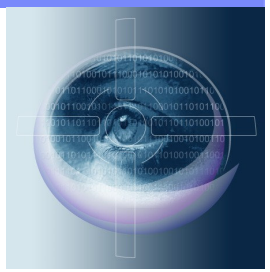


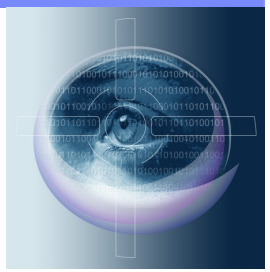
Architecture, Storage, Search/Retrieval, and User Interface for Video Surveillance

YingLi Tian

IBM TJ Watson Research Center

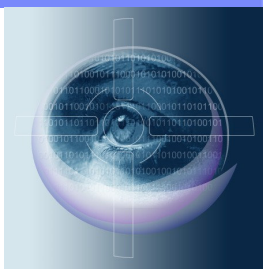
yltian@us.ibm.com

Outline



- Image and Video Retrieval
- Video Surveillance System Architecture
- IBM S3 Data Flow and Storage
- IBM S3 Search and Query
- IBM S3 Interface

Image and Video Retrieval -- 1



- **TREC Video Retrieval Evaluation (NIST, started from 2001, yearly)**

--- http://www-nlpir.nist.gov/projects/trecvid/

- **ACM International Conference on Image and Video Retrieval (started from 1998, yearly)**

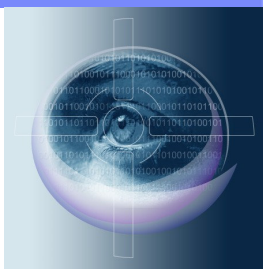
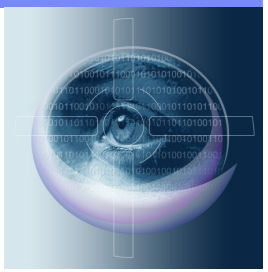


Image and Video Retrieval -- 2

- Feature extraction
  - Emerge from image division
  - *Accumulating* features operate across the entire image or tiles of the image
    - Color histogram
      - Possibly augmented with texture or distance information
    - Compression transforms
      - Intention: preserve all relevant information
      - Should preserve distance information
      - Should carry semantics of image components
      - Should allow indexing of compressed image
  - *Salient* features
    - Weak segmentation yields homogenous regions
    - Store information about the most “conspicuous” regions
    - Goal: store information most robust to changes
      - Often based on invariants
      - Yields regions or points with known location
  - Signs
    - Content of an image with one clear meaning
      - Icon, character, trademark
    - Semantic connection is clear
  - Shape and Object Features

Some contents are from AWM Smeulders, M Worring, S Santini, A Gupta, R Jain “Content-Based Image Retrieval at the End of the Early Years” (http://www.sci.brooklyn.cuny.edu/~sdexter/cis751/CBIREarlyage.ppt)

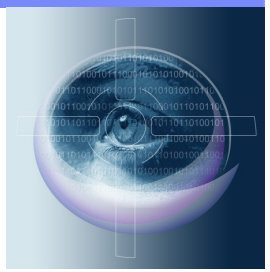
Image and Video Retrieval -- 3



- Feature interpretation and similarity
  - How to give meaning to features?
  - One possibility: assign a *semantic interpretation*
    - *Semantic features aim at encoding interpretations of the image which may be relevant to the application*
    - Each feature may tell us something about the range of possible interpretations of the image, yielding a *weak semantics*
  - Alternatively, try to characterize images/features in terms of *similarity*
- Similarity
  - Could be the distance between two features, considered as vectors
    - Could be a probabilistic measure based on psychological modeling of stimulus, noise
    - Could be any one of a measure of distance between histograms
  - Or a measure of similarity of object shapes
  - Or a measure of similarity of layout or structure (esp in narrow domains)
  - Or a measure of the identity of salient points

Some contents are from AWM Smeulders, M Worring, S Santini, A Gupta, R Jain "Content-Based Image Retrieval at the End of the Early Years" (http://www.sci.brooklyn.cuny.edu/~sdexter/cis751/CBIREarlyage.ppt)

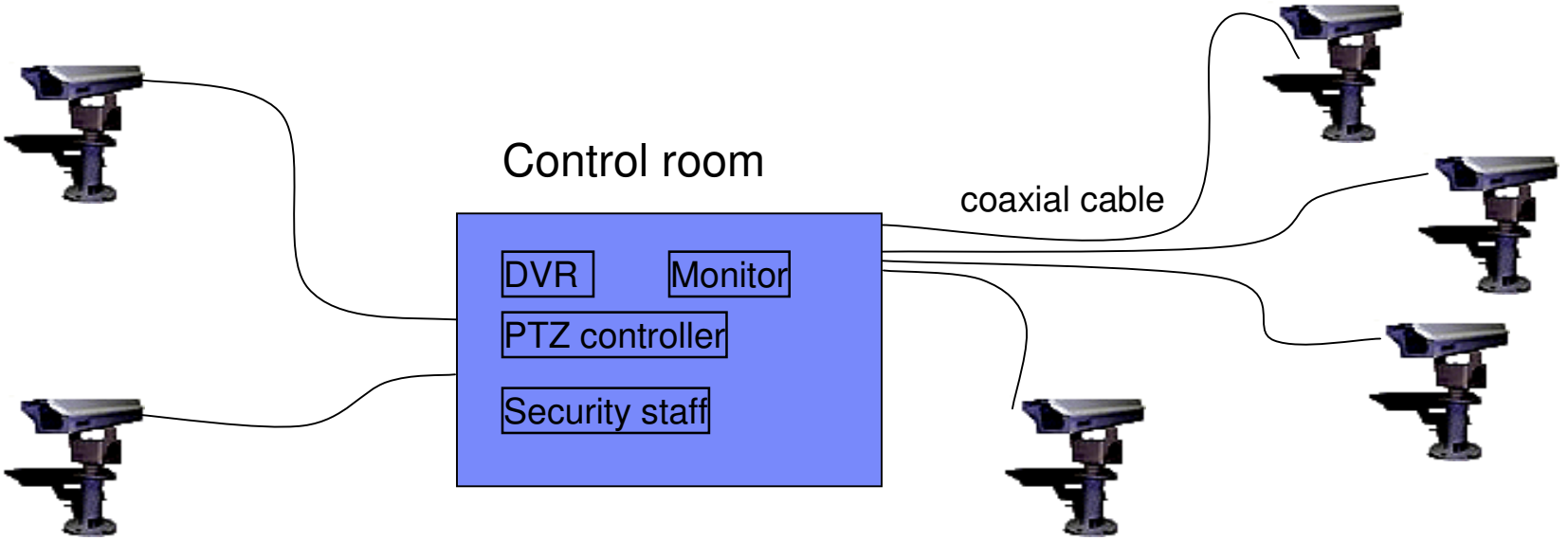
Surveillance architecture



- Legacy systems architecture
- Future systems architecture
- Video sources
- Analytics (software) architecture
- Variations

