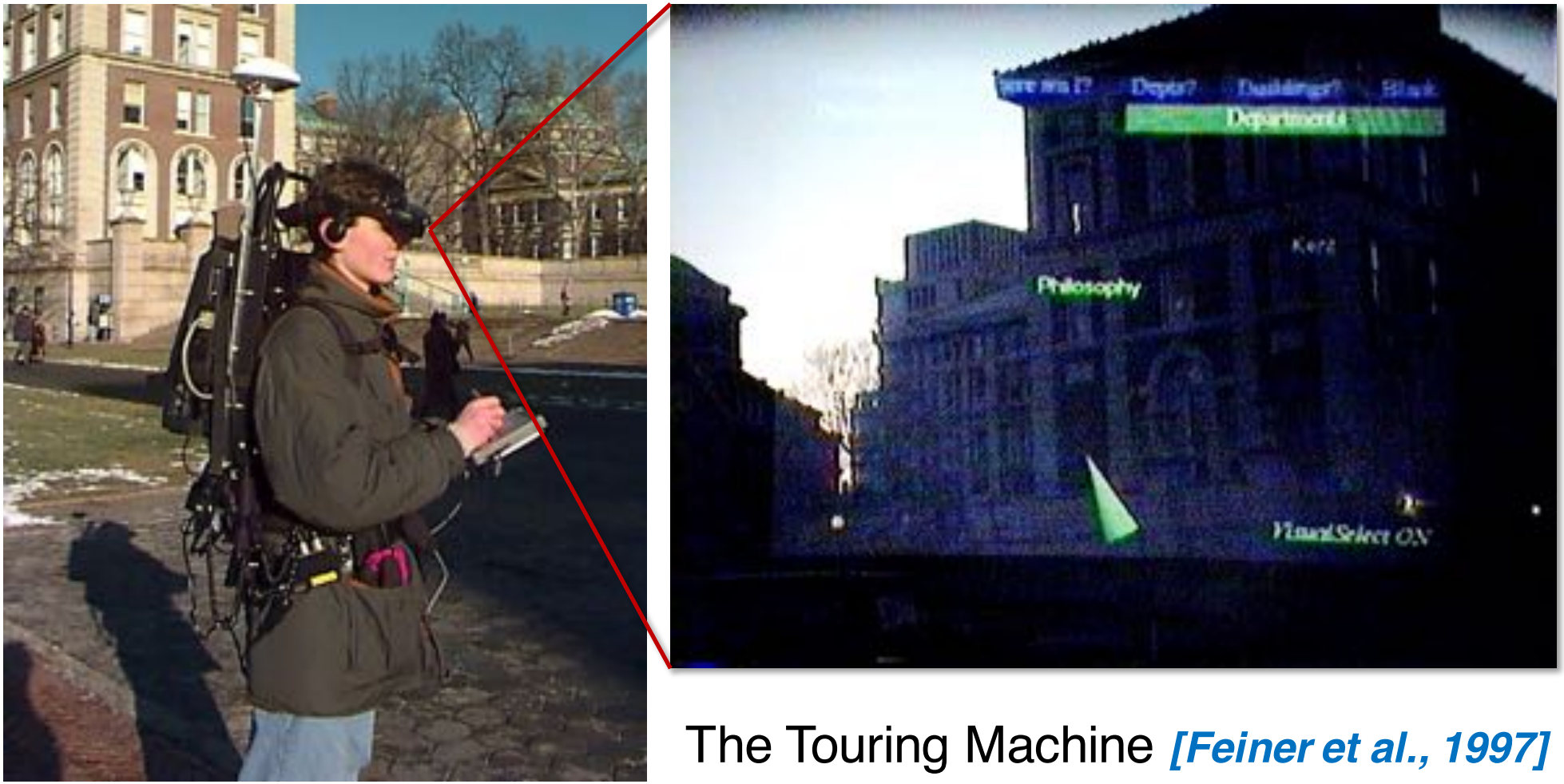
Augmented Reality

Augmented Reality

Future: Smart Contact Lenses

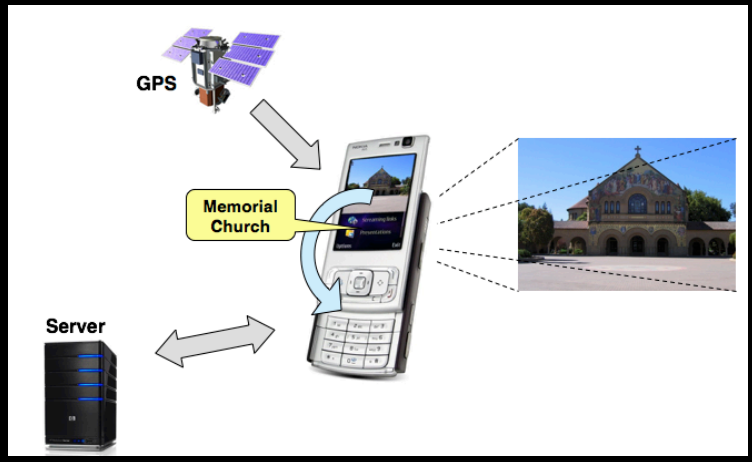
Sight: Contact Lenses with Augmented Reality
[E. May-raz and D. Lazo, 2012]

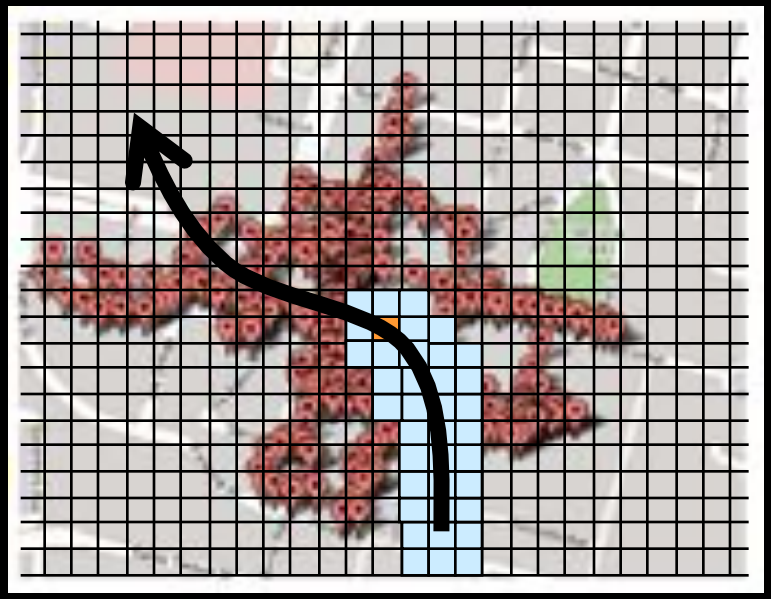
Recognizing What the User Sees



The Touring Machine *[Feiner et al., 1997]*

Stanford Landmark Recognition (2007)



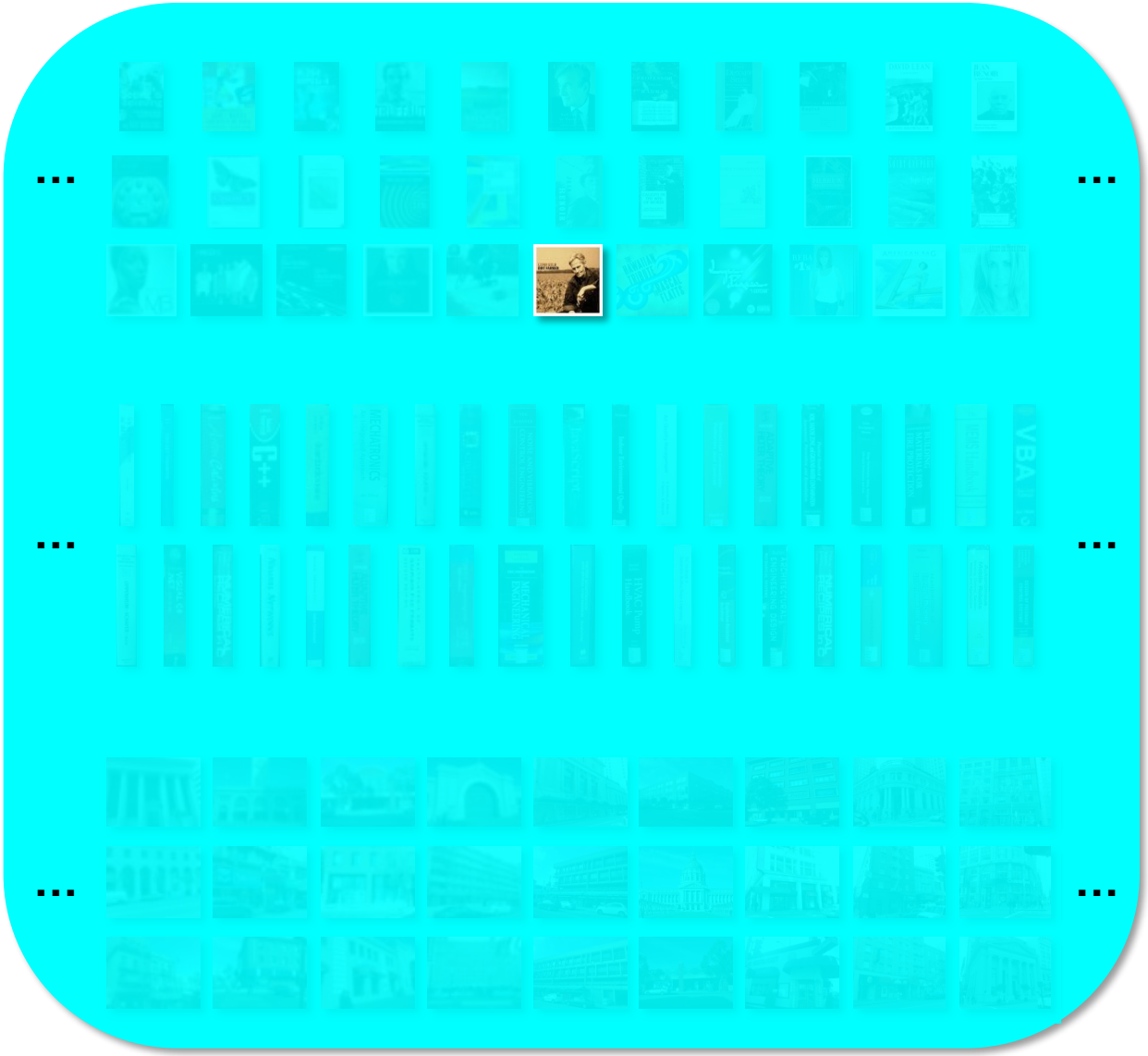


G. Takacs et al., ACM MIR 2008.

Recognizing Objects


Thomas Hill
Jane Stanford
Leland Stanford, Jr.
Who's who?

Image-based Retrieval



Outline

- Review: Computer vision for image-based retrieval

  Invariant local image features (SIFT); matching feature descriptors

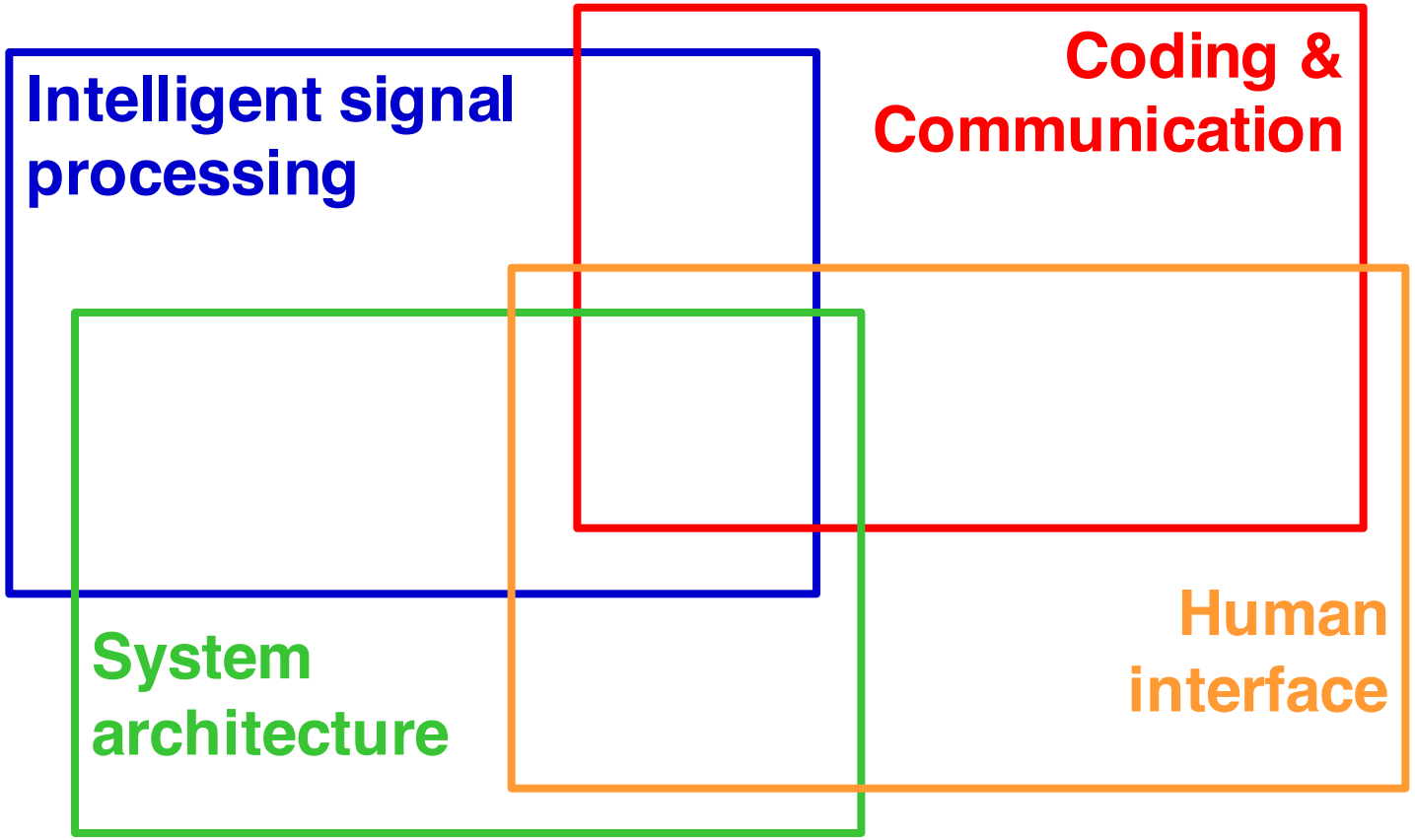
- MPEG CDVS Standard: Compact Descriptors for Visual Search

