Multiple Instance-based Object Retrieval in Video Surveillance : Evaluation and Dataset

Cătălin Mitrea[1,2], Ionuț Mironică[1], Bogdan Ionescu[1,3], Radu Dogaru[1]

[1] LAPI & Natural Computing Labs, University "Politehnica" of Bucharest, 061071, Romania

[2] UTI GRUP, Bucharest, 020492, Romania

[3] DISI, University of Trento, 38123 Povo, Italy

Email: *catalin.mitrea@uti.ro, {imironica,bionescu}@imag.pub.ro, radu d@ieee.org*

10th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania. 4-6 Sept.

ICCP 2014

Outline

- Introduction
- Previous work
- The proposed system
- The proposed dataset (Scouter DB)
- Experimental results
- Conclusions and future work

Introduction

- High volume of video acquisition (~4mil CCTV cameras only in UK);
- Limited human resources.

Solutions

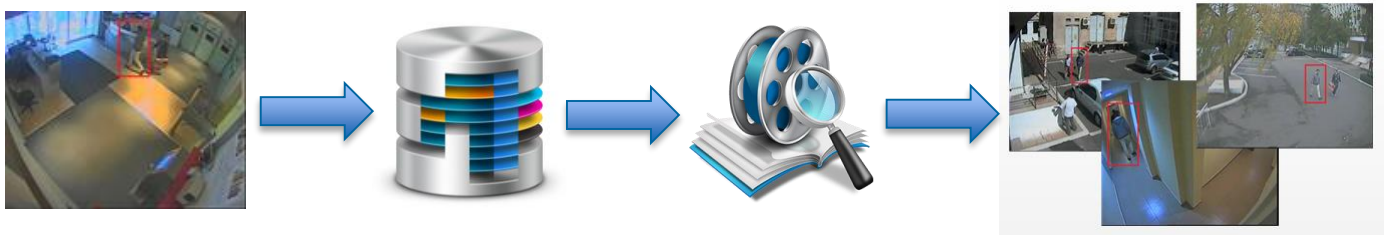
- Intelligent video surveillance techniques represent an important research domain:
  - Real-time identification and tracking of object of interest;
  - Behavior and incident detection;
  - Crowd analysis;
  - **Content-based offline searching and indexing of objects (humans).**

Research area

Introduction (cont.)

Problem statement

- Starting from a small sample (few frames) of the object to-be-found (human) => find (search) all relevant instances into a vast multisource video database.



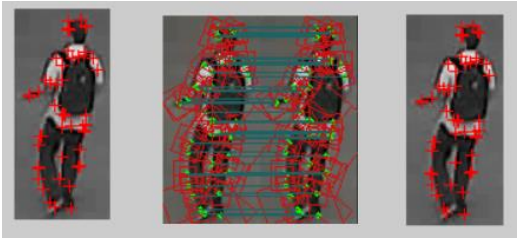
Objectives:

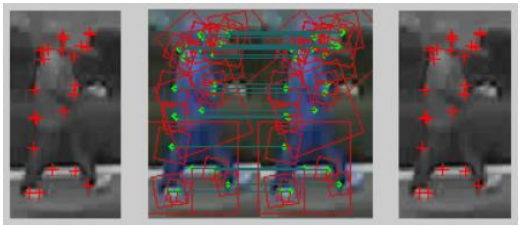
- Develop a system for providing content-based search capabilities within multi-source video surveillance footage.
- Introduce an indexed dataset containing surveillance videos recorded in a real public institution (Scouter DB).

Previous work

Main methods and directions

- Large video databases processing techniques [Snoek,IEEE 2010];
- Intuitive interfaces for video query and data mining [Shah, IEEE MultiMedia 2007];
- Content descriptors extraction (color, texture, shape, temporal and motion, audio [Ionescu, LNCS 2011]);
- Feature points (SIFT, SURF [Stottinger, IEEE 2010]);
- Dictionaries (bag-of-words and fisher kernel representations [Mironica, ACM 2013]).