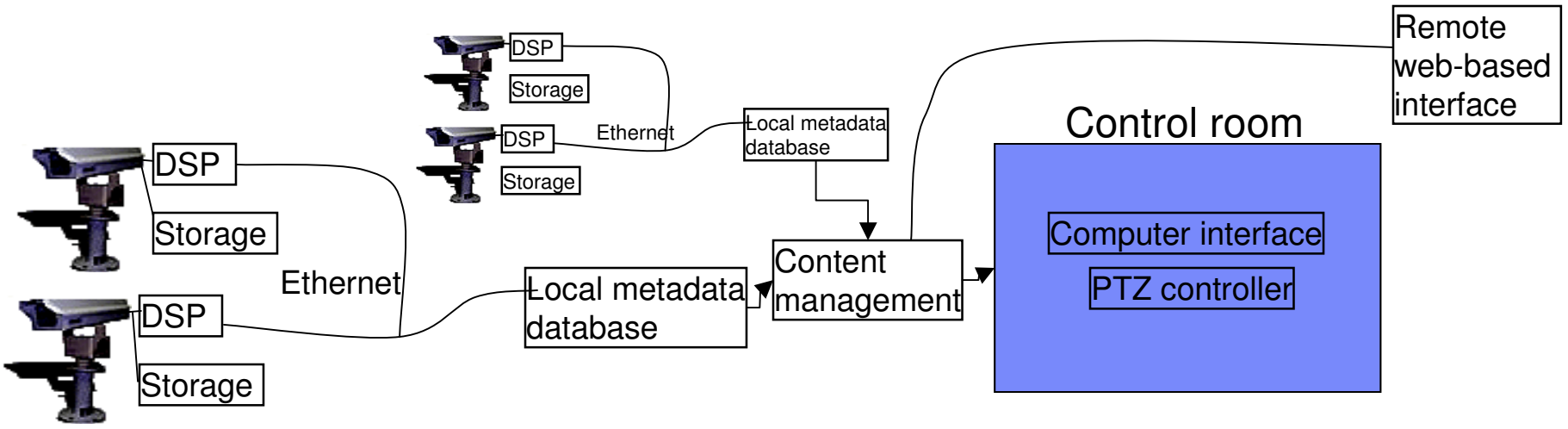
System architecture

- Legacy CCTV systems
  - Analog video, coax cable (maybe multiplexed onto fibre optic)
  - Central control room contains controllers, recorders, monitors, staff



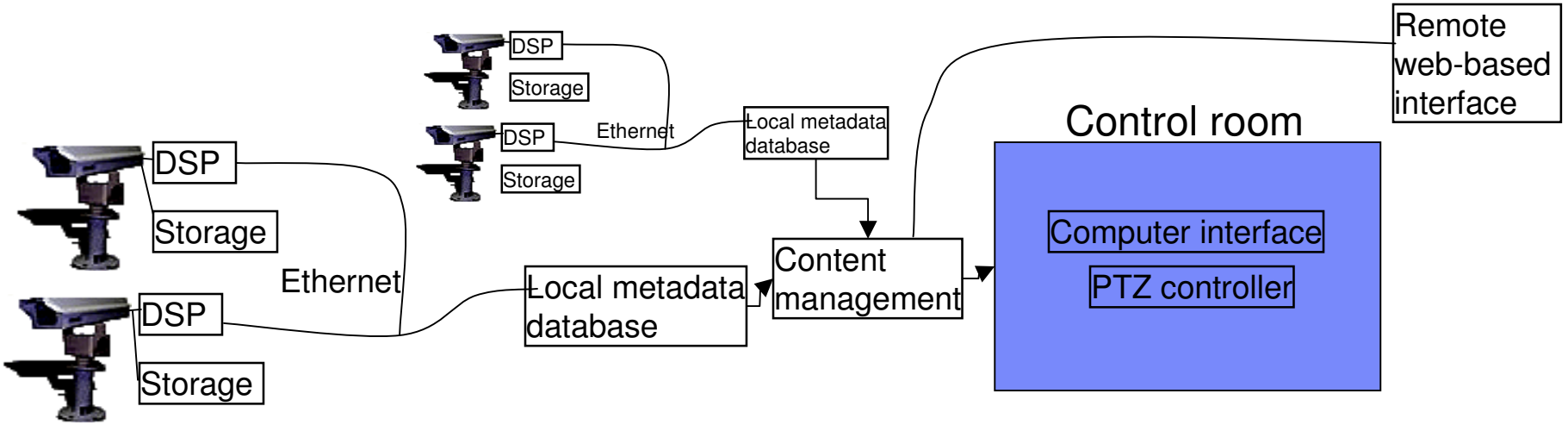
Future system architecture I

- All IP
  - Possibly wireless (though that is subject to attack)
  - Possibly independent of data network
  - All data encrypted
  - Convergence of IT & security departments
  - Convergence of physical and electronic security
  - Central, dynamic, computer-based control
  - Increasingly automated

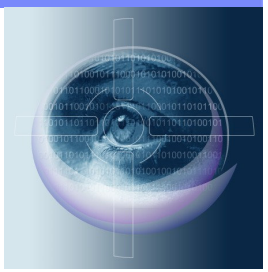


Future system architecture II

- Video storage and processing at the camera
  - DSP encodes, encrypts and interprets video
  - DSP on same wafer as imager
  - Video is not transmitted except when someone needs to view it
- Metadata in distributed clustered content manager

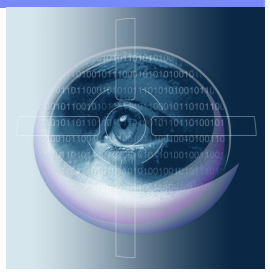


Video sources



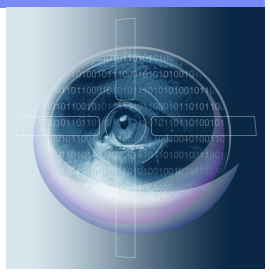
- Analog cameras (NTSC/PAL)
  - To frame-grabber on computer
  - Via encoder box (e.g. Axis analog IP)
  - Constrained ~ 640x480, 30fps
  - Control line for PTZ
- IP cameras
  - Images direct to Ethernet
    - Control over resolution, rate, compression
  - PTZ control over Ethernet
  - Power over Ethernet (1 cable)
- IR/Thermal, multispectral…

Analytics architecture



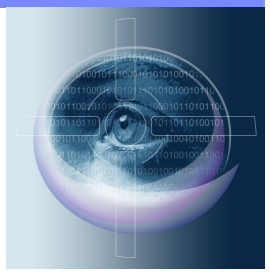
- A series of “modules”
- Our system uses a publish-subscribe architecture
  - Communication through a metadata queue
  - Modules are largely independent and using common protocols to allow recombination