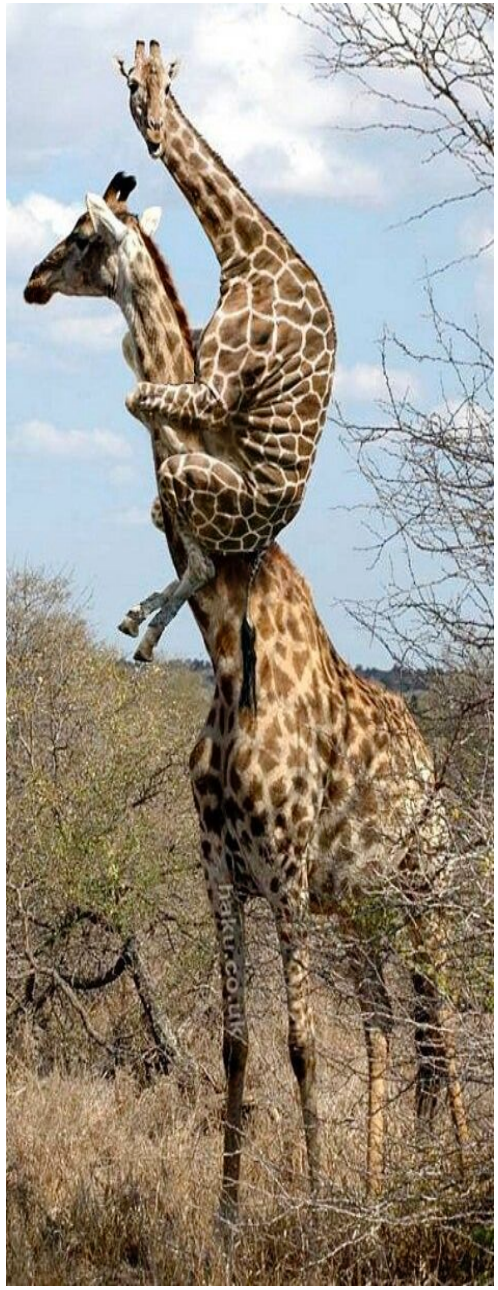
  CDVS framework & pipeline; Fisher vectors as global descriptors

- Current research directions

  Query-by-image video retrieval; interframe compression of local and global descriptors

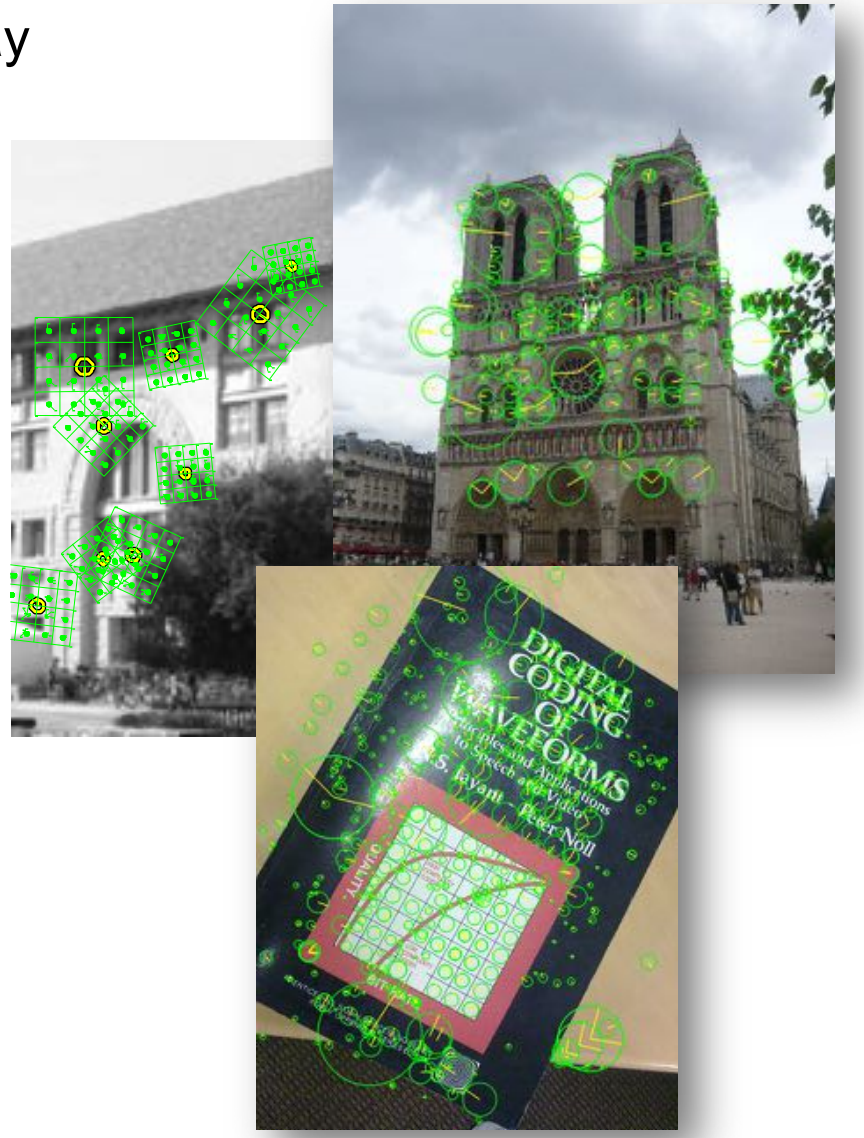
Standing on the Shoulders of …



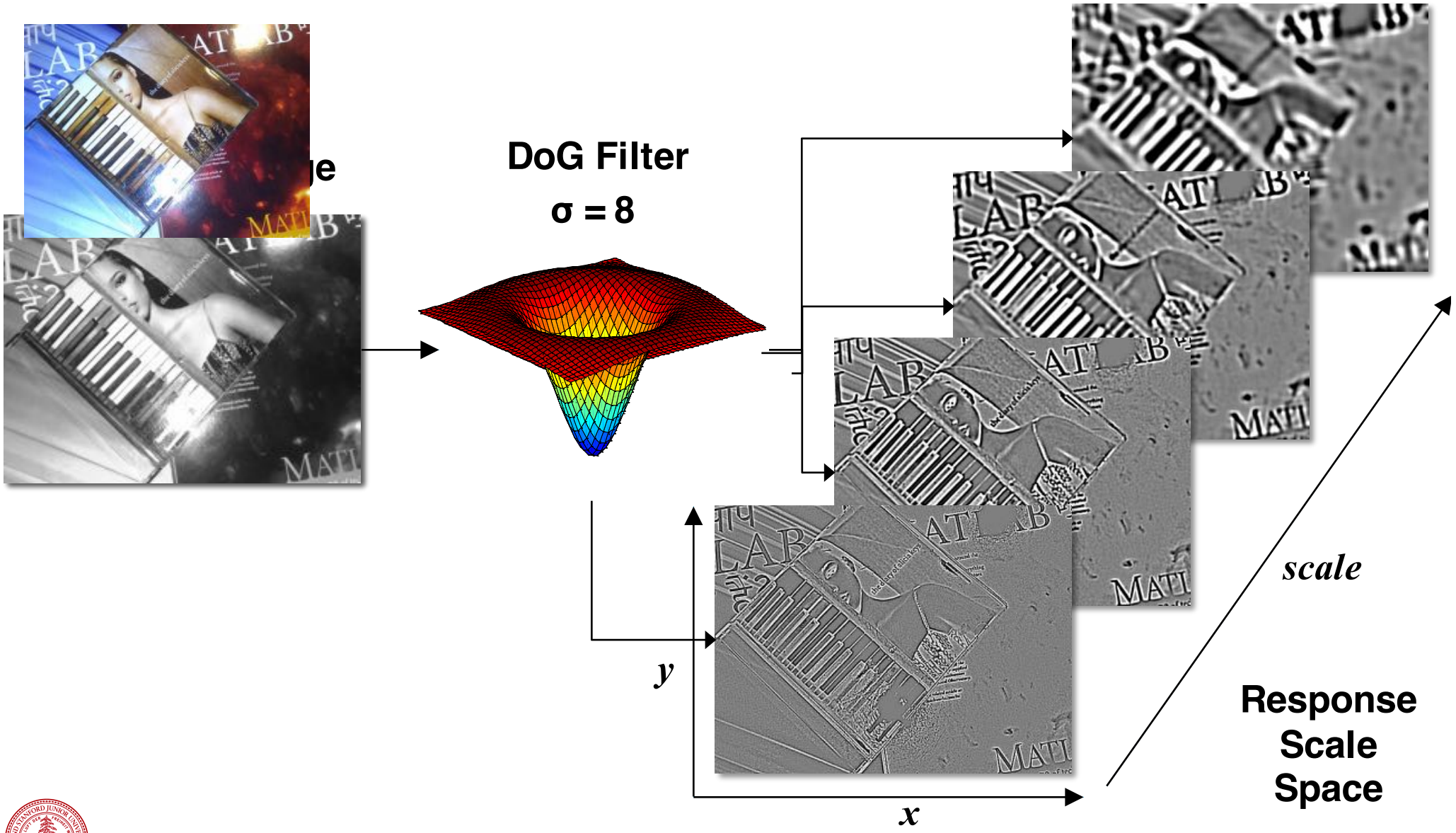


Local Image Features

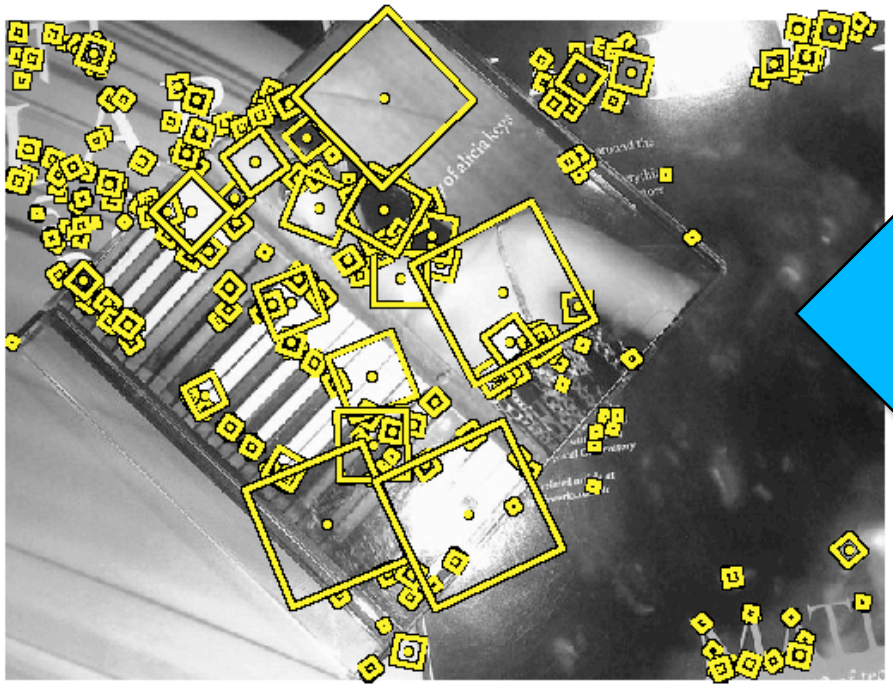
- Vectors that describe local patterns in a way that is both **distinctive** and **invariant** to
  - Brightness changes
  - Contrast changes
  - Shift in x,y
  - Scale change
  - Rotation
  - (Affine distortion)
- Scale Invariant Feature Transform (SIFT)
  [Lowe, 1999, 2004]