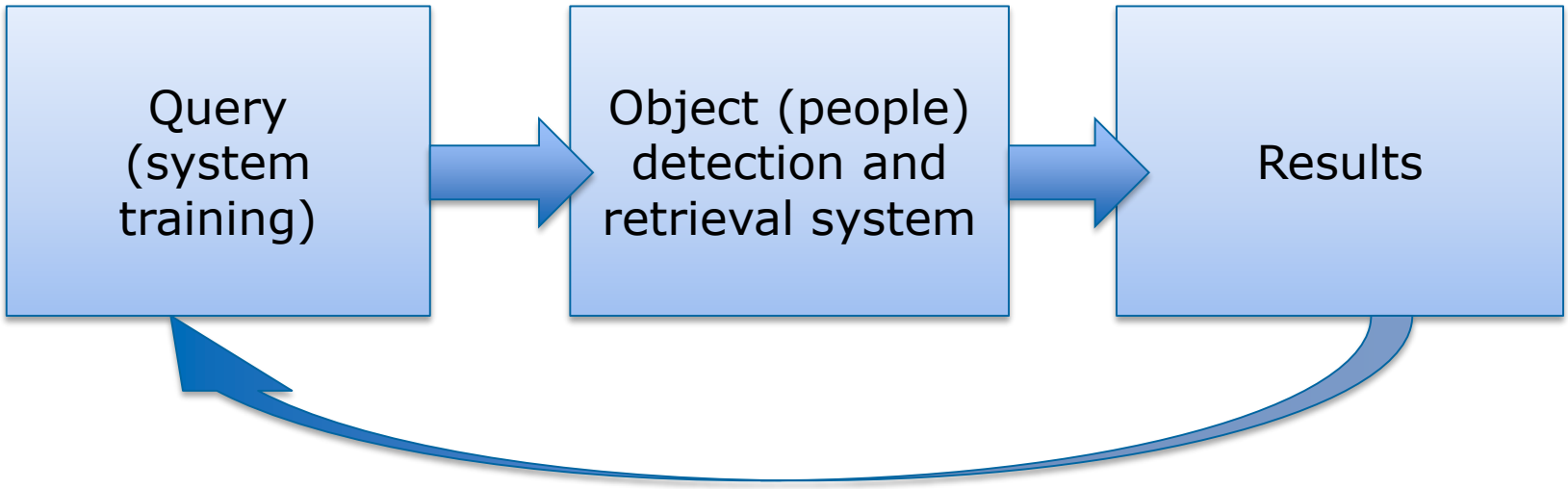
Drawbacks

- Computation complexity
- Difficult to implement for "real field" systems
- Not all methods are suitable for video surveillance datasets

perspectives - e.g., multiple source cameras, different weather conditions, different setups - e.g., indoor vs. outdoor, appearances, etc.