Modularized video analytics



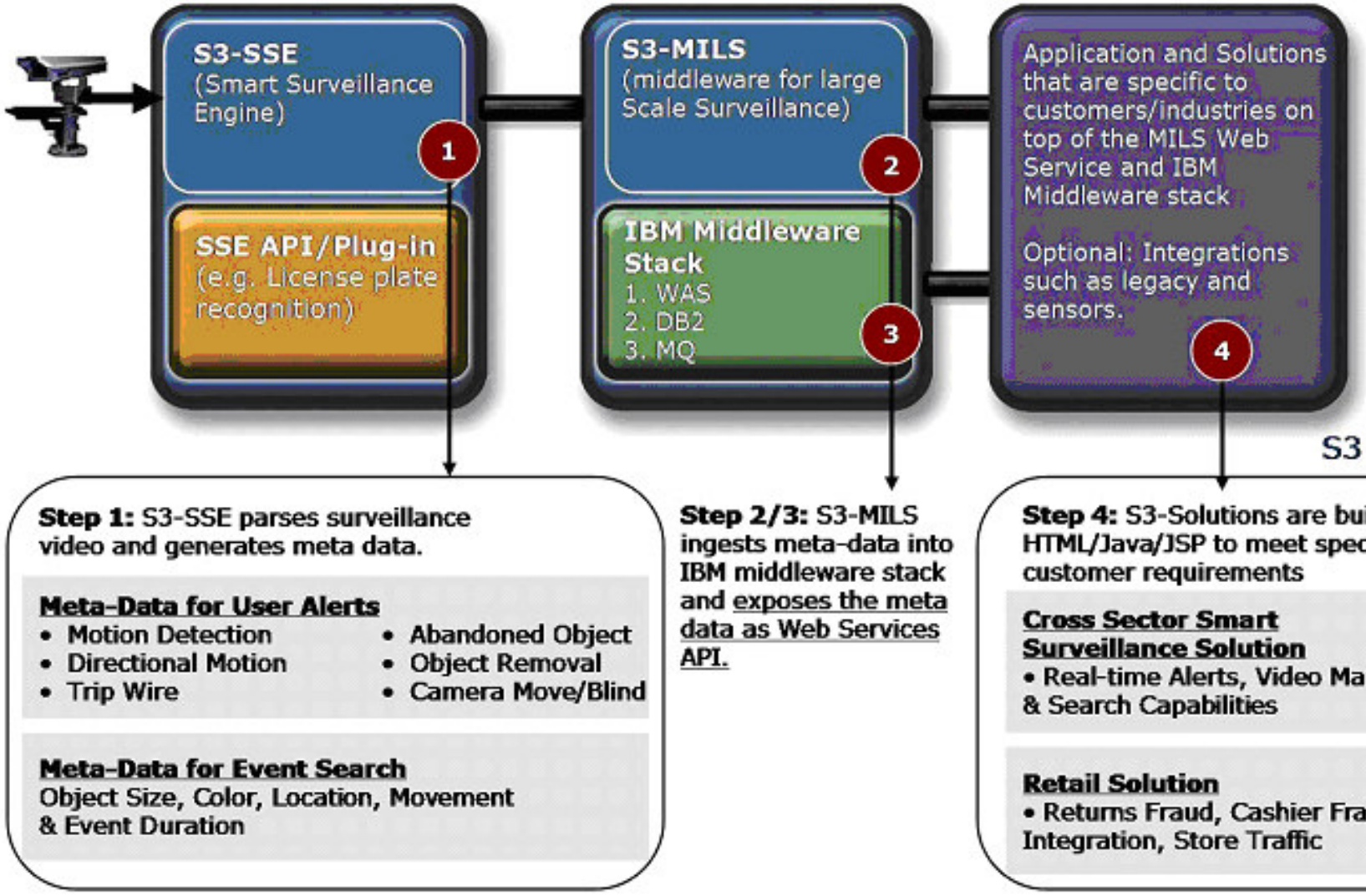
- Data (image) acquisition
- Object detection
  - Background subtraction
  - Motion-based
  - Model-based detection
    - Face, pedestrian, vehicle
- Tracking
- Alert detection
- Classification
- Color indexing
- Recognition (faces, license plate…)
- Behavior analysis
- Communication

Variations

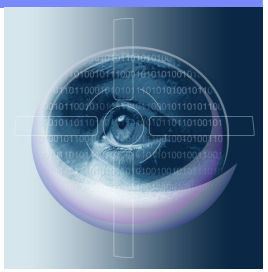


- Steerable cameras
  - Very common in legacy systems
  - Less used in automation
    - Guarantee constant coverage
    - Analytics are easier with static cameras
- Interacting cameras
  - Track from one camera to another
    - Overlapping or with gaps

IBM Smart Surveillance Solutions – Architecture

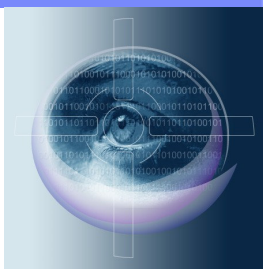


IBM S3 Data Flow



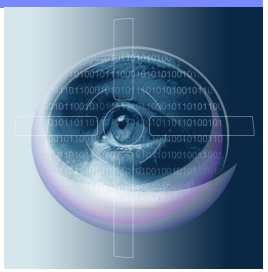
- Sensor data from a variety of sensors is processed in the Smart Surveillance Engines (SSEs). Each SSE can generate real-time alerts and generic event meta-data.
- The meta-data generated by the SSE is represented using XML. The XML documents have some set of fields which are required and common to all engines and others which are specific to the particular type of analysis being performed by the engine.
- The meta-data generated by the SSEs is transferred to the backend MILS system. This is accomplished via the use of web services data ingest APIs provided by MILS.
- The XML meta-data is received by MILS and indexed into predefined tables in the IBM DB2 database. This is accomplished using the DB2 XML extender. This allows for fast searching using the primary keys.
- MILS provides a number of query and retrieval services based on the types of meta-data available in the database.

IBM S3 Event Detection



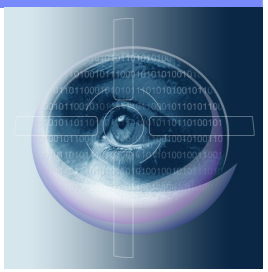
- Camera Stabilization
- Moving Object Detection and Tracking
- Object Classification
- Color Classification
- Alert Detection
- Compound Spatio-Temporal Event Detection
- Face Capture and Tracking
- People Counting
- Behavior Analysis

Compound Spatio-Temporal Event Detection



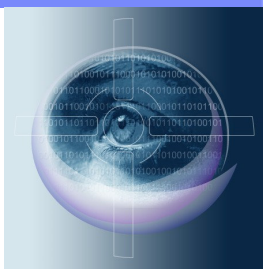
- Definition: multiple events or activities which may occur across different times or multiple cameras based on heterogeneous meta-data
- Solutions:
  - the users specify multiple composite events of high-complexity based on *primitive events* (such as the basic alerts), and then detects their occurrence automatically.
  - Primitive events are connected to each other by an operator using a user-friendly interface.
    - AND
    - OR
    - SEQUENCE
    - Others

IBM S3 **Middleware for Large Scale Surveillance** (MILS)



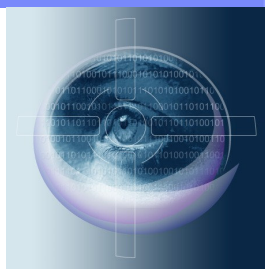
- MILS provides the algorithms needed to take the event meta-data and map it into tables in a relational database.
- MILS provides event search services, meta-data management, system management, user management and application development services.
- MILS uses off the shelf data management (IBM DB2), a web server (IBM Websphere Application Server) and messaging software (IBM MQ) to provide these services.
- Interface -- web applications (written in HTML, Java, JSP, applets, Javascript, and AJAX) which use the web services provided by MILS to provide the functionality needed by the user to query the database and view the results.