Local Features: Keypoint Detection

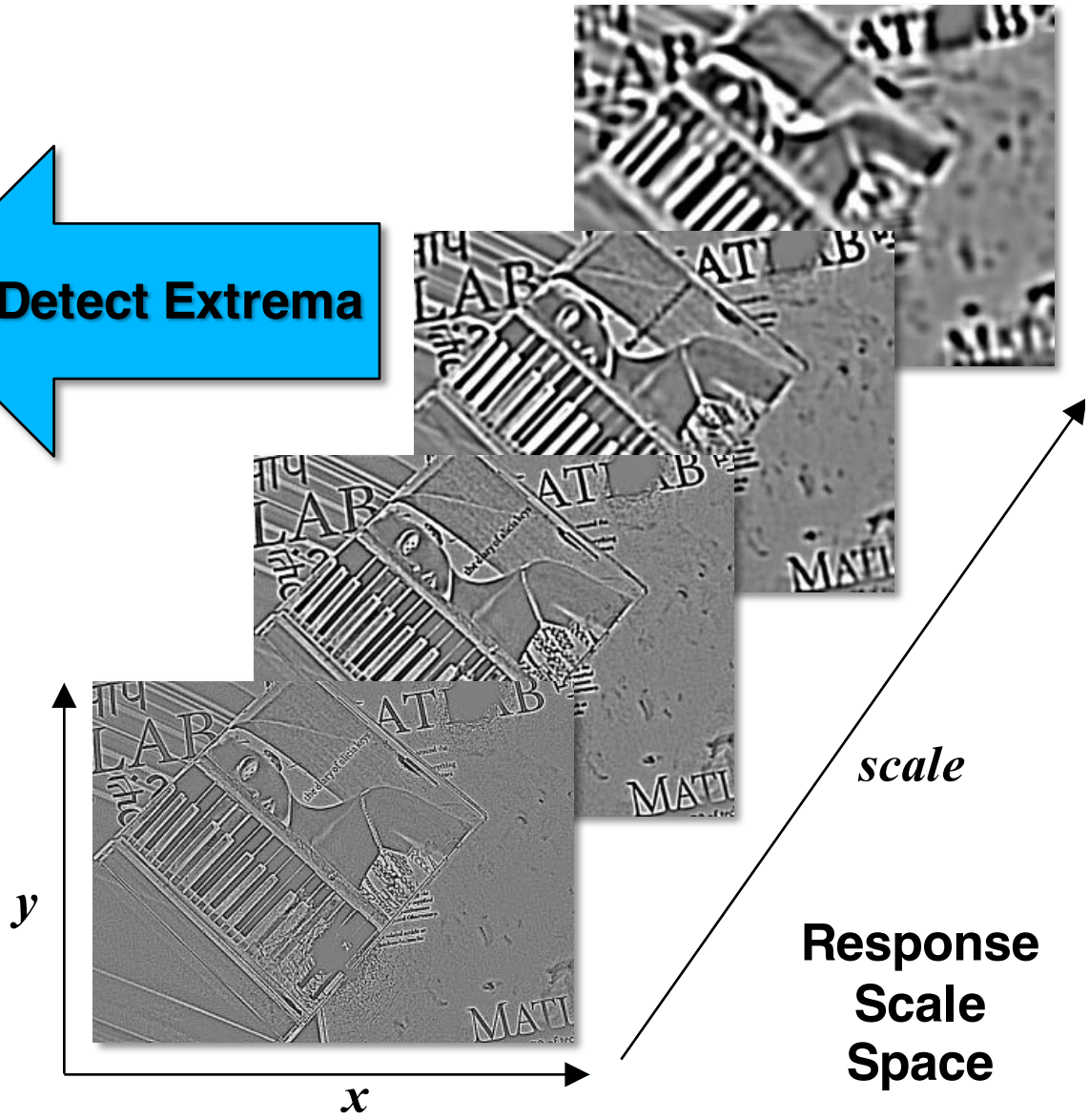


Local Features: Keypoint Detection

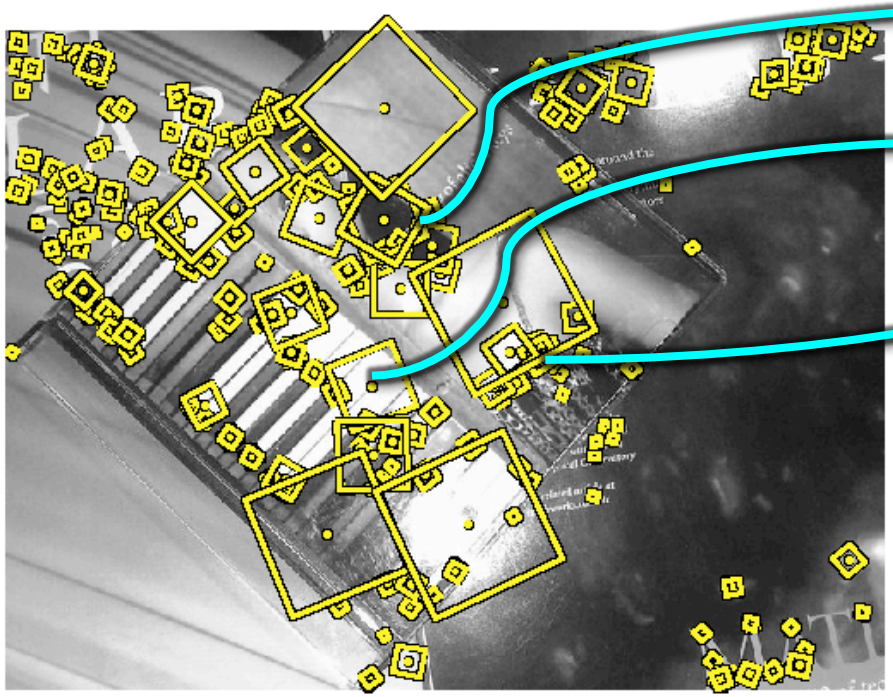


Oriented Feature Keypoints

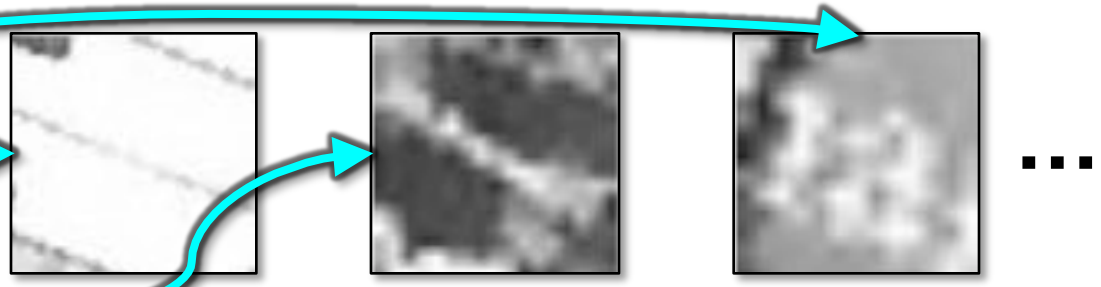
Detect Extrema



Local Features: Descriptor Computation



Oriented Feature Keypoints



Canonical Image Patches

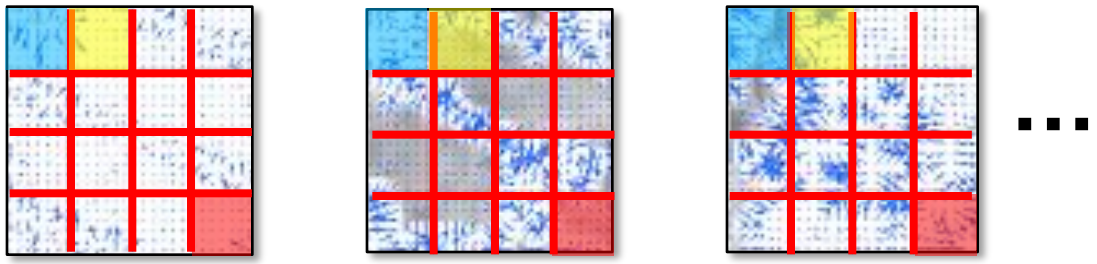
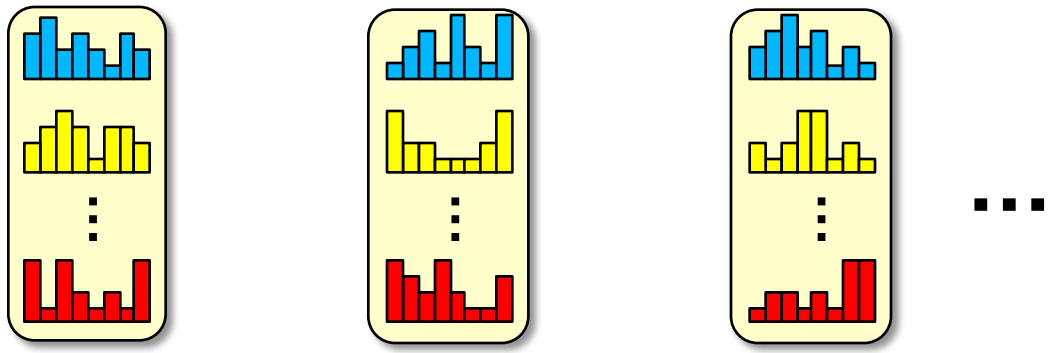
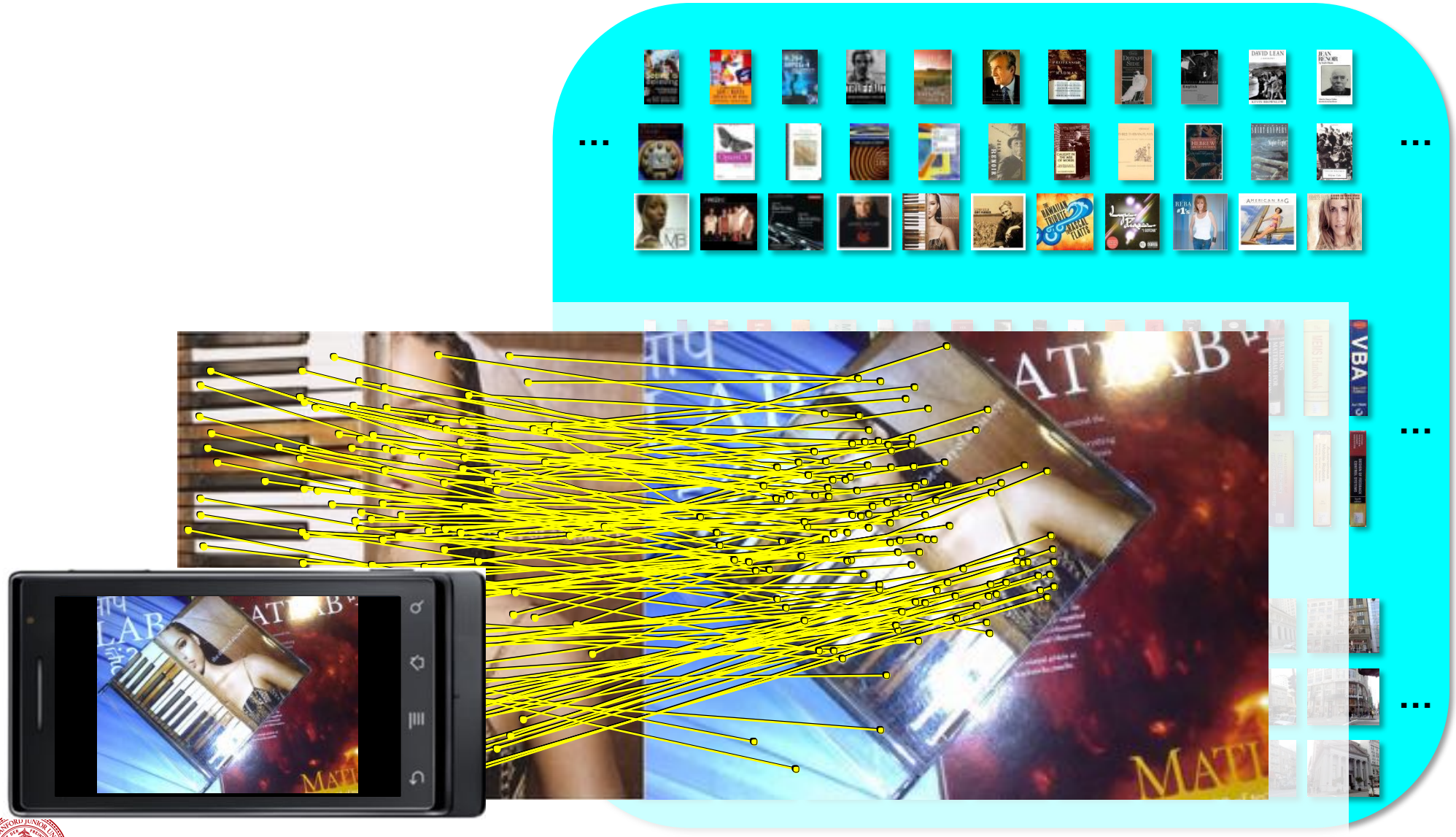


