WARD-CMU @ TRECVID 2015 Surveillance Event Detection

Xingzhong Du, Xuanchong Li,

Xiaofang Zhou and Alexander Hauptmann

DKE Group, The University of Queensland

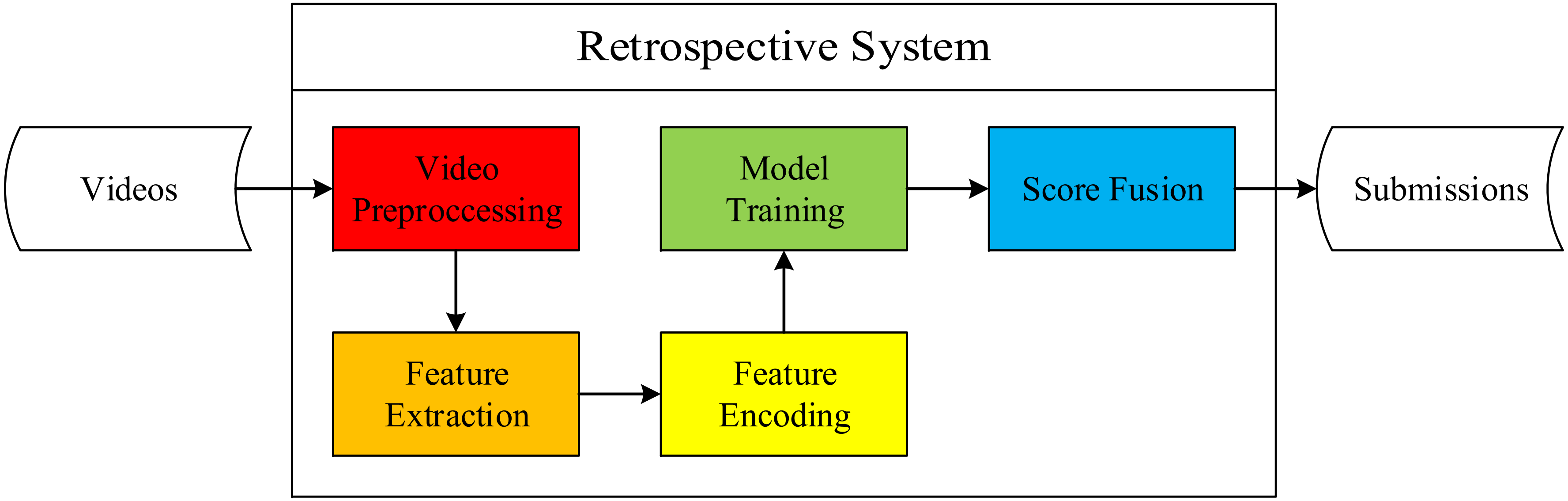
Informedia Group, Carnegie Mellon University

Outline

- Retrospective System
- Change in this year
- This year’s result

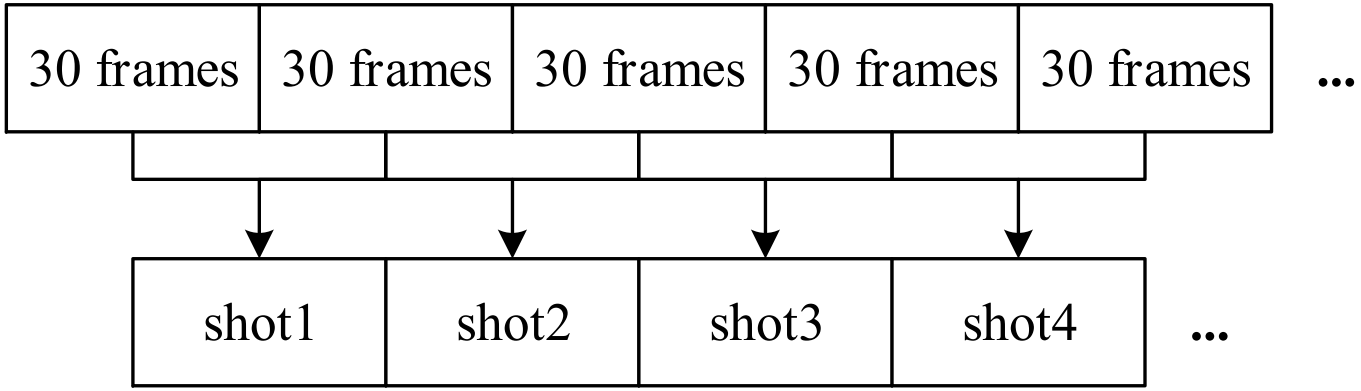
Retrospective System

Retrospective System



Video preprocessing

- Video resize
  - From 720x576 to 320x240
  - Accelerate feature extraction and encoding
  - May lose some motion information
- Video slide
  - window : 60 frames
  - stride : 30 frames