Services Provided by the MILS



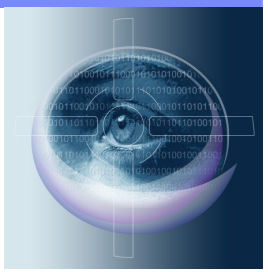
- **Meta-data Ingestion Services**
  - Index Ingestion Services
  - Event Ingestion Services
- **Schema Management Services**
- **System Management Services**
  - Camera Management Services
  - Engine Management Services
  - User Management Services
  - Content Based Search Services

Data Structures in MILS



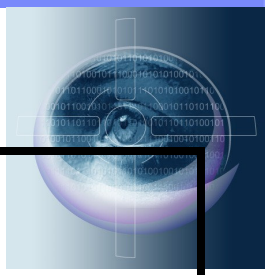
- System data structure
  - captures the specification of a given monitoring system such as geographic location of the system, number of cameras, physical layout of the monitored space, etc
  - Sensor/Camera Data Structure
  - Engine Data Structures
- User data structure
  - contains user names, privileges and user functionality
  - selective access to camera views
  - selective access to camera / engine configuration and system management functionality
  - selective access to search and query functions.
- Event data structure
  - contains the events that occur in a specific sensor or zone in the monitored space
    - Event
    - StartTime
    - Duration
    - Event ID
    - Event Type
    - Event Color

Example Data Models



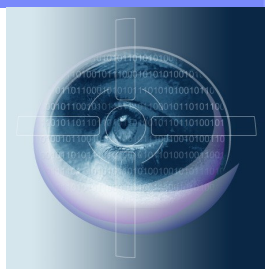
Example Data Models	
Behavior Meta -data	**License Plate Meta -data**
Camera ID: 23	Camera ID: 35
Unique Event ID: 2379406	Unique Event ID: 4926402
Start: 9/10/06:02:22:15:100	Start: 9/10/06:02:12:15:100
End: 9/10/06:02:22:55:300	End: 9/10/06:02:12:25:453
Keyframe : 23567.jpg	Keyframe : 563783.jpg
Video : //mils/xx/file1.wmv	Video : //mils/xx/file3.wmv
Object Type: Car	License Plate #: 525sds
Additional Fields: (trajectory, color, shape, size, etc)	Additional Fields: (e.g State of Origin)

Storage



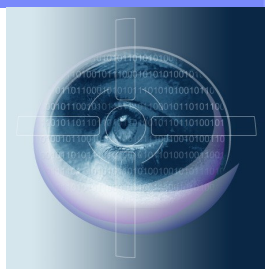
- 1 Million events** utilizes approximately 300 GB
- Guess-timates have a large city deployments (300 camera) generating up to 5 million events per day
  - This equates to ~1.5 TB per day!
  - Retaining 2 weeks of metadata equates to ~21 TB
  - Retaining 1 month of metadata equates to ~45 TB
- Monthly Summary and Track Data/Trajectory data counts
  - 150 Million Summary records
  - 11 – 18 **Billion** Trajectory records ***
- Databases
  - DB2, Oracle, SQL
- Storage
  - Hard Disks
  - Optical Disks
  - Video Servers

Video Query and Search for Surveillance



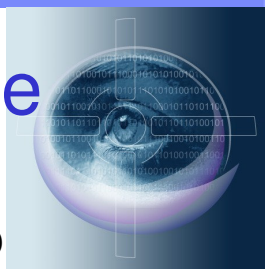
- **Text Based query (LPR)**
- **Content based query (Face Recognition)**
- **Metadata based query**

Query and Search



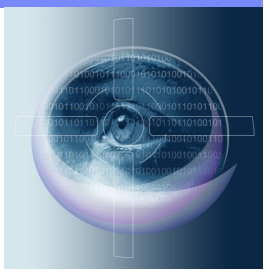
Query: Show me all the **people** who entered the building from **11am to 5pm**, in **November 8th, 2007**.

Challenges for Search and Query for Surveillance



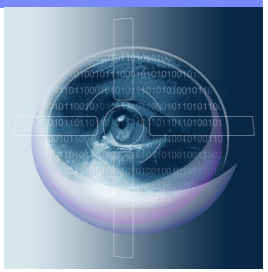
- using knowledge of time and deployment conditions to improve video analysis
- using geometric models of the environment and other object and activity models to interpret events
- using learning techniques to improve system performance and detect unusual events
- Overcome the semantic gap between the feasible low level feature set and the high level semantics or ontology desired by the system users.

Systems with search query capability



- 3VR (3VR) does provide capabilities to search for a person based on face recognition.
- IBM S3
- There is a very limited number of both research and commercial systems focused on searching surveillance video.

Key search criteria



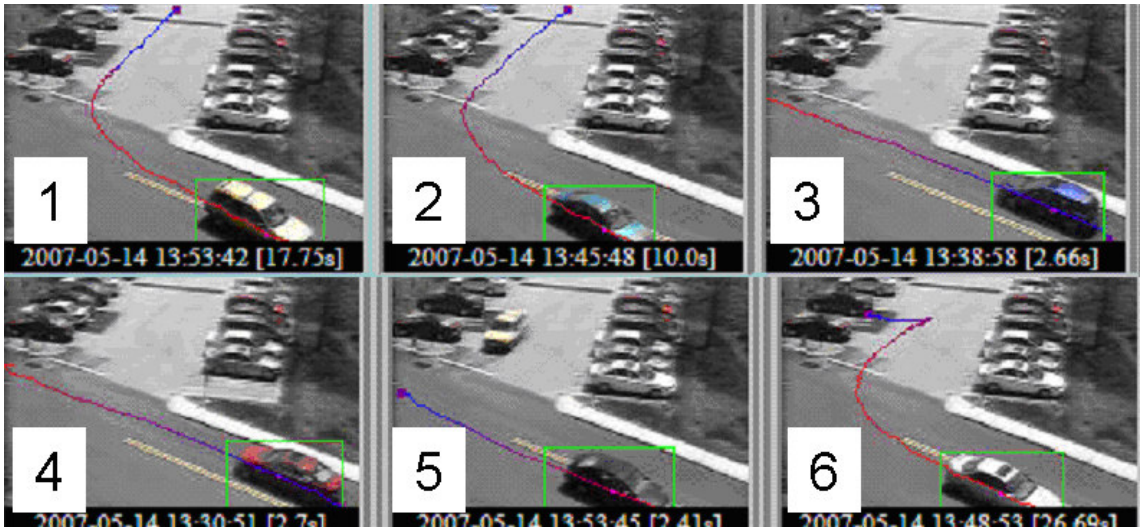
- Specific search for people and vehicles
- Search for alerts
- Search composite events/alerts
- Search across multiple cameras distributed over a spatial region
- Generic search for objects and events of interest
  - Color
  - Size
  - Type
  - Shape
  - Location
  - Duration
  - Velocity
  - Time