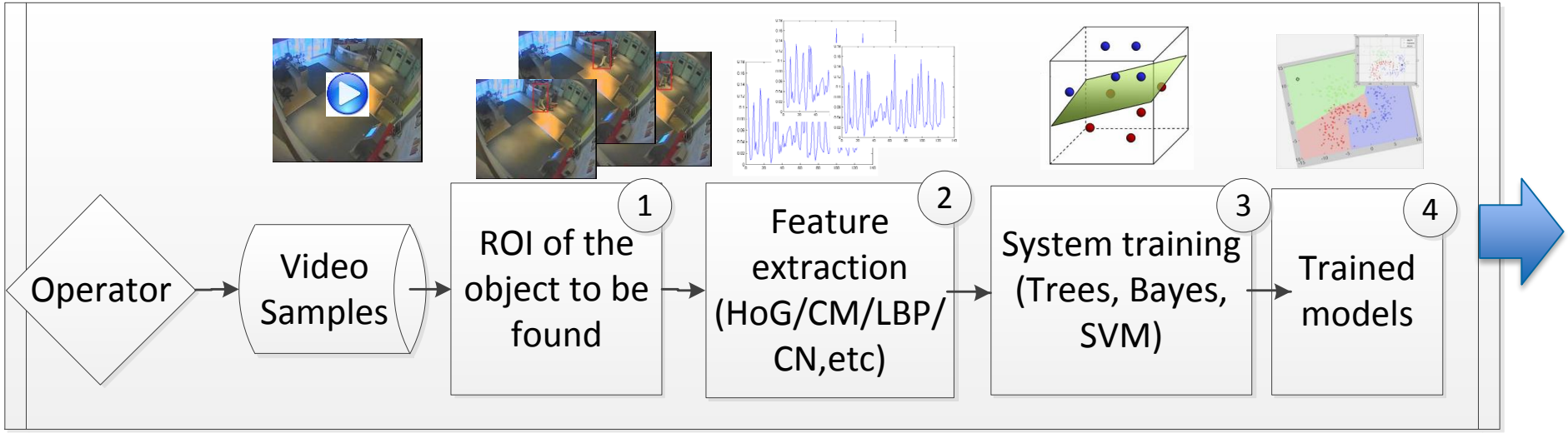
Proposed system

- Block diagram



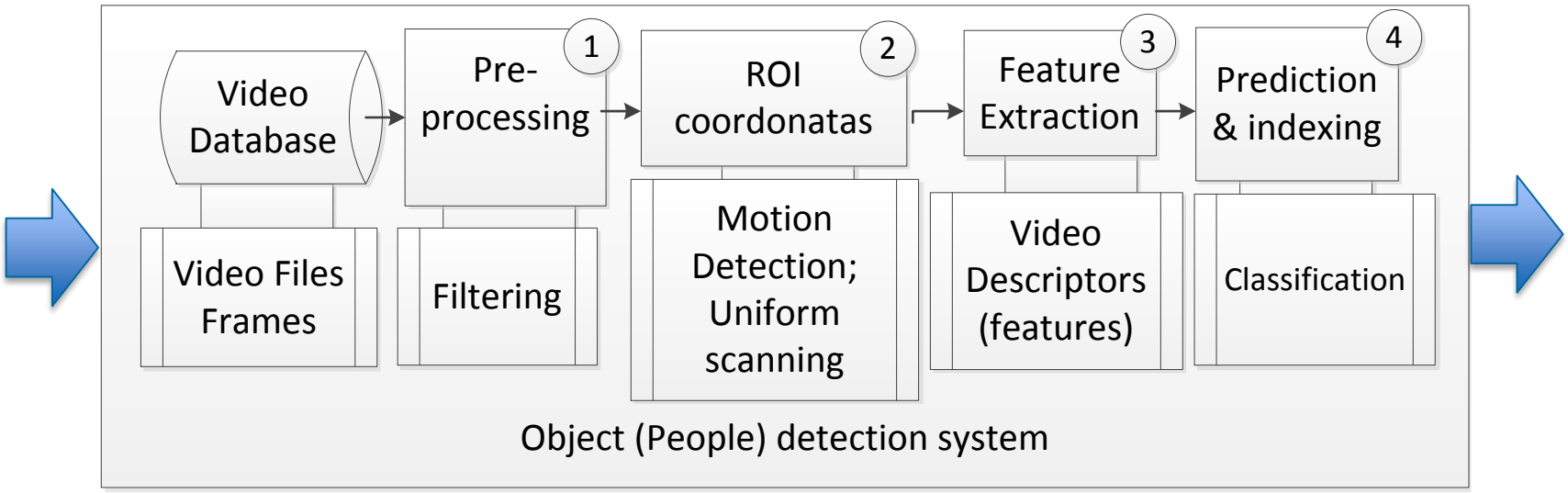
Proposed system (2)

- Query (system training)



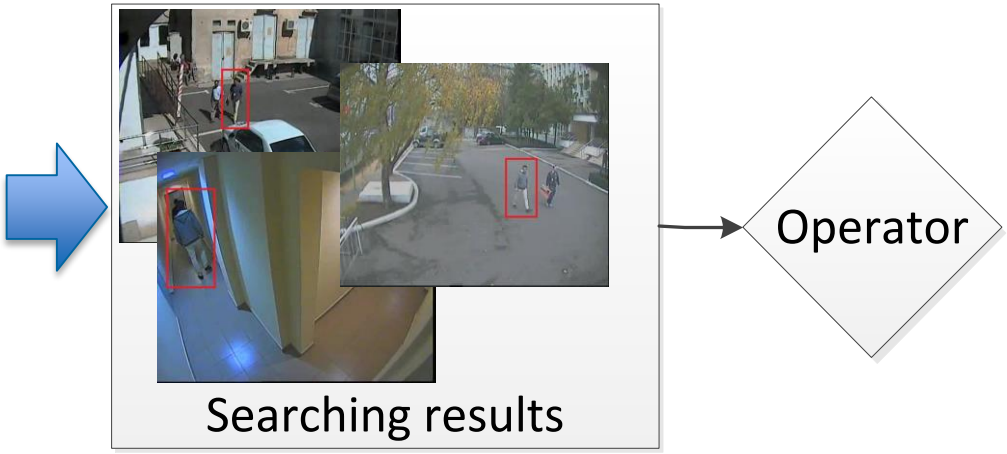
Proposed system (3)

- Object (people) detection and retrieval system



Proposed system (4)