30 frames 30 frames 30 frames 30 frames 30 frames ...

shot1 shot2 shot3 shot4 ...

Feature extraction & encoding (1)

- Feature in use last year:
  - Improved Dense Trajectory (idt)
  - idt has five parts : trajectory (tra), hog, hof, MBHx, MBHy
- Encoding method:
  - Fisher vector (fv)[1]
  - Spatial-temporal information is also encoded by fisher vector (sfv)[2]

tra		hog		hof		MBHx		MBHy	
sfv	fv	sfv	fv	sfv	fv	sfv	fv	sfv	fv

[1] Perronnin, Florent, Jorge Sánchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification." Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010. 143-156.

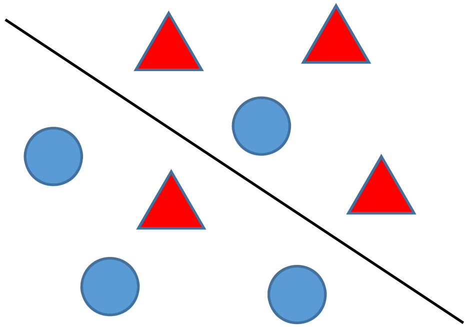
[2] Krapac, Josip, Jakob Verbeek, and Frédéric Jurie. "Modeling spatial layout with fisher vectors for image categorization." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011

Feature extraction & encoding (2)

- Learn PCA
  - Dimension reduction (to half)
  - Make the co-variance matrix be diagonal
- Learn GMM
  - 256 components
- Calculate the derivatives with regards to the means and variances then concatenate them into vector
- Normalization
  - Power normalization
  - L2 normalization

Model Learning (1)

- Event detection as one-vs-all classification



- One model per event and camera
- Positive and Negative
  - Get the event spans from the annotation files
  - The video shots whose middle frames locate in the event spans are positive
  - The other video shots are negative

Model Learning (2)

- The dimension of the fisher vector in use is 116736
  - Each vector costs 456KB
- If we use LIBSVM:
  - Each model has around 8000 support vectors
  - Each model costs around 3.65 GB
- Using LIBLINEAR instead:
  - Each model only contains the *weights* and *bias*
  - Each model costs around 456KB
- So we use LIBLINEAR in the retrospective system

Model Learning (3)

- However, LIBLINEAR in python does not support probability output

- In last year system, we estimate the probability distribution of decision values by curve fitting

$$P(x) = e^{-(Ax+B)}$$

x is the decision value, *A* and *B* are the parameters need to learn by curve fitting.

- The decision values in use are from the training data after the model is obtained.

Score Fusion

- Last year the final system has three features in total:
  - Improved Dense Trajectory
  - STIP
  - MoSIFT
- Each feature provides a ranking list, we fuse them into one list by average fusion