Search for face

Total # of people approaching camera		445
Total # of people receding from camera		40
Face Detection	Faces Captured	351
	Faces Missed	94
	False Poitives	7
Person Detection	Persons captured	134
	Approaching	94
	Receding	40
	False Positives	19
Overall People False Negatives		0
Overall People False Positives 26/445		5.6%

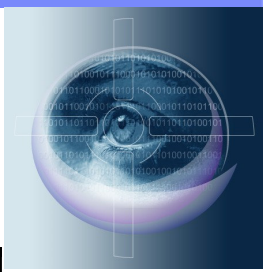
Results obtained from ten hours of surveillance video for entrance of a building

Search by Color -- 1



Retrieved keyframes (cross indexed to video by time) (1) yellow, (2) green, (3) blue, (4) red, (5) black and (6) white vehicles. (Trajectory color indicates direction of movement, blue is track start, red is track end)

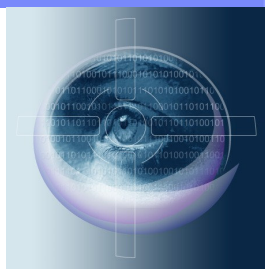
Search by Color -- 2



	COLOR SEARCH GROUND TRUTH							
COLOR SEARCH RESULTS		BL	WH	RE	YE	BU	GR	
	BL	119	36	20	0	1	0	165
	WH	3	102	1	0	1	0	107
	RE	0	0	18	0	0	0	18
	YE	0	0	0	1	0	0	1
	BU	0	0	0	0	7	0	7
	GR	0	0	0	0	0	2	2
		122	138	39	1	9	2	

Color Results: BL-Black, WH-White, RE-Red, YE- Yellow, BU-Blue, GR-Green

Search by Object Class and Size

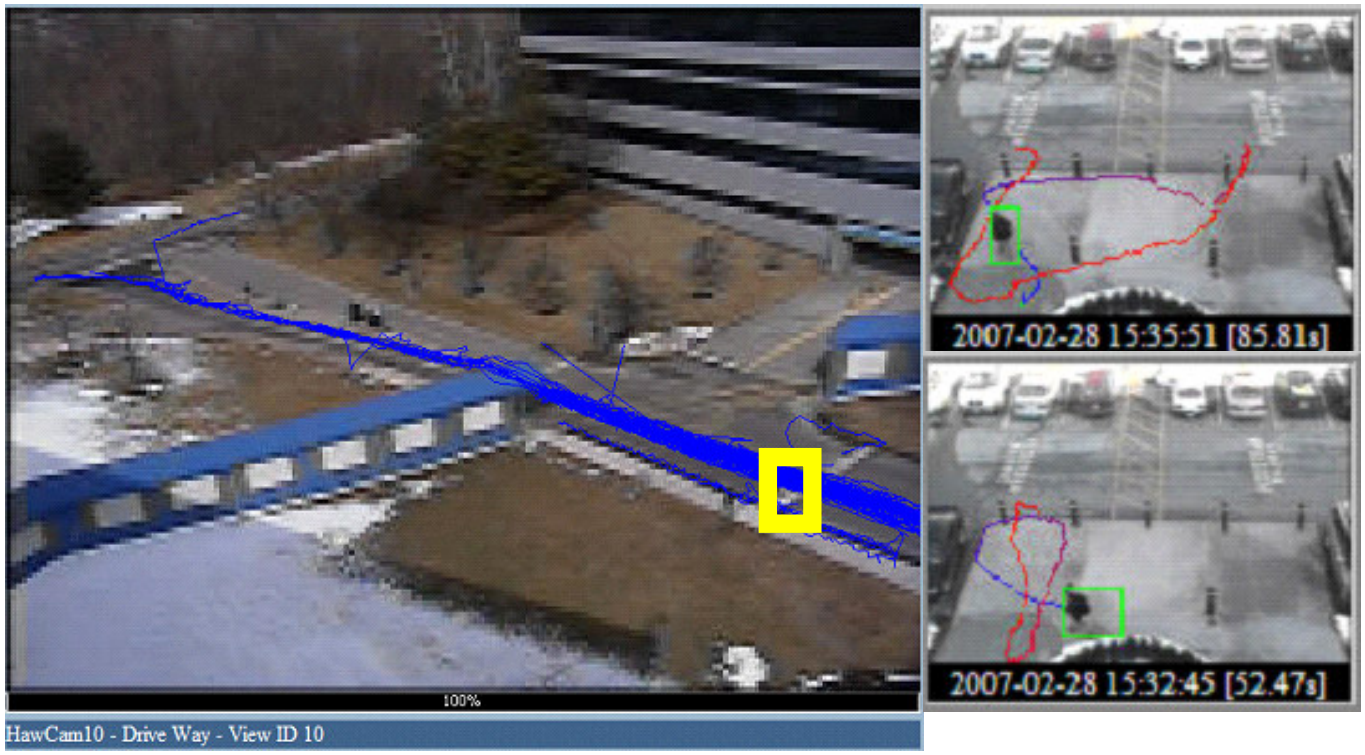


OBJECT CLASS RESULTS	Object Class Ground Truth			
		V	P	O
	V	77	0	1
	P	9	230	19
	O	1	18	53
		86	248	73

SIZE SEARCH RESULTS	SIZE SEARCH GROUND TRUTH				
	P	C	MS-T	L-T	O
P	17	2			
C	3	39			20
MS-T		3	1		
L-T				1	
O					11
Total	20	43	1	1	31

Left: Object Classification Result: V: Vehicles, P: Person, O: Other
Right: Search using object size. P: Person, C: Car, MS-T: Medium Sized Truck, L-T: Large Truck, O: Other.