Image Gradients

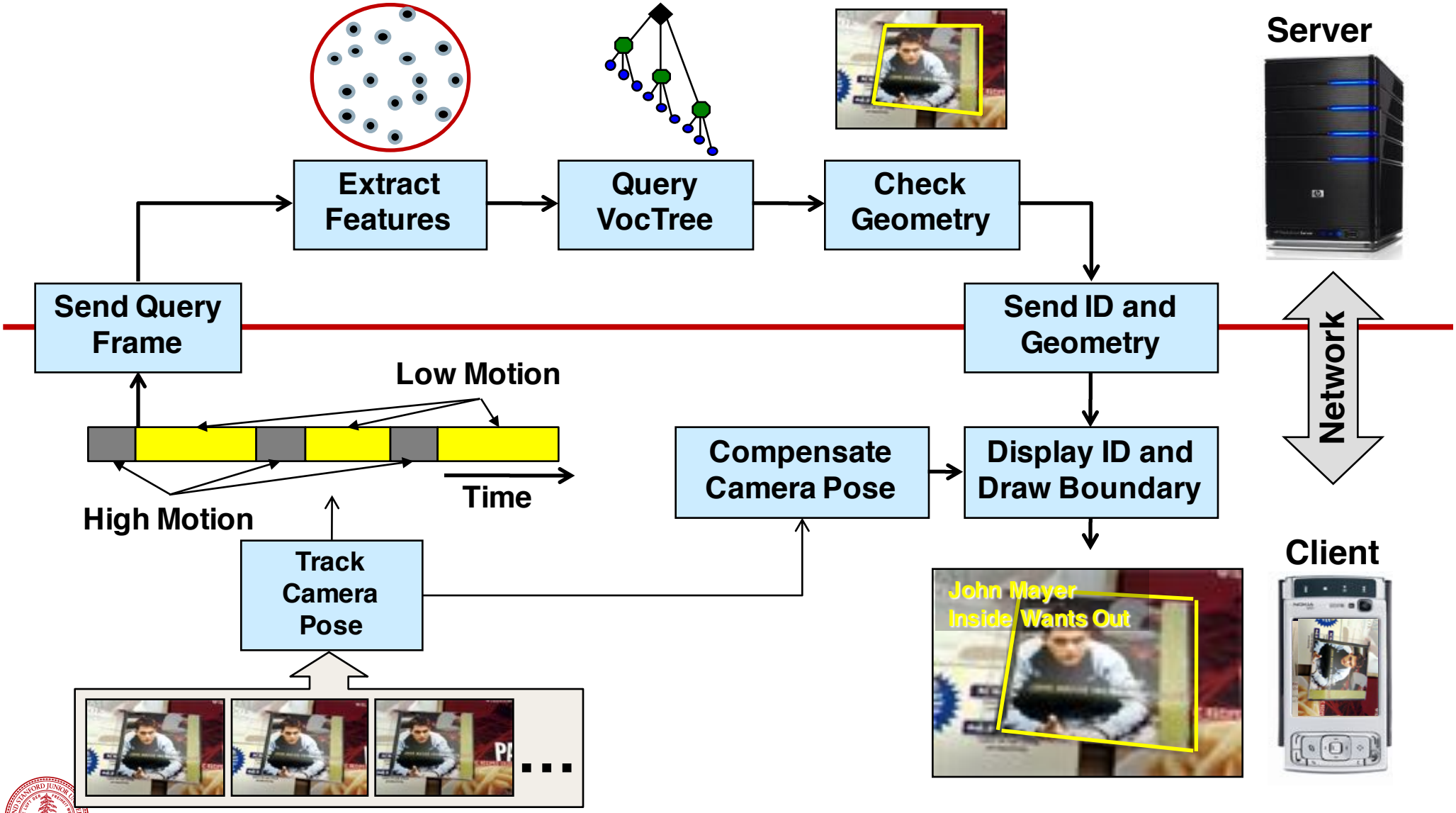


Gradient Orientation Histograms

Matching Local Feature Descriptors



Mobile Augmented Reality



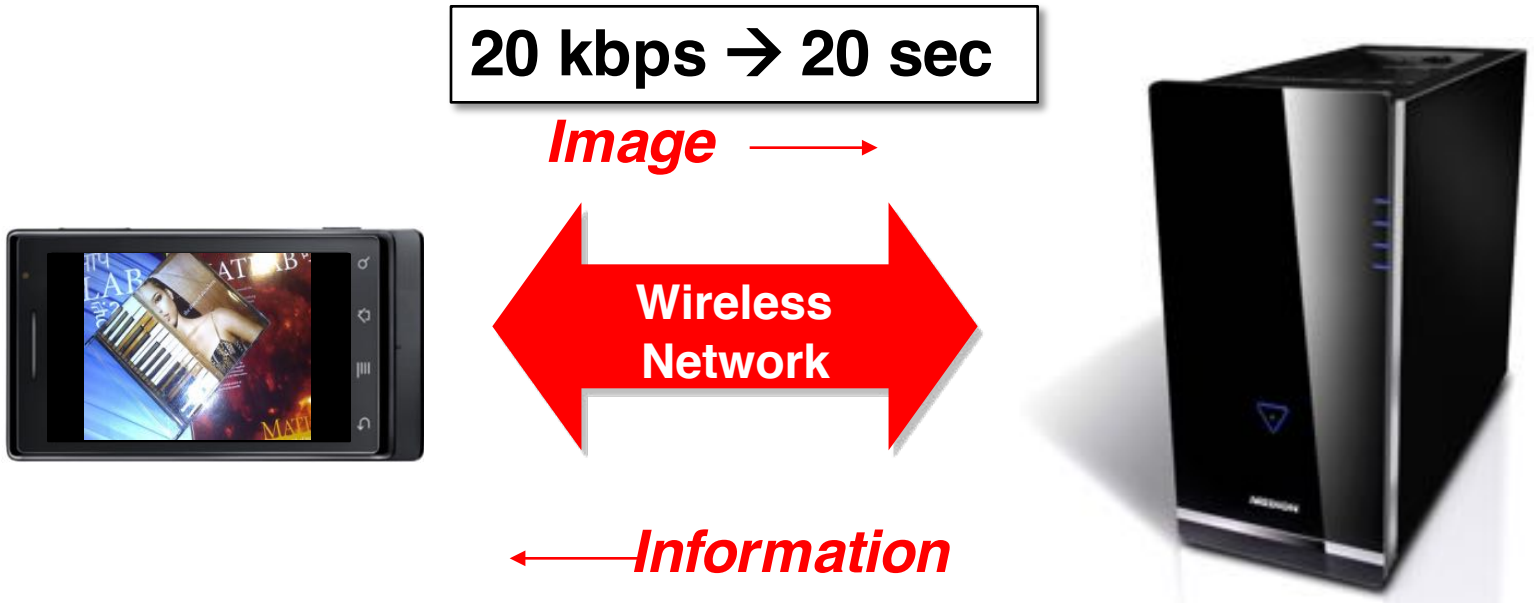
Media Cover Recognition

Nokia N95 Smartphone

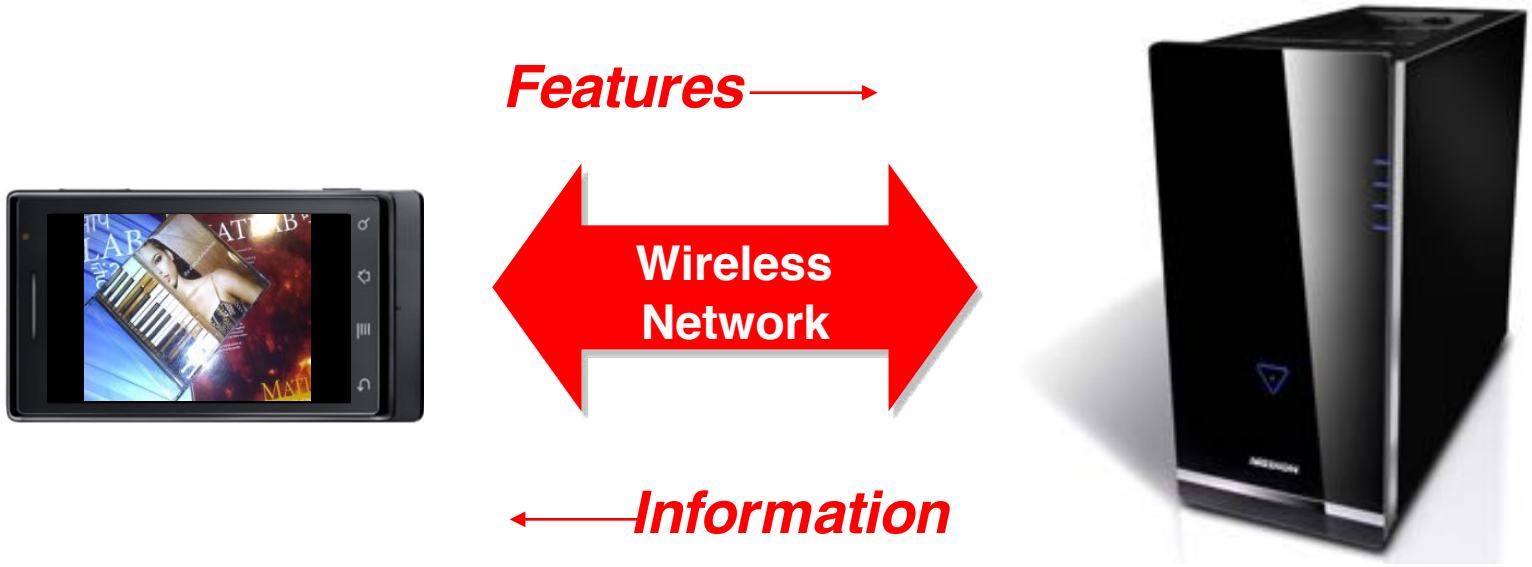
Recognizing Books on A Shelf

Motorola Droid Smartphone

Architecture A: Send Image



Architecture B: Send Features

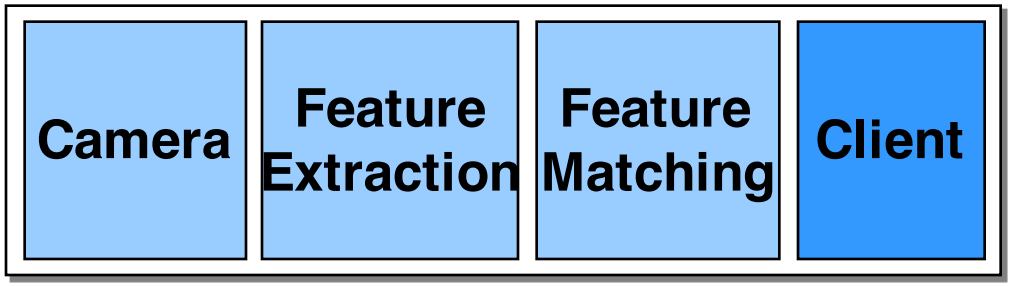


Architecture C: Features on Mobile Device

Wireless Network

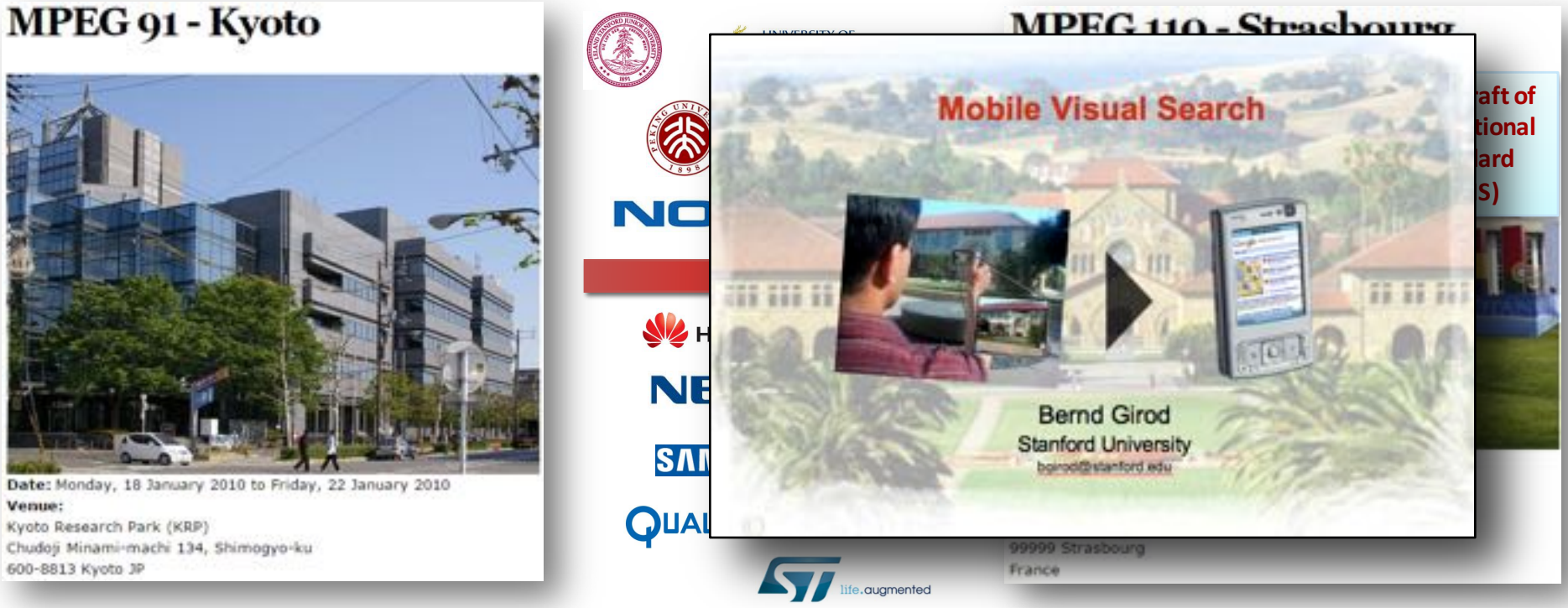
Features

Information



CDVS Standardization

- Moving Picture Experts Group (MPEG - ISO/IEC JTC1 SC29 WG11) initiated the **Compact Descriptors for Visual Search (CDVS)** standard activity at the 91st MPEG meeting (Kyoto, Jan. 2010).



CDVS Evaluation Framework

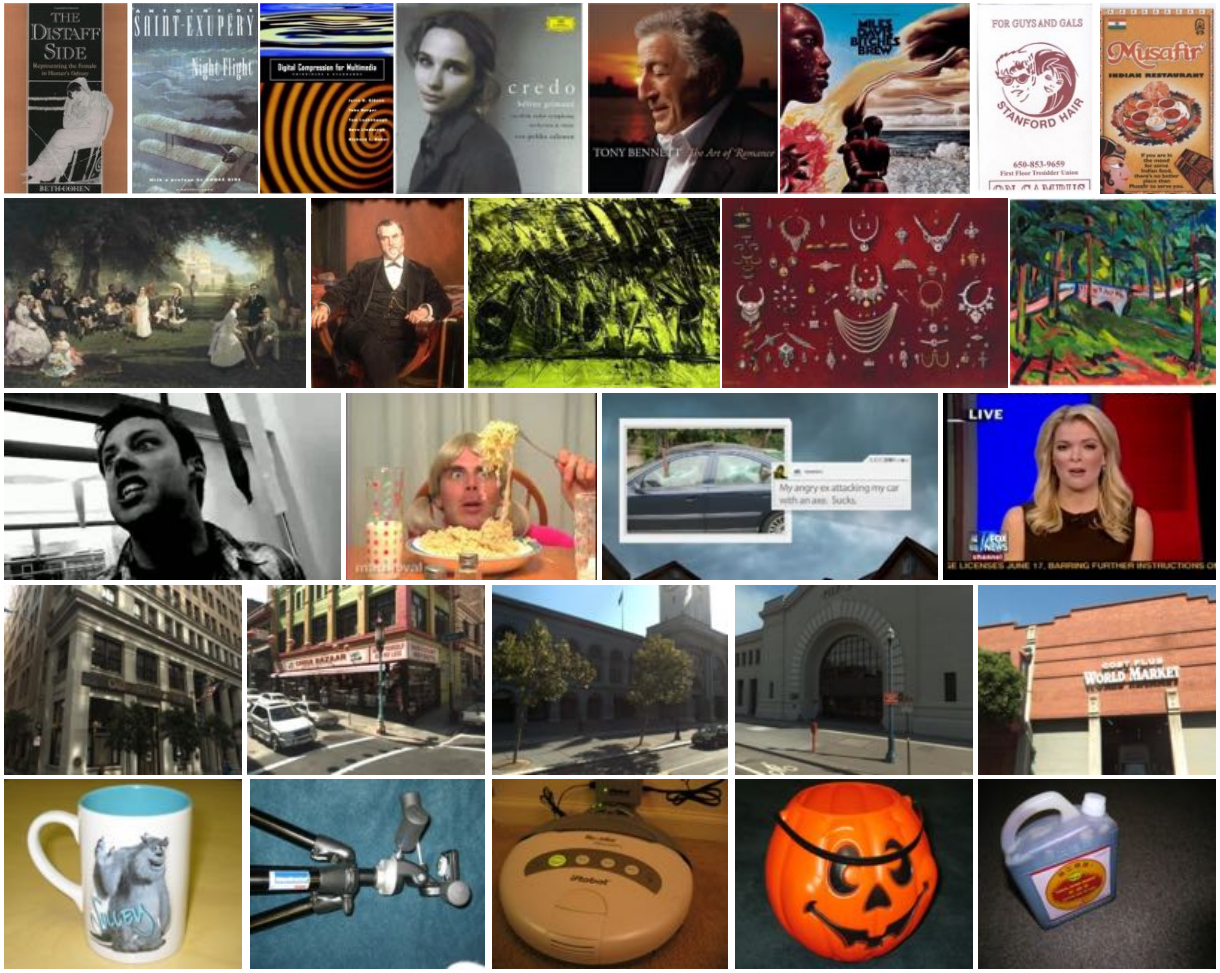
Graphics

Paintings

Video Frames

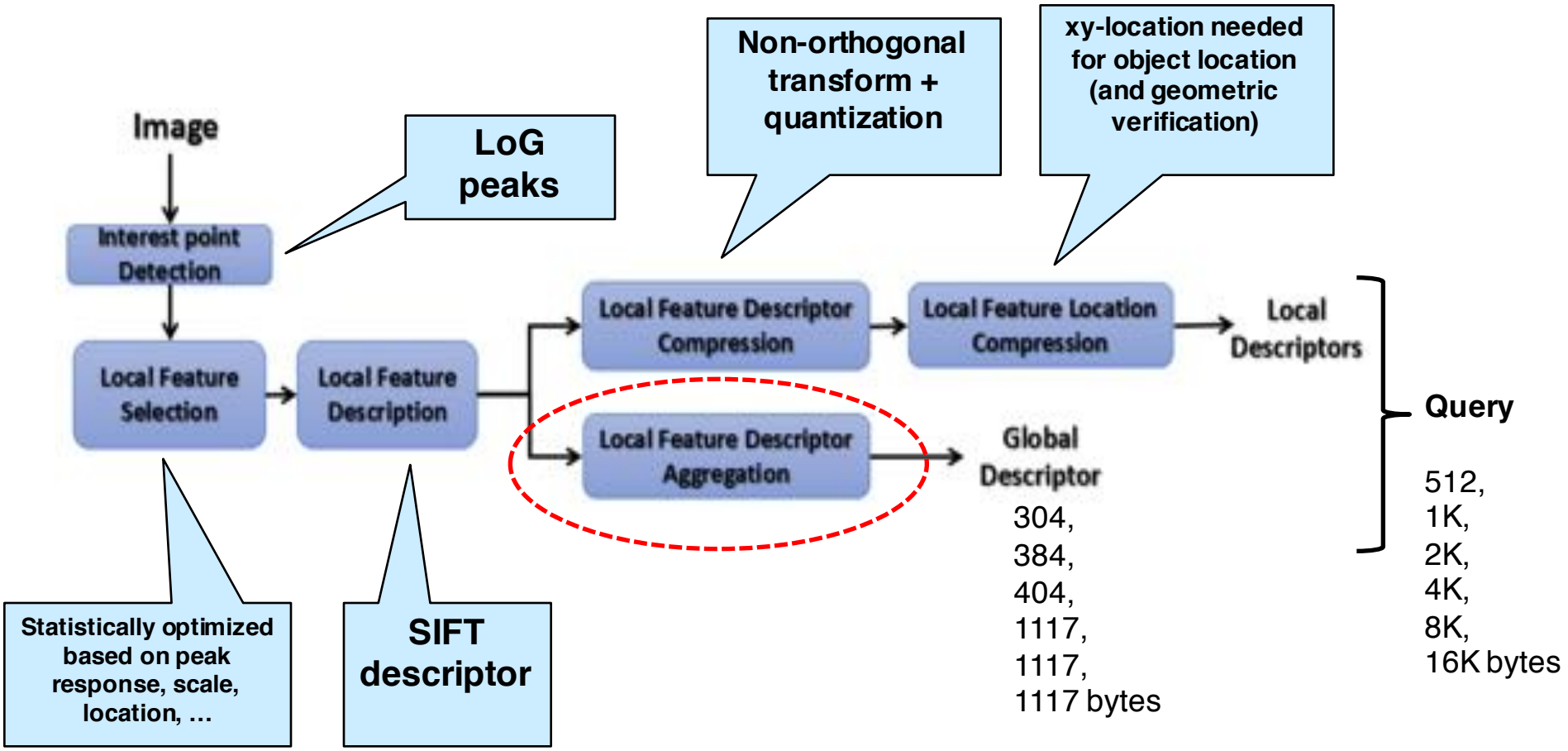
Landmarks

Common Objects




THE DISTAFF SIDE
SAINT-EXUPERY
Night Flight
Digital Compression for Multimedia
credo
TONY BENNETT
FOR GUYS AND GALS
STANFORD HAIR
650-853-9659
Musafir
INDIAN RESTAURANT
1M Distractor Images
My angry ex attacking my car with an axe. Sucks.
LIVE
WORLD MARKET