Change in this year

Feature in Use

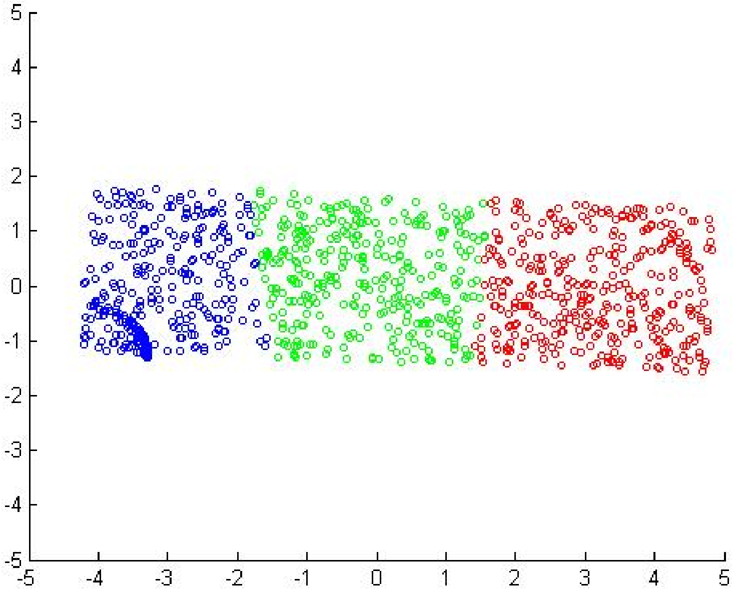
- Dense Trajectory (dt)
  - Do not warp the dominant motion between the adjacent frames
  - Fit for event detection where only several persons appear in the surveillance

- Improved Dense Trajectory (idt)
  - Warp the dominant motion between the adjacent frames
  - Fit for event detection where a crowd of persons appear in the surveillance

We think they are complementary features in surveillance event detection

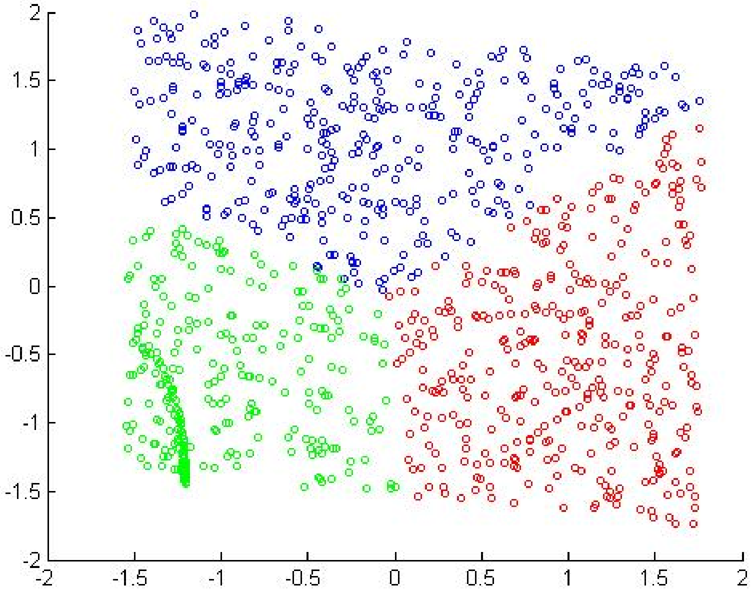
Improved Encoding

- PCA



Space is quantized by the high variance components. Soft assignment maybe degrades to hard assignment

- whiten PCA



Space is quantized by the equal variance components. Soft assignment often works

Probability Estimation by Cross-validation

$$P(x) = e^{-(Ax+B)}$$

- Using the decision values for the training data to learn A and B causes overfitting
- Improve overfitting by 5-fold cross validation
  - Use 4-fold to train classifier
  - Get the decision values for the rest fold
  - After each fold gets the decision value, get A and B by curve fitting

Fusion on selected features

- We get four ranking lists before the submission
  - dt-fv : ranking list based on dense trajectory with normal PCA
  - Idt-fv : ranking list based on improved dense trajectory with normal PCA
  - dt-wfv : ranking list based on dense trajectory with whiten PCA
  - idt-wfv : ranking list based on improved dense trajectory with whiten PCA
- After fusing any combination of ranking list and evaluating, we found average fusing dt-wfv and idt-wfv is the best