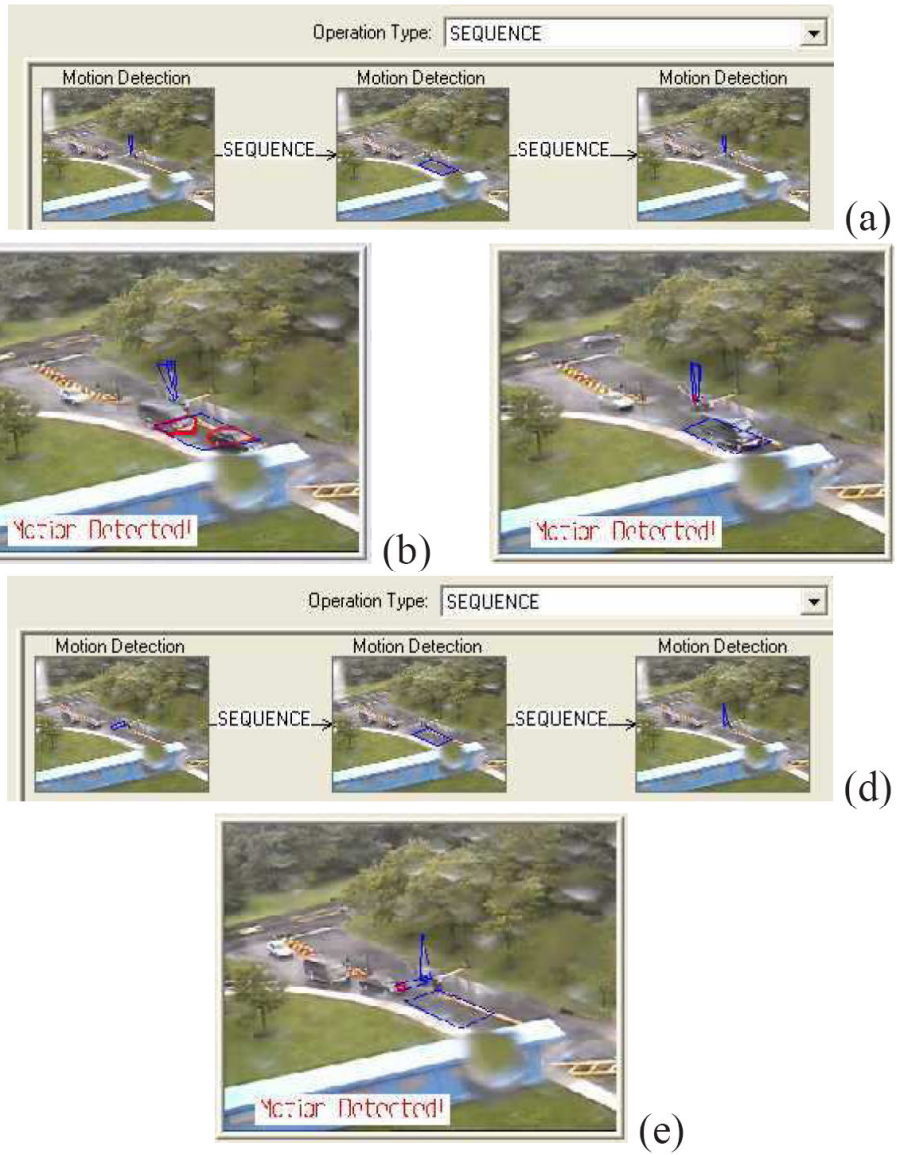
Search by Object Location and Duration



(Left) Results of spatial search, showing the trajectories of all objects that passed through the user-selected yellow region. The user can click on the trajectory to view the corresponding video clip.

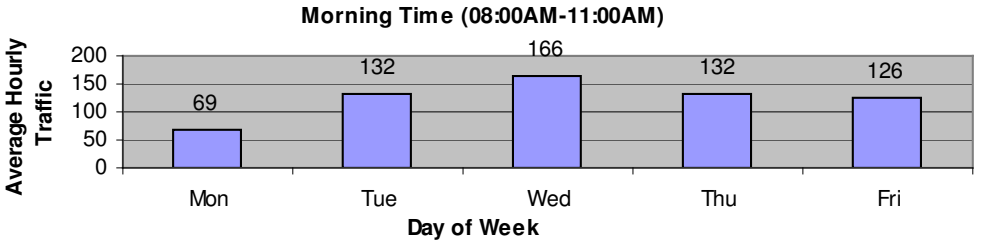
(Right) Loitering events (note the long person trajectories) retrieved by using the event duration query.

Compound Spatio Temporal Search



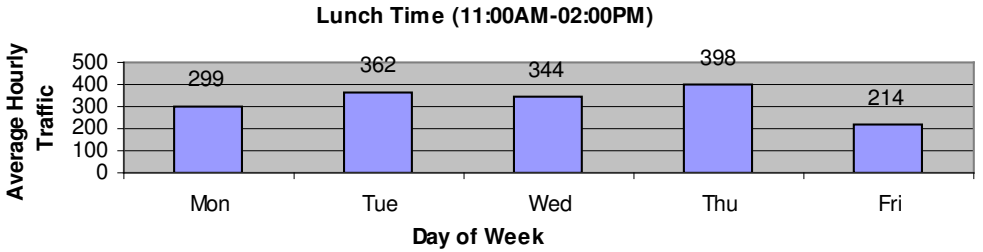
An example to detect a tailgate event at the entrance by using the spatio-temporal event detection method.

PeopleCounting

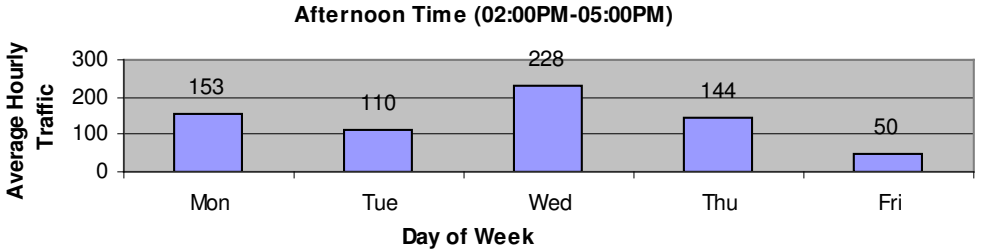


(a)

Results of people counting for a week in a cafeteria

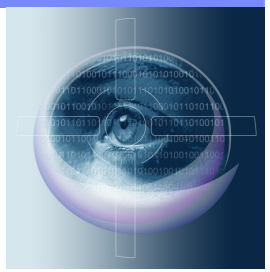


(b)