CDVS Pipeline



Local Feature Descriptor Aggregation

- Nearest-neighbor matching of variable-size sets of local features is costly
- Compare images based on a global binary signature of constant size ("hash") instead
- **Naïve:** VQ of feature vectors to generate histogram, compare non-empty histogram bins ("bag of features," "bag of visual words")
- **Better:** binarize gradient of log likelihood of w.r.t. to parameter vector ("Fisher vector")

Fisher Vector

- Discriminative score function

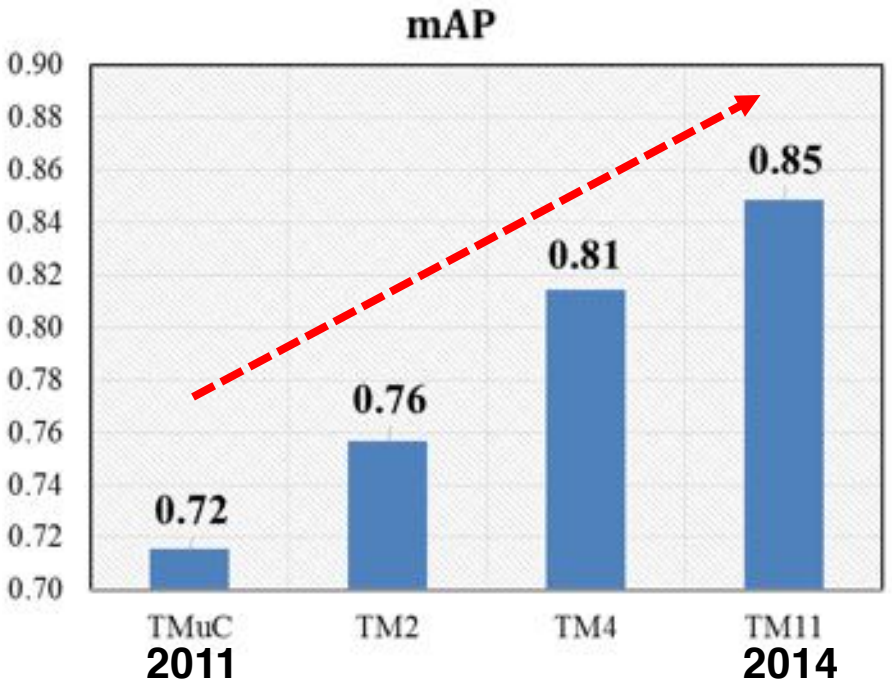
$$U(X) = \frac{\partial}{\partial \Theta} \log p_{X|\Theta}(X|\Theta)$$

d-dimensional vector $\quad \boldsymbol{d \gg k} \quad$ k-dimensional feature vector $\quad$ d Parameters

- Typical, we use Gaussian mixture model (GMM) for $p_{X|\Theta}(X|\Theta)$
- Parameters Θ: mean (and variance) of Gaussian clusters
- For GMM, feature scores $U(X)$ are soft-assigned distance vectors (and squared distance vectors) relative to cluster centers
- Sums of feature scores of an image are "Fisher vector" that can be used to compare images
- Binarization & Hamming distance comparison results in only minor performance loss ("Binarized Fisher vector")

CDVS Evolution

Average performance over all datasets and test conditions



TMuC first reference software (based on SIFT)

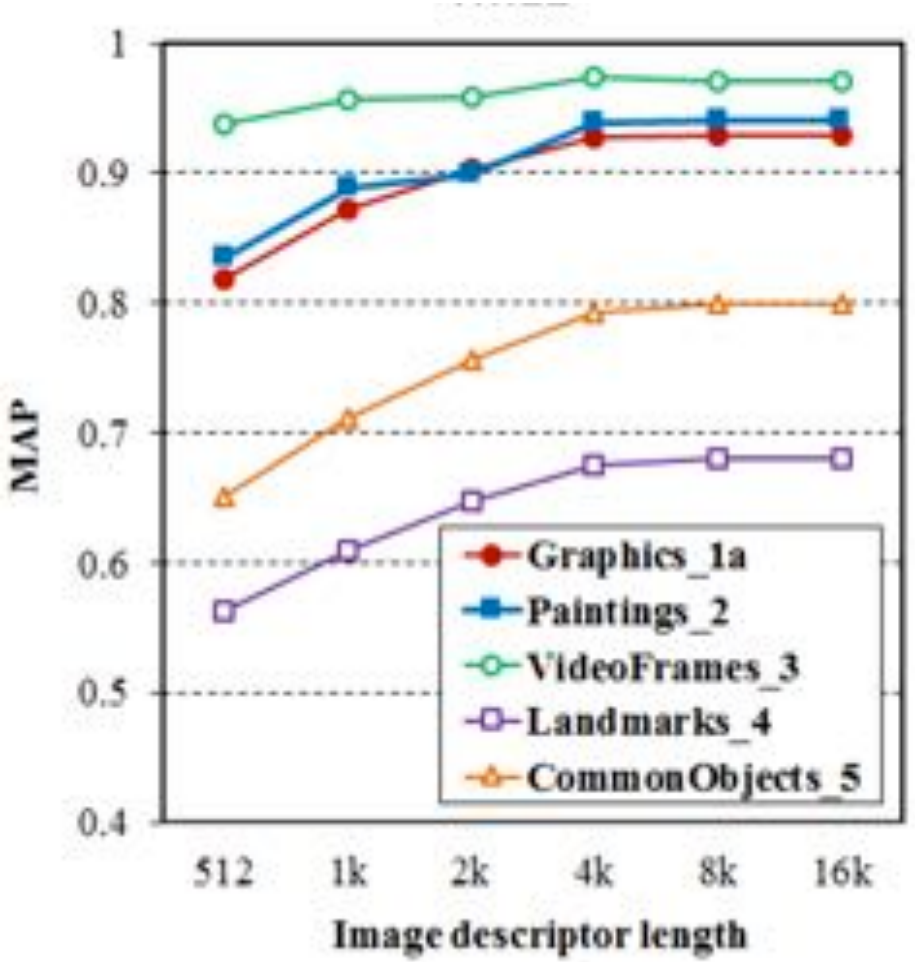
TM2 Global descriptor ("REVV") based on Fisher vector framework introduced

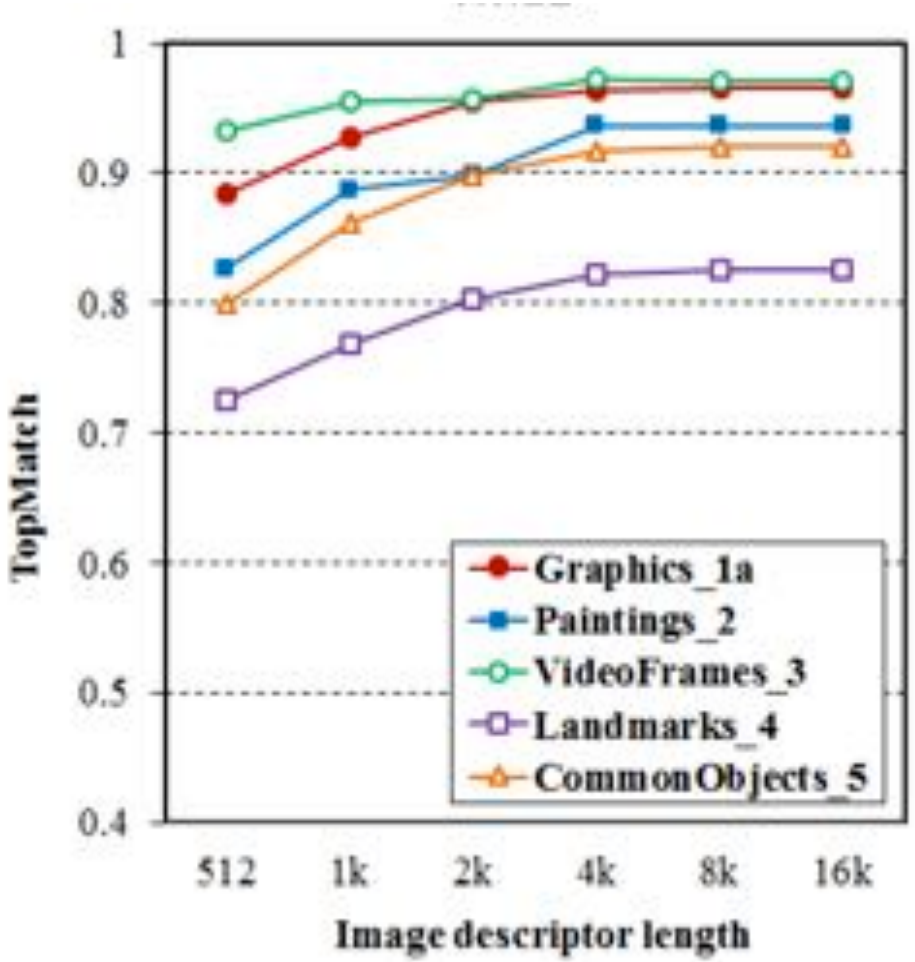
TM4 Scalable Fisher Vector (SCFV)

TM11 Technology development complete

Reduced algorithm memory requirements from **~400 MB** to **~1MB** at the same time

CDVS Performance (TM11)





Architecture C: Features on Mobile Device

Wireless Network

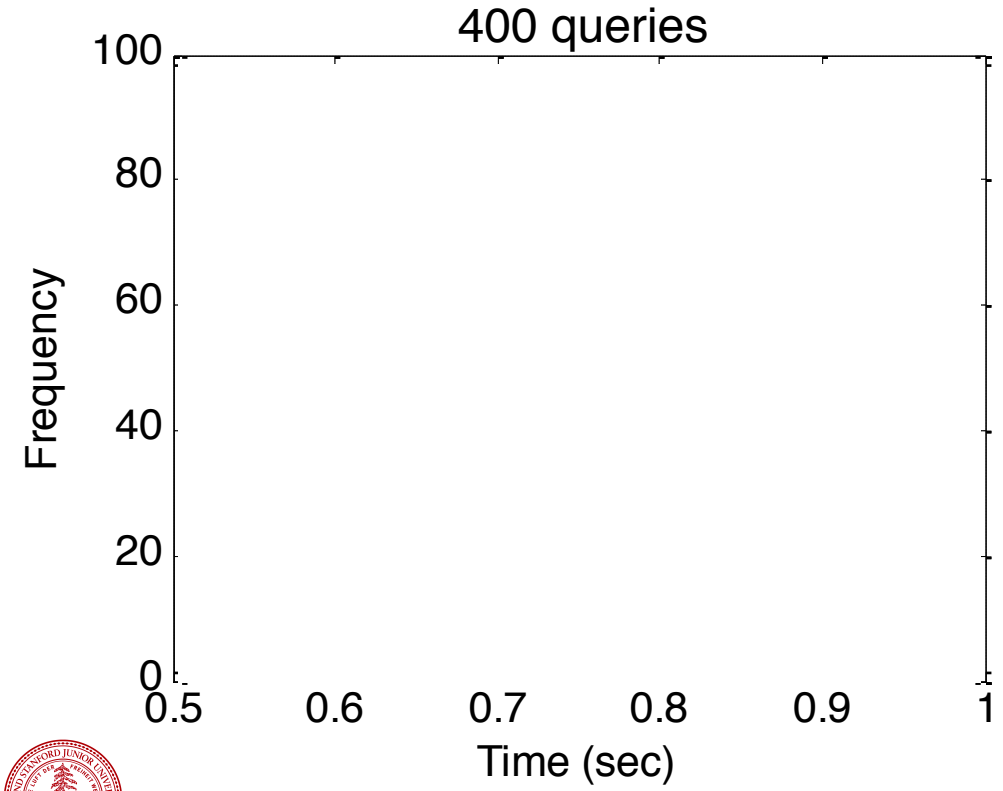
Features

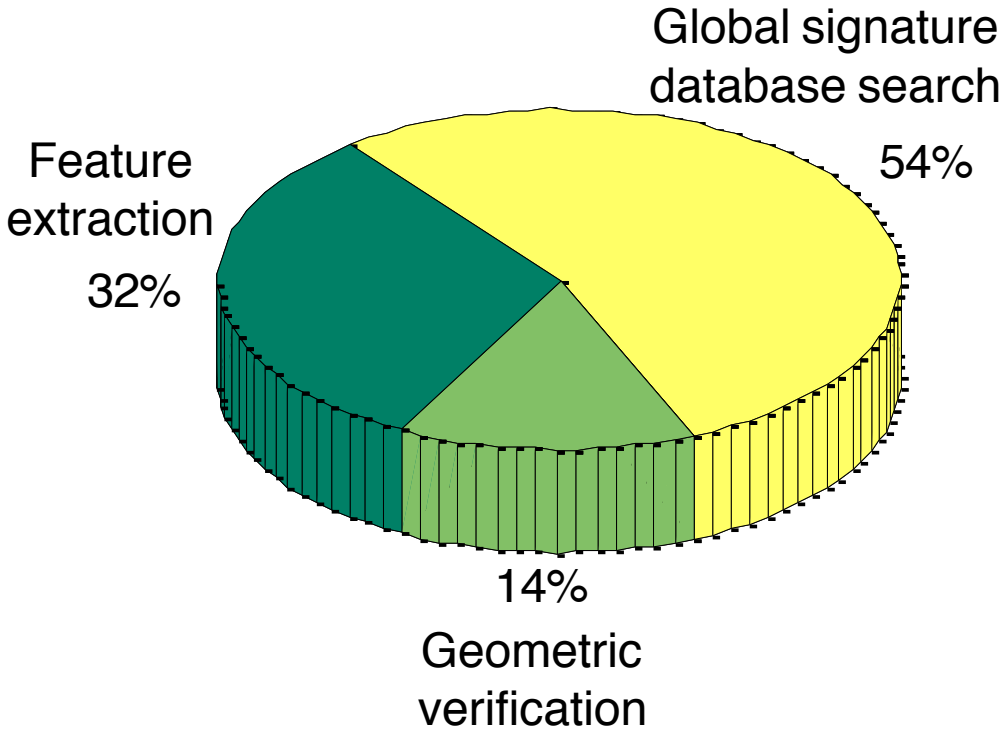
Information

Camera	Feature Extraction	Feature Matching	Client

On-Device Timing Measurements

Samsung Galaxy S3 Smartphone
1.4 GHz Processor
1 GB RAM
Database of 100K Images





On-Device Image Matching Demo
Database of 100K Images

Samsung Galaxy S3 Smartphone

Augmented Reality Glasses

Augmented Reality Glasses

Augmented Reality Glasses

AR w/ Head-Mounted Camera

[Baidu Eye, 2014]