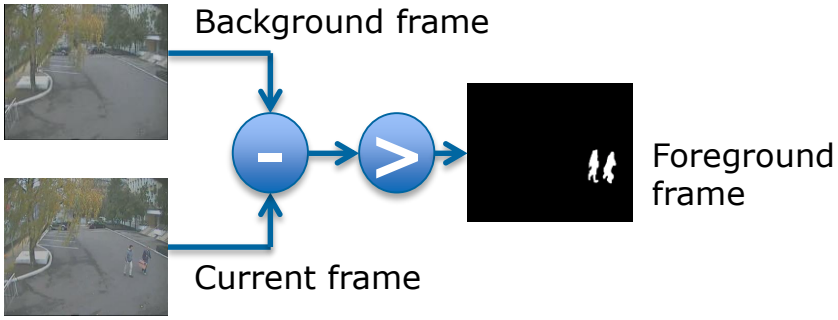
- Results



- Different instances of the object-to-be-found returned to user from the entire database

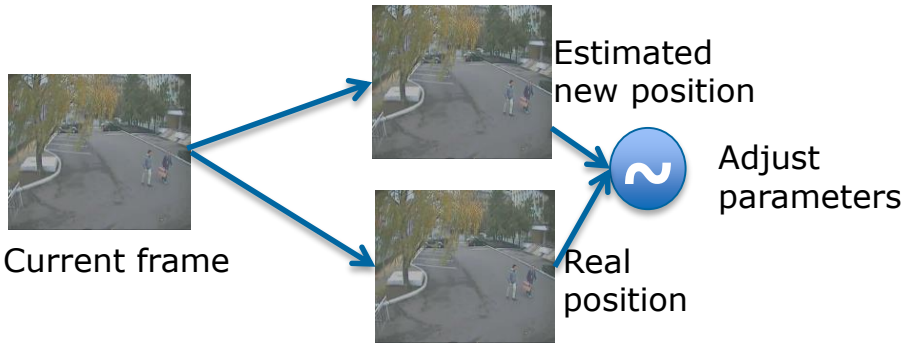
Motion detection

- Background subtraction motion detector



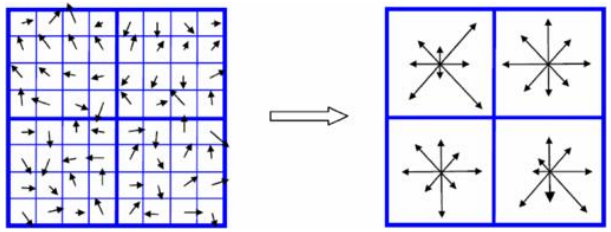
- Accumulative optical flow method

- Kalman filter-based motion detector



Content descriptors

- **HoG** features (Histogram of Oriented Graphs – shape-based descriptor, 81 values) - [Dalal, CVPR 2005]



- **CN** features (Color Naming histogram – color descriptor, 11 dimensions) [Van De Weijer, CVPR 1994]

  11 colors distribution:

  "black", "blue", "brown", "gray", "green", "orange", "pink", "purple", "red", "white" and "yellow".

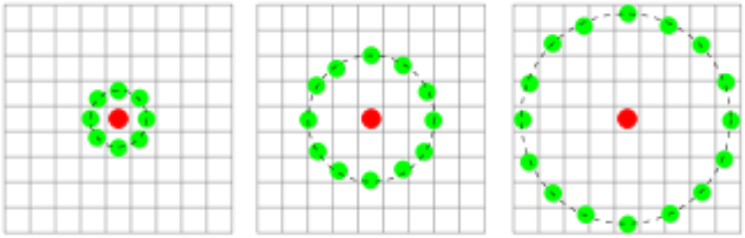
Content descriptors

- **CM** (Color moments –color descriptor, 225 dimensions) [Stricker, SPIE 1995];

Color similarity:

Three central moments of an image's color distribution: mean, standard deviation and skewness.

- **LBP** (Local Binary Pattern – texture descriptor, 256 dimensions) [Ojala, ICPR 1994]



Classifiers

- 5 classifiers used, 6 sec. of video sample with true class, 12 sec. with false class

1. Naive Bayes
2. Nearest Neighbor
3. Decision trees
4. Random forests
5. Support vector machines

The proposed dataset

- Manually indexed Scouter Database
  - 30 video files (3 different days x 10 cameras).
  - recorded at 6 to 10 fps
  - 704 x 576 resolution
  - ~36,000 annotated frames (2 people–scenarios!)
  - humans varies from 50 x 50 pixels to ~250 x 350 pixels.

frame#, frame_name, width, height, has_object?, no_of_objects, nameObj1, nameObj2, ob1_x1,ob1_y1, ob1_x2, ob1_y2, obN_x1,obN_y1, obN_x2, obN_y2