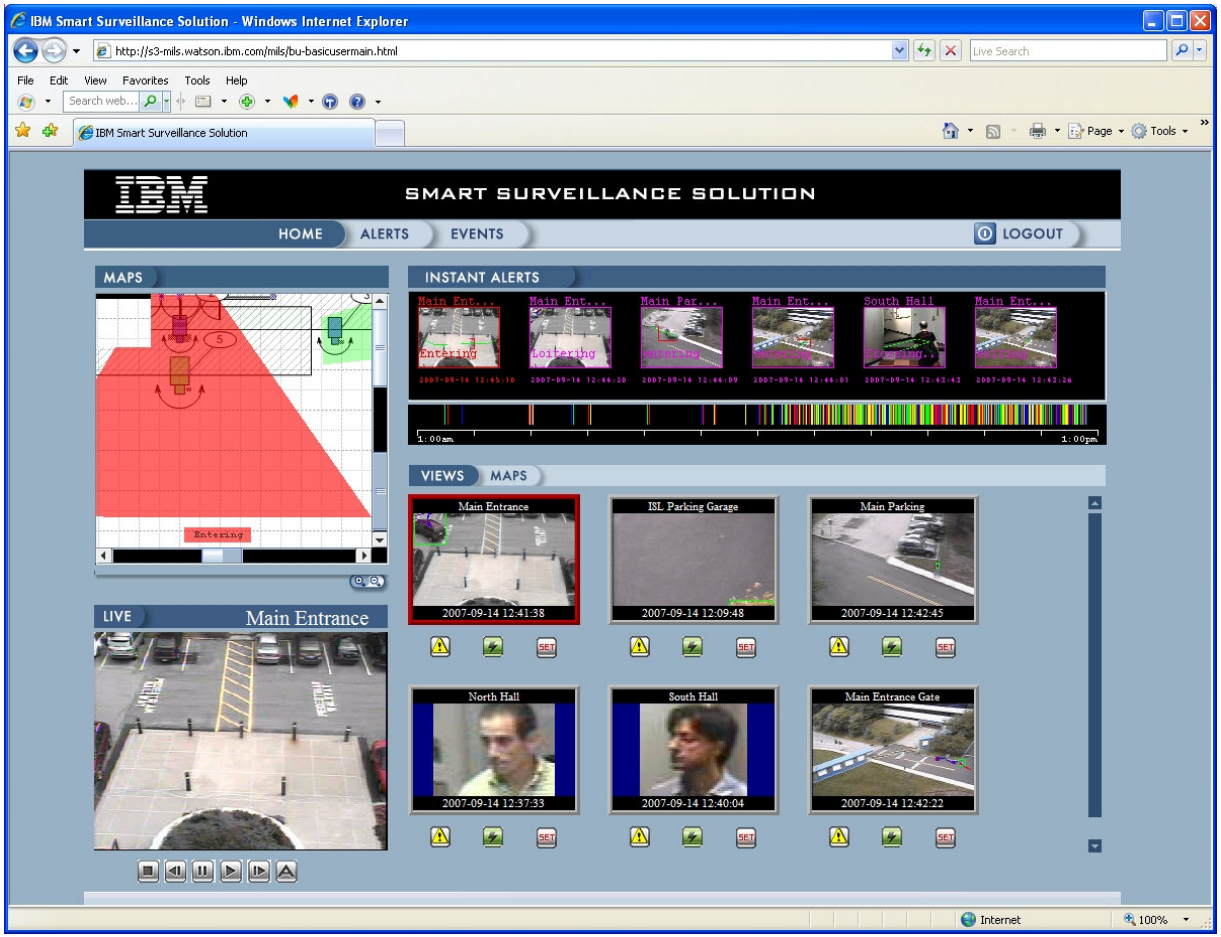
(c)

Retrieval Time Summary



Database Server, Dual Xeon, 3.8Ghz with 4GB Ram running IBM DB2	
Total number of events on Main Parking Lot Camera	From Apr 30, 07 to May 14, 07 10997 events over 15 days
Red car search	219 events retrieved in under 5secs