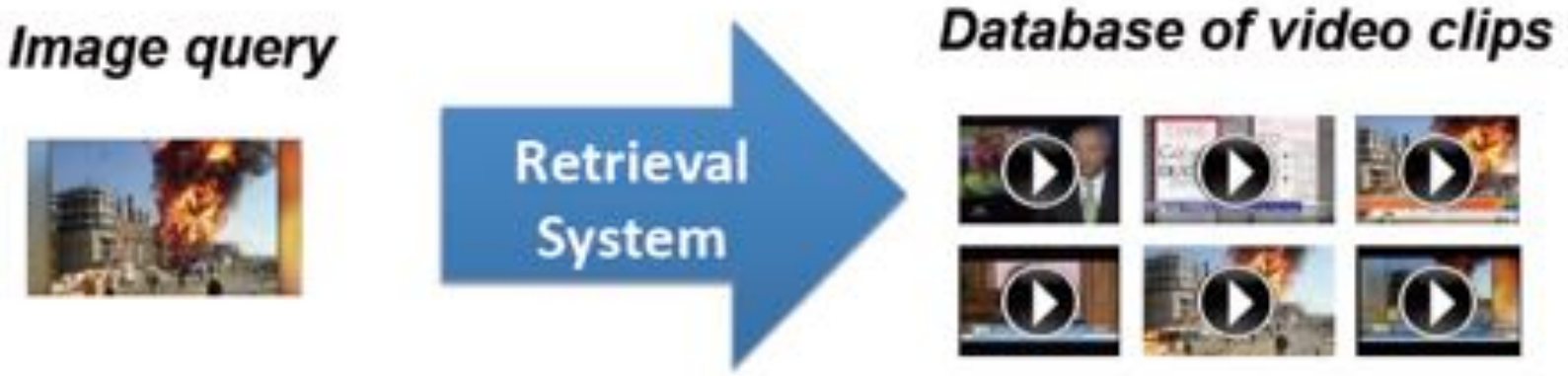
Visual Search: Where Do We Go From Here?

	Database: **Images**	Database: **Videos**
Query: **Image**	*Limitations of SIFT/CDVS framework* • Scale to very large databases • Dense text • Non-planar 3d objects	*Search "Dark matter of the Internet"* • Temporal redundancy of database • Asymmetric comparisons
Query: **Video**	*"Streaming" augmented reality* • Exploit temporal redundancy of queries • Database caching in mobile device	*Tracking of copies* • Leverage audio • Largely solved

Visual Search: Where Do We Go From Here?

	Database: Images	Database: Videos
Query: Image	*Limitations of SIFT/CDVS framework* • Scale to very large databases • Dense text • Non-planar 3d objects	*Search "Dark matter of the Internet"* • Temporal redundancy of database • Asymmetric comparisons
Query: Video	*"Streaming" augmented reality* • Exploit temporal redundancy of queries • Database caching in mobile device	*Tracking of copies* • Leverage audio • Largely solved

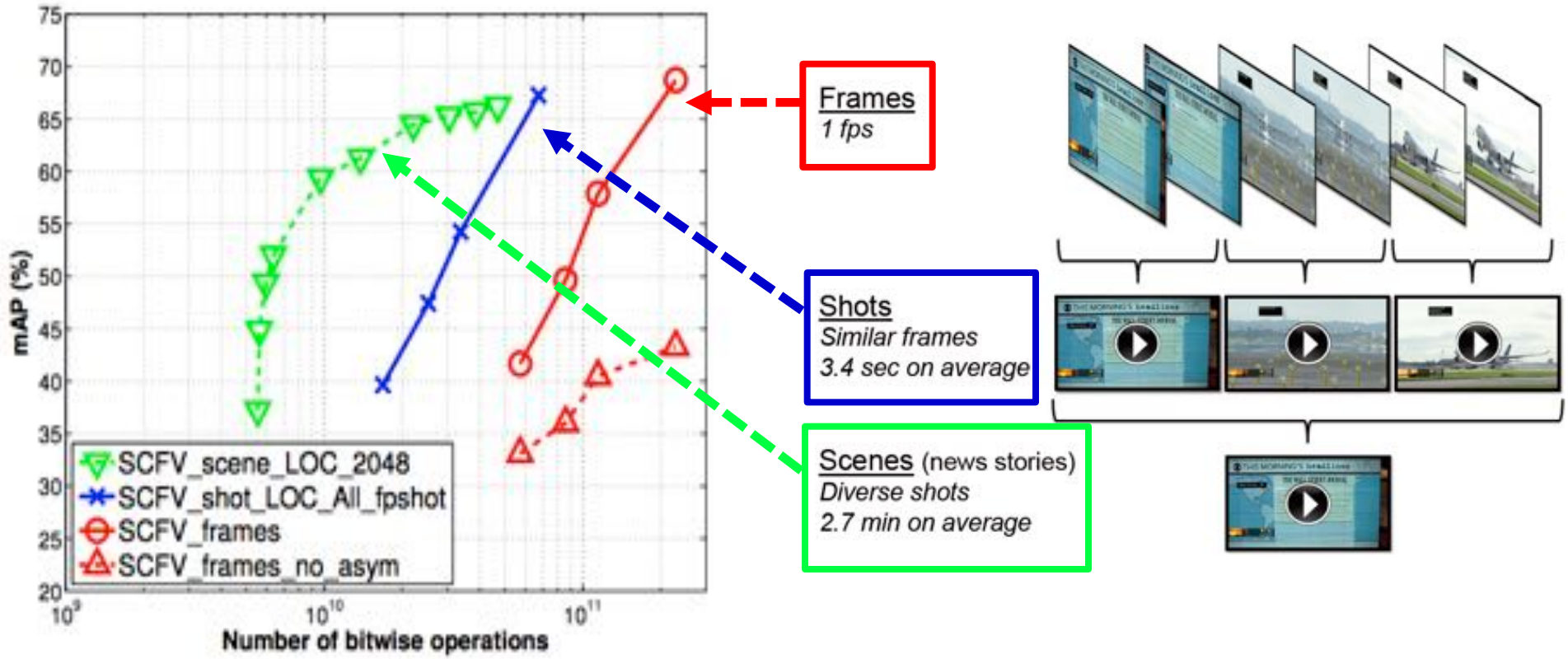
Query-by-Image Video Retrieval



Applications

- News videos: search event footage using photos
- Online education: search lecture videos using slides
- Brand monitoring: search web videos for product placement
- …

Fisher Vector Aggregation



Stanford I2V dataset, 3,800 hours of news videos, 229 query images

[Araujo et al., ICIP 2015]

Asymmetric Comparisons

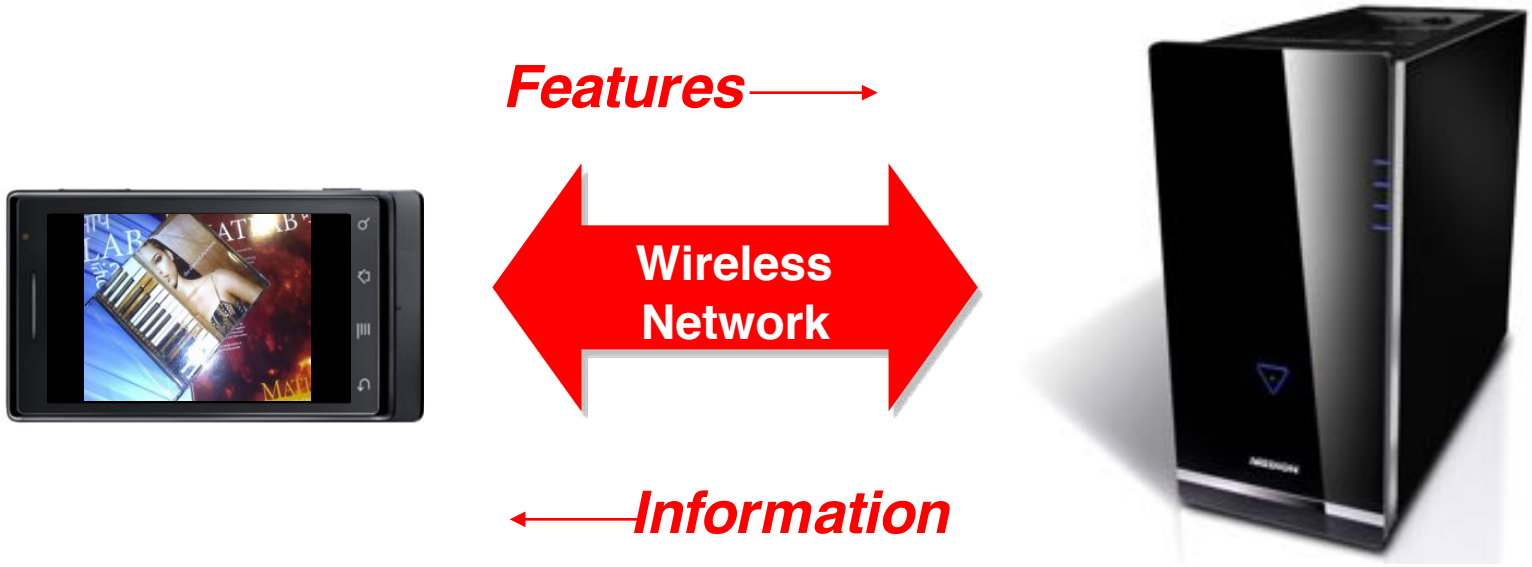
Query Images | **Database Frames**

- Problem becomes more pronounced with temporal aggregation
- **Solution:** omit Fisher vector components of Gaussian clusters that the query does not visit *[Araujo et al., ICIP 2015]*
- Might have to use more Gaussian clusters to accommodate larger number of features on the database side

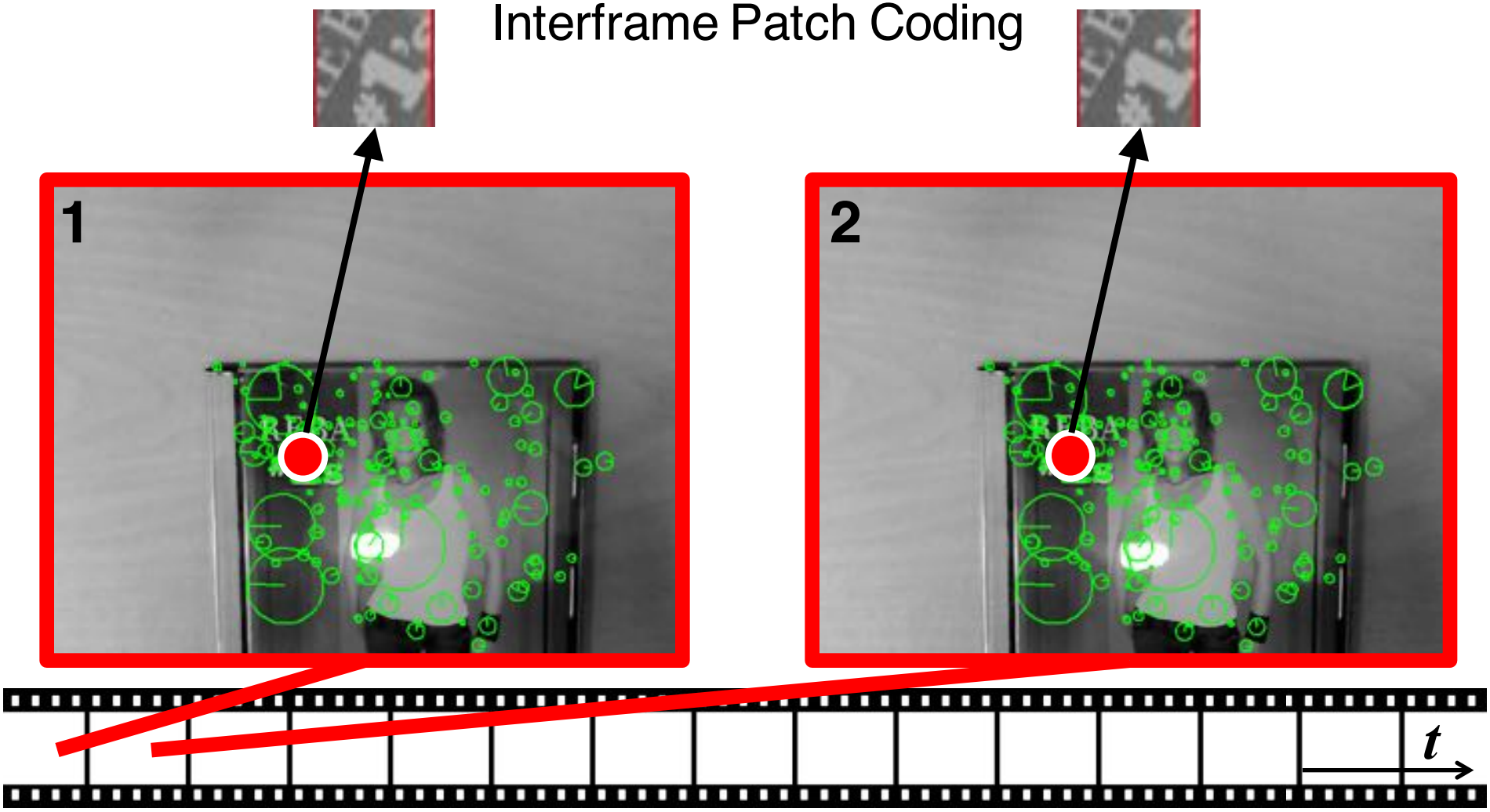
Visual Search: Where Do We Go From Here?