This year’s result

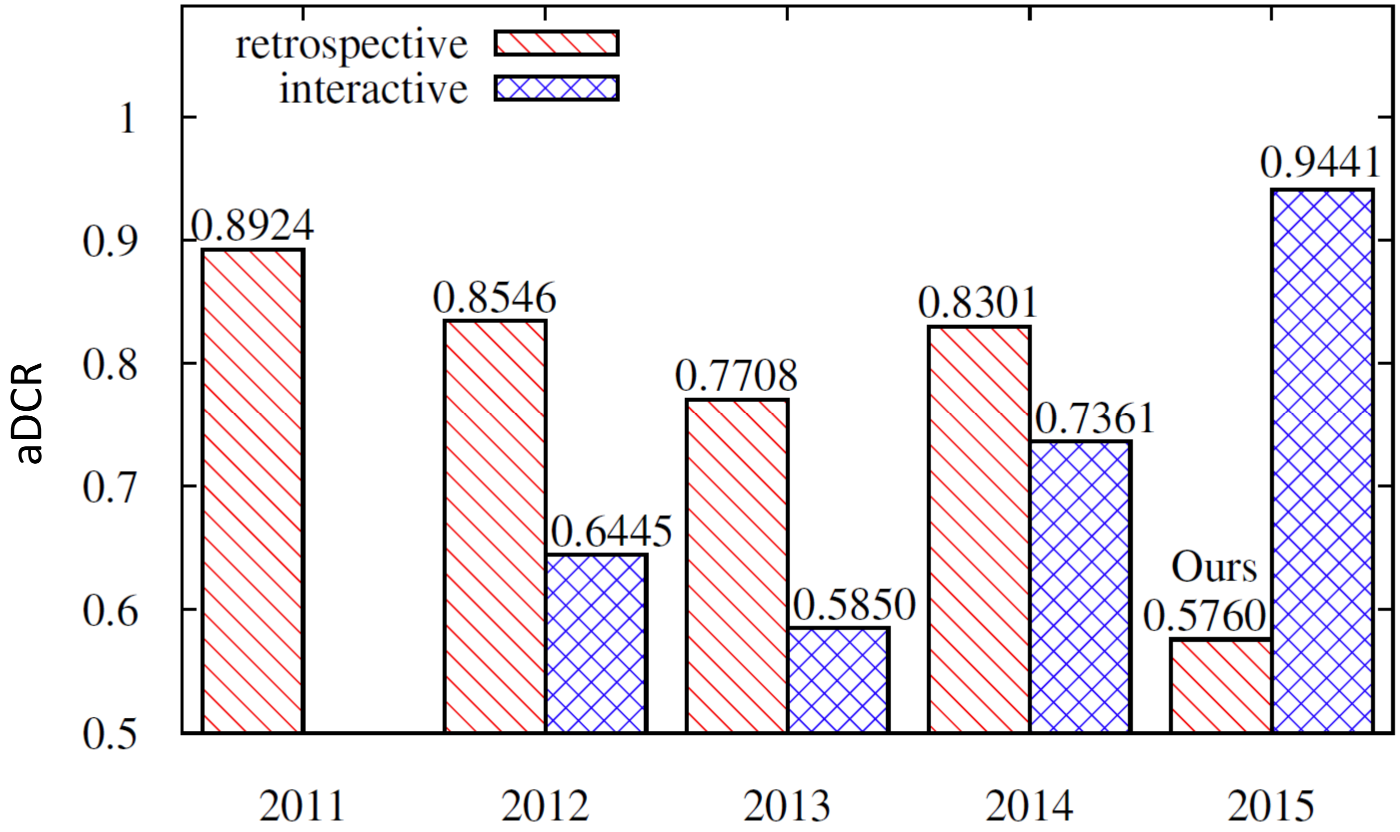
Comparison with the best results

Event	Our retro results		Others' retro results		Others' inter results	
	aDCR	mDCR	aDCR	mDCR	aDCR	mDCR
CellToEar	1.0046	1.0006	1.3071	1.0006	2.1010	1.0006
Embrace	0.8680	0.8453	0.7909	0.7909	0.8540	0.8540
ObjectPut	1.0160	0.9884	1.0120	0.9965	0.9930	0.9867
PeopleMeet	0.8939	0.8848	1.0426	0.9981	0.9978	0.9919
PeopleSplitUp	0.8934	0.8785	0.9387	0.9253	0.9164	0.9164
PersonRuns	0.5768	0.5466	0.9700	0.9545	0.9411	0.9411
Pointing	1.0140	0.9940	1.0040	0.9989	0.9939	0.9939

retro = retrospective, inter = interactive, aDCR = actual DCR, mDCR = minimum DCR

In total, we won 4 events in this year's competition.

PersonRuns gets a new record



- With this year’s retrospective system, the automatic detection for PersonRuns reaches a new level, which is better than previous interactive results.

Thank you