Interface of IBM S3 -- 1

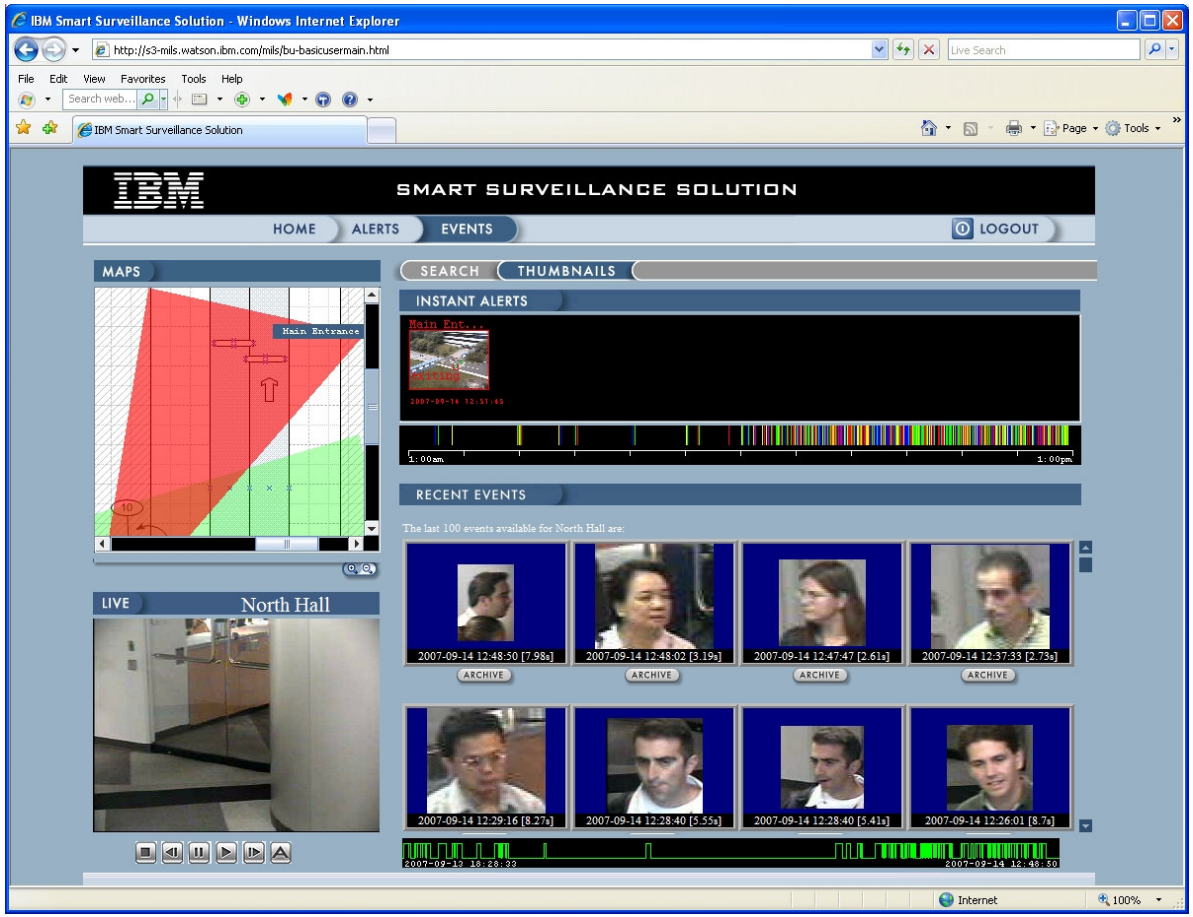


An Interface showing the various camera views currently available in the system

Interface of IBM S3 -- 2

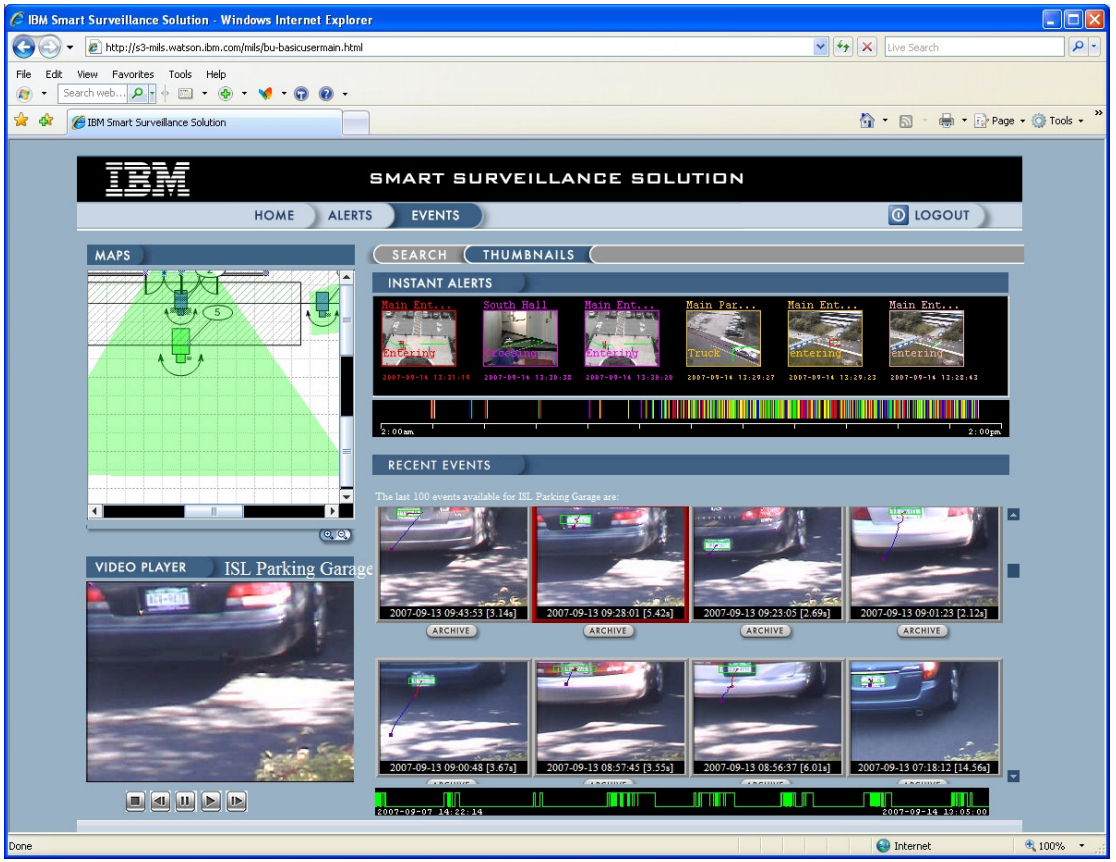
An Interface showing the Results from a "Find Person" Query

Interface of IBM S3 -- 3



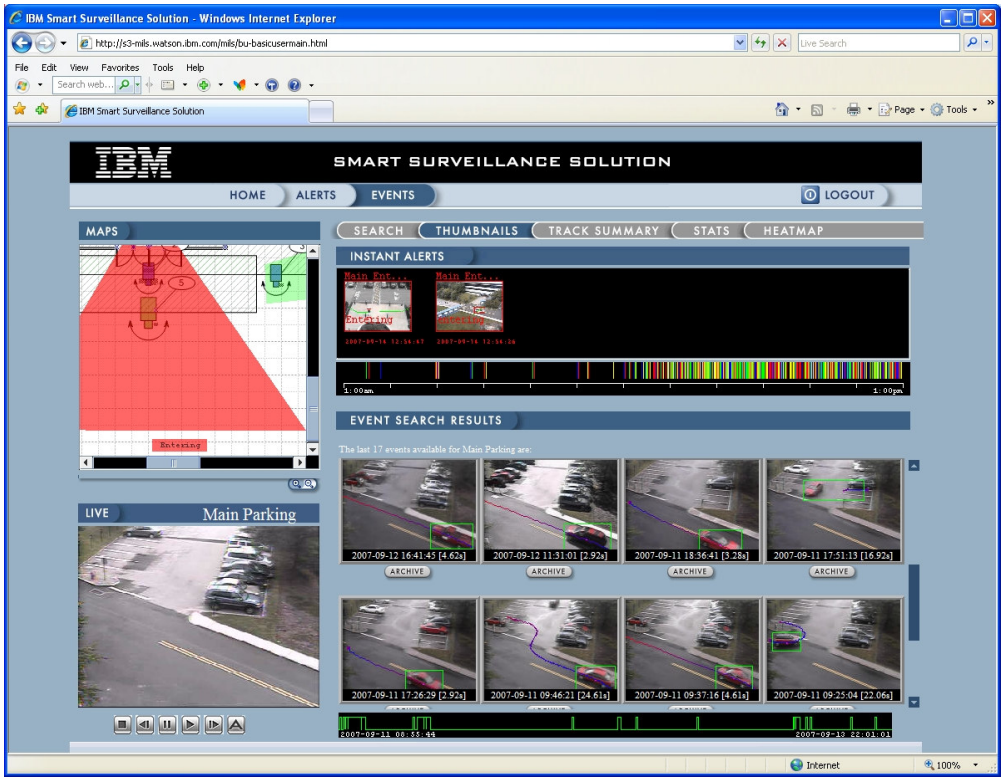
An Interface showing the Results of "Find Faces"

Interface of IBM S3 -- 4



An Interface showing the Results of "License Plate Recognition"

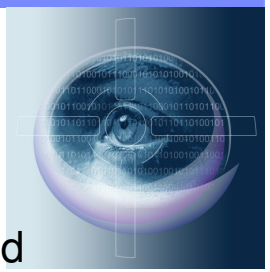
Interface of IBM S3 -- 5



An Interface showing the Results of "Red Car" search

Interface of IBM S3 -- 6

An Interface showing the track summary of one day data



References

- Berriss, W.P., Price, W.G., & Bober, M.Z. (2003) Real-Time Visual Analysis and Search Algorithms for Intelligent Video Surveillance, *International Conference on Visual Information Engineering* (pp. 226-229) July 2003.
- Arun Hampapur, Lisa Brown, Rogerio Feris, Andrew Senior, Chiao-Fe Shu, Yingli Tian, Yun Zhai, and Max Lu, "Searching Surveillance Video," IEEE International Conference on Advanced Video and Signal based Surveillance, Sept. 2007
- Meessen, J., Coulanges, M., Desurmont, X., & Delaigle, J.F., (2006) Content-Based Retrieval of Video Surveillance Scenes, Multimedia Content Representation, Classification and Security. (pp.785-792.)
- Naphade, M. & Smith, J.R. (2004) On the Detection of Semantic Concepts at TRECVID, ACM International Conference on Multimedia. (pp. 660-667.)
- Shu, C., Hampapur, A., Lu, M., Brown, L. Connell, J. Senior, A. & Tian, Y. (2005), IBM smart surveillance system (S3): a open and extensible framework for event based surveillance, *IEEE Conference on Advanced Video and Signal Based Surveillance* (pp. 318 – 323.)
- Content-Based Image Retrieval at the End of the Early Years (http://www.sci.brooklyn.cuny.edu/~sdexter/cis751/CBIREarlyage.ppt)

- Arvind Karunanidhi, David Doermann ,Survey of Content Based Access To Surveillance Video (http://www.cfar.umd.edu/~arvind/vidsurvey.htm), 2002

Project report – mid-evaluation