	Database: **Images**	Database: **Videos**
Query: Image	*Limitations of SIFT/CDVS framework* • Scale to very large databases • Dense text • Non-planar 3d objects	*Search "Dark matter of the Internet"* • Temporal redundancy of database • Asymmetric comparisons
Query: Video	*"Streaming" augmented reality* • Exploit temporal redundancy of queries • Database caching in mobile device	*Tracking of copies* • Leverage audio • Largely solved

Architecture B: Send Features

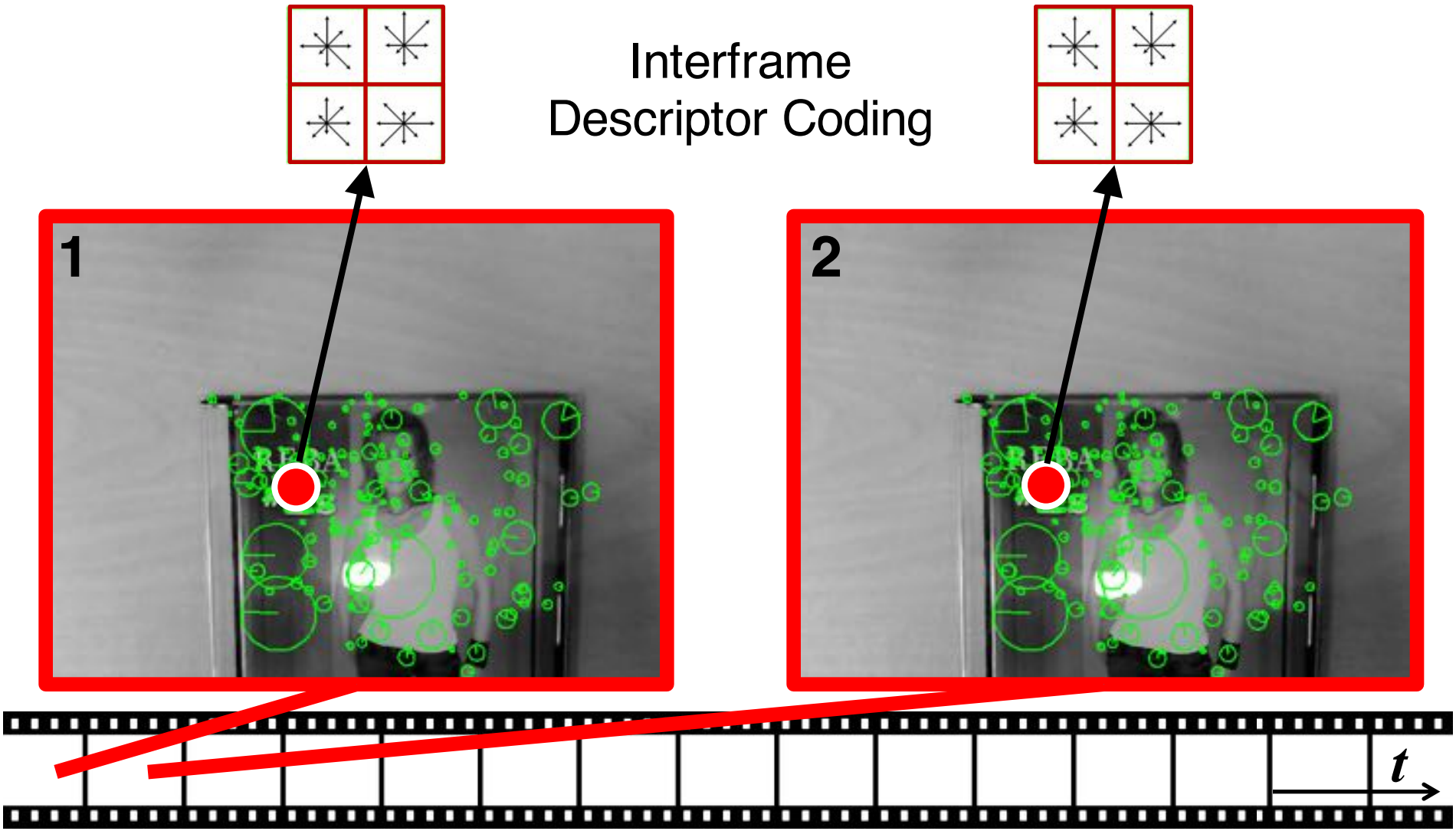


Interframe Compression of Features



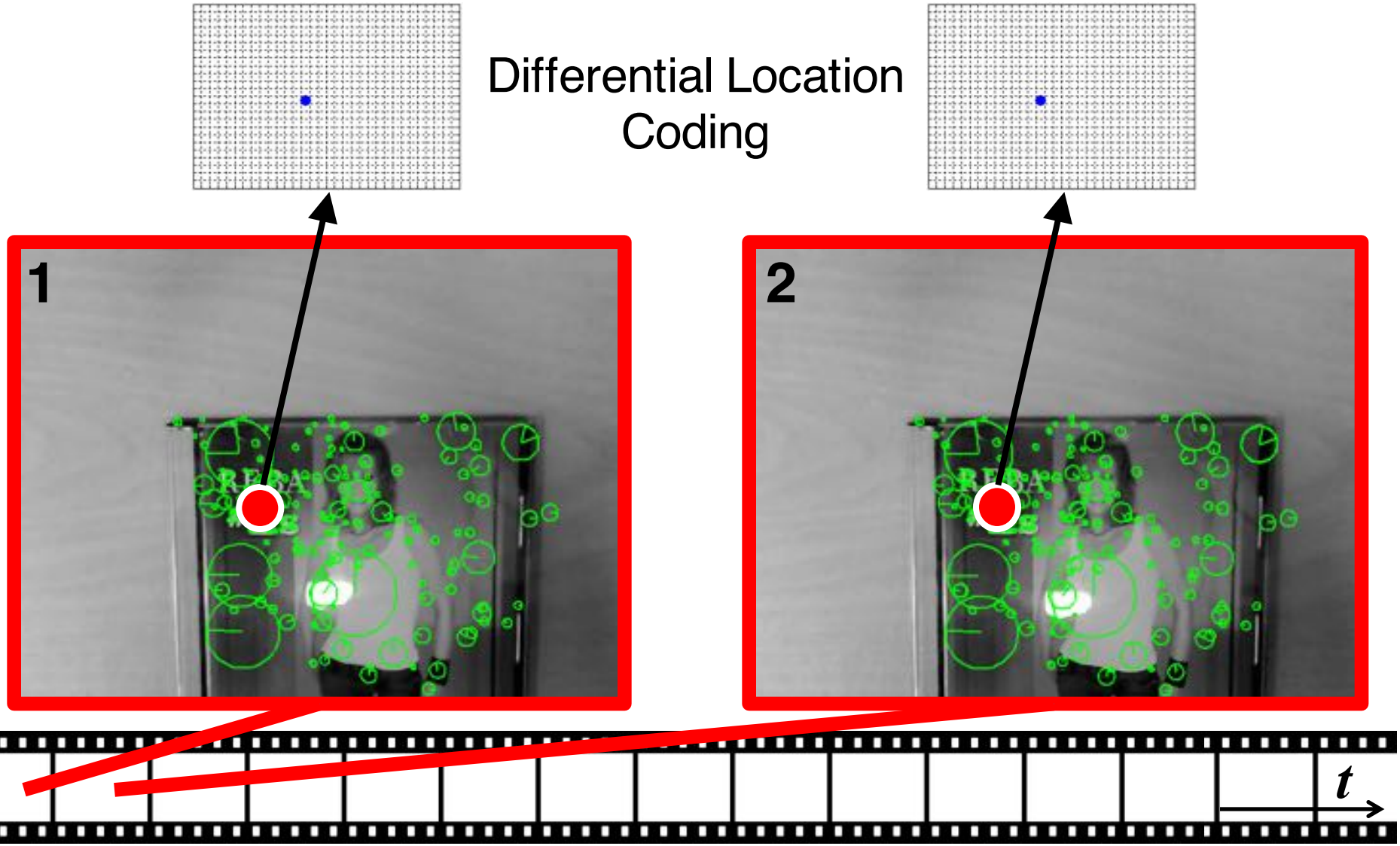
[Makar et al., IEEE Trans. Image Processing, 2014]

Interframe Compression of Features



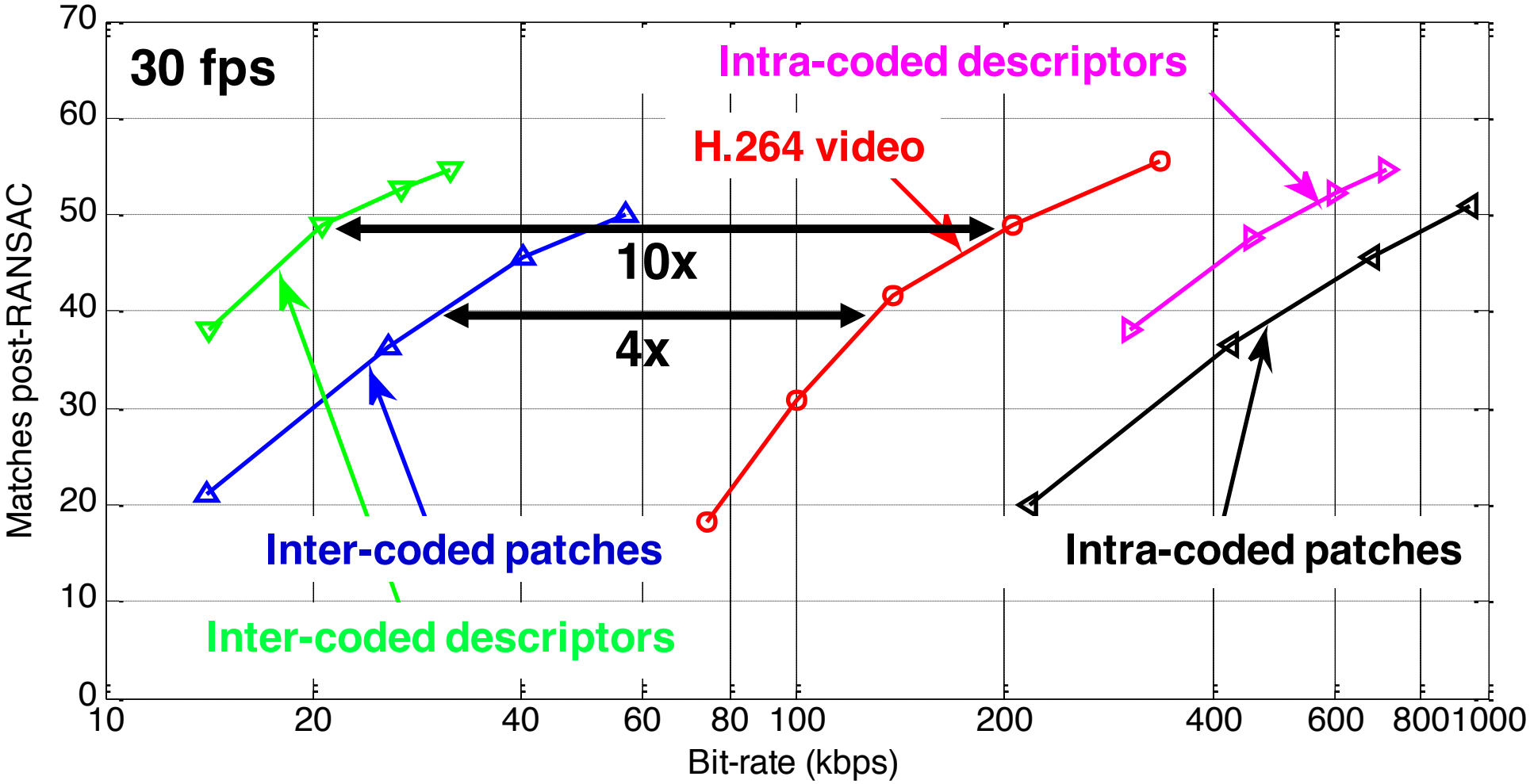
[Makar et al., IEEE Trans. Image Processing, 2014]

Interframe Compression of Features



[Makar et al., IEEE Trans. Image Processing, 2014]

Interframe Compression of Features



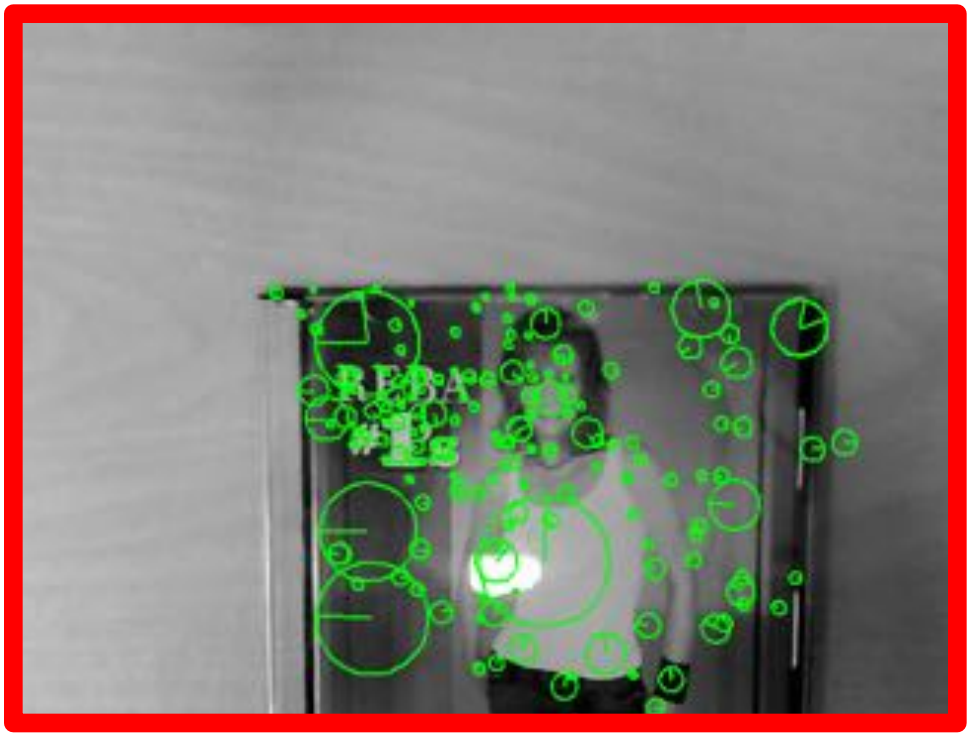
[Makar et al., IEEE Trans. Image Processing, 2014]

Temporally Coherent Keypoint Detection

Conventional keypoint detection

Reba keypoints, frame 2

Temporally coherent

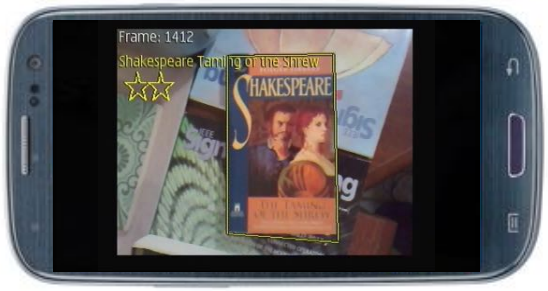


Reba keypoints, frame 2

[Makar et al., IEEE Trans. Image Processing, 2014]