125, .\WS00_Cam0001_2013_07_25_10_53_16(1)_0125.jpg, 704, 576, 1, 2, Catalin, Daniel, 501, 163, 546, 277, 534, 153, 566, 144

126, .\WS00_Cam0001_2013_07_25_10_53_16(1)_0126.jpg, 704, 576, 1, 3, 1, 2, 3, 496, 160, 545, 278, 496, 160, 545, 278, 496, 160, 545, 278

……

283, .\WS00_Cam0001_2013_07_25_10_53_16(1)_0283.jpg, 704, 576, 0

GT file sample

	KTH	Weizmann	Scouter
Max. Resolution (W x H)	160 x120	180 x 144	704 x 576
Human Height in Pixels	80 – 100	60 - 70	50 - 350
Human to video height ratio	65 to 85%	42 to 50%	10 to 60%
Scenes Viewpoint Type	Side	Side	Varying
Natural Background Clutter	No	No	Yes
Incidental Objects/Activities	No	No	Yes
Multiple annotations on movers	No	No	Yes

The dataset was publicly released an can be downloaded at: **http://uti.eu.com/pncd-scouter/rezultate.html**

Video dataset examples

Smocking place

Main entrance turnstiles

Hallway

Main entrance reception

Evaluation

- TP - True Positives
- FP - False Positives
- TN - True Negatives
- FN - False Negatives

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F2Score = 5*Precision*Recall/(4*Precision + Recall)

Motion detectors evaluation

Motion detection algorithm	Precision	Recall
Background subtraction motion	74%	**86%**
Accumulative optical flow method	58%	55%
Kalman filter motion detector	**75%**	48%

Evaluation of the system

Recall (%)	HoG	LBP	CM	CN3x3	FUSION
1KNN	61.787	58.608	49.817	68.574	68.869
3KNN	66.19	59.561	49.881	72.254	72.144
5KNN	68.5	60.139	51.327	75.568	75.773
Decision Trees	61.473	62.55	55.822	71.865	68.127
Naïve Bayes	61.679	61.087	58.929	71.2	65.665
Random Forest	61.744	63.188	61.824	71.576	65.894
Linear SVM	63.448	65.015	61.483	69.005	71.743
SVM with RBF kernel	85.026	**92.179**	67.652	87.856	79.916
Precision (%)	HoG	LBP	CM	CN3x3	FUSION
1KNN	38.186	39.078	37.336	34.132	**46.305**
3KNN	37.981	39.244	37.116	34.186	43.811
5KNN	37.898	39.361	36.935	34.375	42.400
Decision Trees	37.23	37.526	35.321	35.128	39.718
Naïve Bayes	37.652	37.873	34.912	35.201	41.004
Random Forest	37.736	37.639	34.744	35.065	42.603
Linear SVM	37.406	36.763	34.402	35.282	39.001
SVM with RBF kernel	34.423	33.114	34.991	33.243	43.912
F2Score (%)	HoG	LBP	CM	CN3x3	FUSION
1KNN	54.99	53.28	46.70	57.06	62.75
3KNN	57.63	53.97	46.67	59.09	63.88
5KNN	58.98	54.40	47.62	60.96	65.47
Decision Trees	54.39	55.19	50.02	59.43	59.60
Naïve Bayes	54.70	54.42	51.80	59.11	58.61
Random Forest	54.77	55.64	53.49	59.24	59.40
Linear SVM	55.69	56.35	53.12	57.93	61.43
SVM with RBF kernel	65.71	67.94	57.01	66.13	**68.66**

Examples of system classification responses

Correct classification. Results and Ground Truth agrees

True Positives examples (TP)

Misclassification reasons: signal noise, illumination conditions (insufficient, over exposed), partial object view (out of frame, junction with another object) or dimension too low.

False Negatives examples (FN)

Conclusions and future work

- A labeled “real-field” video surveillance dataset is proposed to conduct the experiments.
- The classification-based approach seems a suitable perspective to solve multi-instances object retrieval (search).
- Good results are achieved in terms of recall measure using selected descriptors or their combination (fusion).

Conclusions and future work

- Drawbacks
  - The performance of the system is closely related to the number of frames and the diversity of training sample (different perspective, object size, the quality of the images)
  - The method tends to fails when too fewer samples are used for training.
- Future work
  - New types (optimized) of video descriptors (Cuboid, Hessian, MoSIFT, FK, BoW, Late fusion, etc.)
  - New classifiers (RBF-M, super fast training)
  - Co-training techniques (very few training samples)

Acknowledgement

- This work is supported by the SCOUTER project (PN-II-IN-DPST-28DPST/30.08.2013);

Thank you!