Streaming MAR at ~15 kbps

Hybrid Query Mode



Extract Local Features

Aggregate Global Descriptor

Match with Global Descriptors

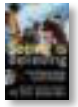
Perform Geometric Verification

Send global descriptors in uplink

Descriptor Database (Mobile)

Wireless Network

0.62



0.51

0.50

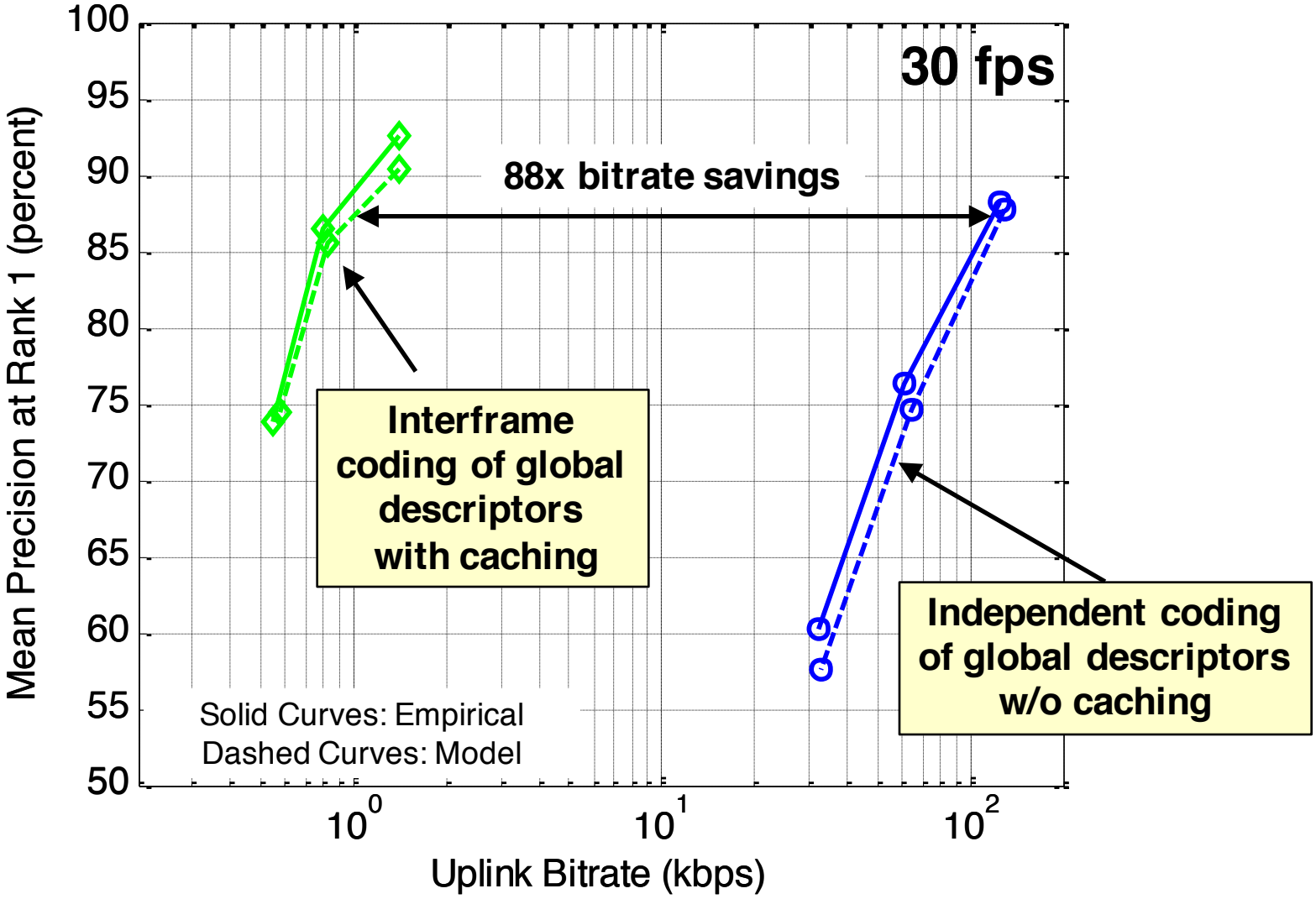
0.49

Match with Global Descriptors

Send labels and local features for top-ranked database candidates in downlink

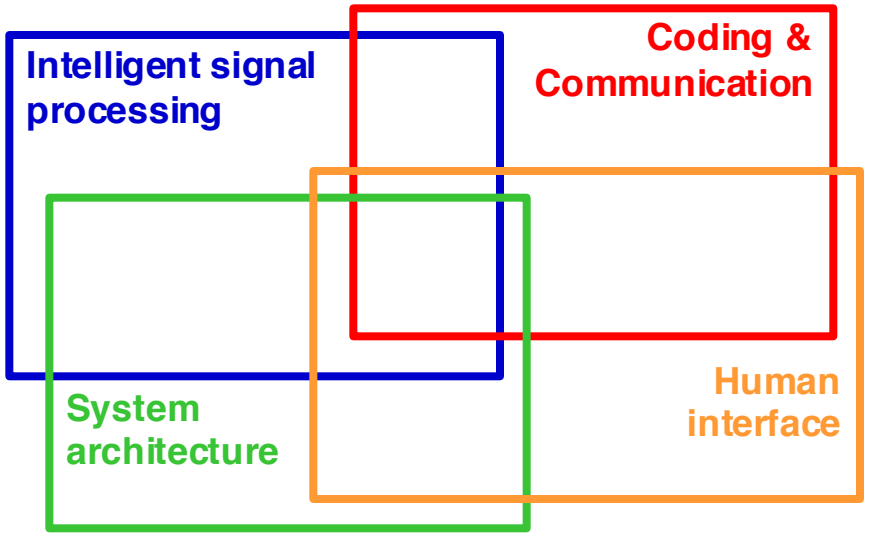
Descriptor Database (Cloud)

Hybrid Query Mode



Conclusion: An Exciting Area!

- Mobile visual search is ready for prime-time
- Wide-spread use of augmented reality with HMDs probably still some years away
- Compression for visual matching is a key problem
  - MPEG standardization “Compact Descriptors for Visual Search” (CDVS)
  - Video is next: MPEG-CDVA
  - Akin to video coding 1980 – still mostly uncharted territory.



Bernd Girod, Vijay Chandrasekhar, David M. Chen, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, Gabriel Takacs, Sam S. Tsai, and Ramakrishna Vedantham

Mobile Visual Search

Linking the virtual and physical worlds

Mobility in Media Search

Mobile phones have evolved into powerful image and video processing devices equipped with high-resolution cameras, color displays, and hardware-accelerated graphics. They are also increasingly equipped with a global positioning system and connected to broadband wireless networks. All this enables a new class of applications that use the camera phone to initiate search queries about objects in visual proximity to the user (Figure 1). Such applications can be used for identifying products, comparison shopping, finding information about movies, compact disks (CDs), real estate, print artworks. First deployments of such systems include Google Goggles [1], Nokia Point and Find [2], Kooaba [3], Ricoh iCandy [4]–[6], and Amazon Snaptell [7].

Mobile image-retrieval applications pose a unique set of challenges. What part of the processing should be performed on the mobile client, and what part is better carried out at the server? On the one hand, transmitting a Joint Photographic Experts Group (JPEG) image could take few seconds over a

Digital Object Identifier 10.1109/MSP.2011.940881
Date of publication: 15 June 2011

Multimedia Data Management in Mobile Computing

Memory-Efficient Image Databases for Mobile Visual Search

David M. Chen
Stanford University

Bernd Girod
Stanford University

Storing a memory-efficient database of image signatures entirely on a mobile device can enable fast local queries regardless of external conditions, such as a slow network or congested server.

Mobile visual search (MVS) systems recognize objects in the user's local environment, retrieve interesting and important information about the objects, and overlay the information in the mobile device's viewfinder. Figure 1 shows a typical example of an MVS system. The system recognizes outdoor buildings, overlays the address and phone number of each building, and shows the building's location on a map of the local neighborhood. MVS systems have also been developed for recognizing and augmenting media covers, product packages, billboards, artwork, and clothing, among other categories of objects. Recent commercial deployments of MVS technologies include Amazon Flow, Kooaba Visual Search, Google Goggles, Nokia Point and Find, and Layar Browser.

For accurate o… images captured b… compared against a… A near-real-time re… seamless and contin… ing MVS systems …

hosted on a remote server and can achieve a low latency, around 1 second, when the network connection is fast and when the server is highly responsive. However, slow transmissions over a wireless network or congestion on a busy server can severely degrade the user experience.

To address this problem, we explain how a memory-efficient database of image signatures stored entirely on a mobile device can enable fast local queries. A locally stored database can provide fast recognition anywhere and anytime, regardless of conditions outside the mobile device. To realize this goal, the image signatures stored in the local database must be extremely compact to fit in the small amount of memory available on the mobile device, capable of efficient comparisons across a large database, and highly discriminative to provide robust recognition for challenging queries. With compact image signatures, a mobile device can store a database containing images of outdoor landmarks, book covers, or product packages, among many more practical examples. When the database requires an update in response to changes in the user's environment or interests, the same signatures should support incremental database updates. Ideally, when server and network conditions improve, these compact signatures can be transmitted to a remote server for expanded queries against a remote database.

In this article, we present four methods recently developed for constructing a compact database from local image-based features and compare their retrieval performances: tree histogram coding (THC),[1] inverted index coding (IIC),[2] residual enhanced visual vector (REVV),[3] and scalable compressed fisher vector (SCFV).[4] Both THC and IIC use compression techniques in conjunction with a bag-of-visual-words histogram to generate compact and discriminative global image signatures. These two methods require the storage of a codebook in the mobile device's memory and decoding of compressed signatures during a query. In contrast, compact REVV and SCFV signatures are generated from bag-of-visual-words residuals. While achieving the same high-level retrieval performance as THC and IIC, REVV and SCFV utilize a much

1070-986X/14/$31.00 © 2014

Overview of the MPEG-CDVS Standard

Ling-Yu Duan, *Member, IEEE*, Vijay Chandrasekhar, *Member, IEEE*, Jie Chen, Jie Lin, *Member, IEEE*, Zhe Wang, Tiejun Huang, *Senior Member, IEEE*, Bernd Girod, *Fellow, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*— Compact descriptors for visual search (CDVS) is a recently completed standard from the ISO/IEC moving pictures experts group (MPEG). The primary goal of this standard is to provide a standardized bitstream syntax to enable interoperability in the context of image retrieval applications. Over the course of the standardization process, remarkable improvements were achieved in reducing the size of image feature data and in reducing the computation and memory footprint in the feature extraction process. This paper provides an overview of the technical features of the MPEG-CDVS standard and summarizes its evolution.**

***Index Terms*— Compact descriptors, feature compression, MPEG-CDVS, visual search.**

I. Introduction

Over the past decade, mobile phones and tablets have become devices that are suitably equipped for visual search applications. With high-resolution cameras, powerful CPUs and pervasive wireless connections, mobile devices can use images as search queries for objects observed by the user. Emerging applications include scene retrieval, landmark recognition, and product identification, among others. Examples of early commercial mobile visual-search systems include Google Goggles [1], Amazon Flow [2] and Layar [3].

The requirements for mobile visual search, such as faster searches, higher accuracy and better user experience, pose a unique set of challenges. Normally, a mobile visual search system transmits JPEG-encoded query images from the mobile end to the remote server, where a visual search is performed over a reference image database. However, image transmission could take anywhere from a few seconds to a minute or more over a slow wireless link, and wireless upload might even time-out in the case of an unstable connection. On the other hand, on-device image analysis, either for mobile image matching or for the transmission of a compact signature to the cloud, might be computationally demanding and hence slow.

In Figure 1, we present four typical client-server architectures, as follows:

- In Figure 1(a), a JPEG-encoded query image is transmitted to the server. Visual descriptor extraction and matching/retrieval are performed entirely on the server;
- In Figure 1(b), visual descriptors are extracted and compressed on the mobile client. Matching/retrieval is performed on the server using the transmitted feature data as the query;
- In Figure 1(c), a cache of the database is maintained on the mobile device, and image matching is performed locally. Only if a match is not found does the mobile device send the query to the server for a remote retrieval;
- In Figure 1(d), the mobile device performs all the image matching locally, which is feasible if the database is small and can be stored on the mobile device.

In each case, the retrieval framework must adapt to stringent mobile system requirements. First, the processing on the mobile device must be fast, lightweight and have low power consumption. Second, the size of the data transmitted over the network must be as small as possible to reduce the network latency. Finally, the algorithms used for retrieval and matching must be scalable to potentially very large databases and robust to allow reliable recognition of objects captured under a wide range of conditions, such as partial occlusions, changes in vantage point, camera parameters, and lighting.

Initial research on the topic [4]–[9], [21] demonstrated that one could reduce transmission data by at least an order of magnitude by extracting compact visual features efficiently on the mobile device and sending descriptors at low bitrates to a remote server for performing the search. A significant reduction in latency could also be achieved when performing all processing on the mobile device itself.

Following initial research on the topic, an exploratory activity in the Moving Picture Experts Group (MPEG) (formal title "ISO/IEC JTC1 SC29 WG11") was initiated at the 91st meeting (Kyoto, Jan. 2010). As MPEG exploratory work progressed, it was recognized that the suite of existing MPEG technologies, such as MPEG-7 Visual, did not include tools for robust image retrieval and that a new standard would therefore be needed [10]. It was further recognized that, among several component technologies for image retrieval, such a standard should focus primarily on defining the format of descriptors and those parts of their extraction needed to ensure interoperability. Such descriptors need to be compact,

Manuscript received July 24, 2015; revised November 2, 2015; accepted November 3, 2015. Date of publication November 11, 2015; date of current version December 3, 2015. This work was supported in part by the National Science Foundation of China under Contracts 61271311, 61390515, 61210005, and in part by the National High-Tech Research and Development Program of China (863 Program) under Grant 2015AA016302. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jie Liang. Ling-Yu Duan and Vijay Chandrasekhar are joint first authors. (*Corresponding author: Ling-Yu Duan.*)

L.-Y. Duan, J. Chen, Z. Wang, T. Huang, and W. Gao are with the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing 100871, China (e-mail: lingyu@pku.edu.cn; cjie@pku.edu.cn; zhew@pku.edu.cn; tjhuang@pku.edu.cn; wgao@pku.edu.cn).

V. Chandrasekhar and J. Lin are with the Institute for Infocomm Research, Singapore 138632 (e-mail: vijay@i2r.a-star.edu.sg; lin-j@i2r.a-star.edu.sg).

B. Girod is with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: bgirod@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2500034