Incorporating Object and People Information to Improve Video Activity Recognition

Niranjan Viladkar
Under the guidance from

Dr. Parag Singla
Department of Computer Science, IIT Delhi

M.Tech Project Presentation – June, 2014

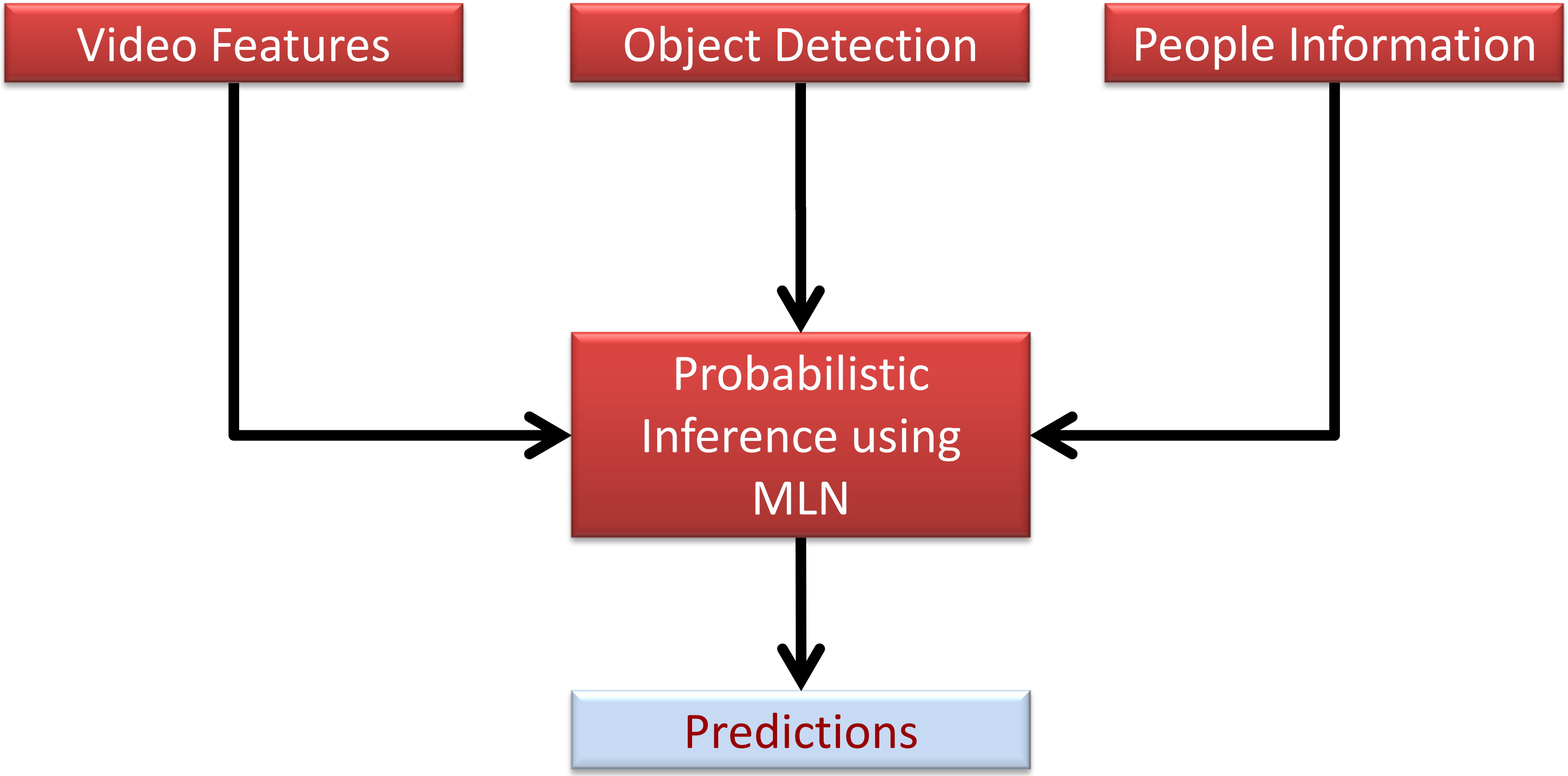
Outline

- Problem Definition
- Background
  - Video recognizer, Object Detector, MLN
- Approach
  - MLN Model details
- Results

Problem Definition

- Existing Frameworks
  - Pure HoGHoF features (Laptev, CVPR'08)
  - HoGHoF + Scene context (Laptev, CVPR'09)
  - HoGHoF + Object detection(Mooney, ECAI'12)
- Problem with low level features
  - Partial or full occlusion
  - Noisy training data
- To capture semantic relationship between Activity and object & people information.

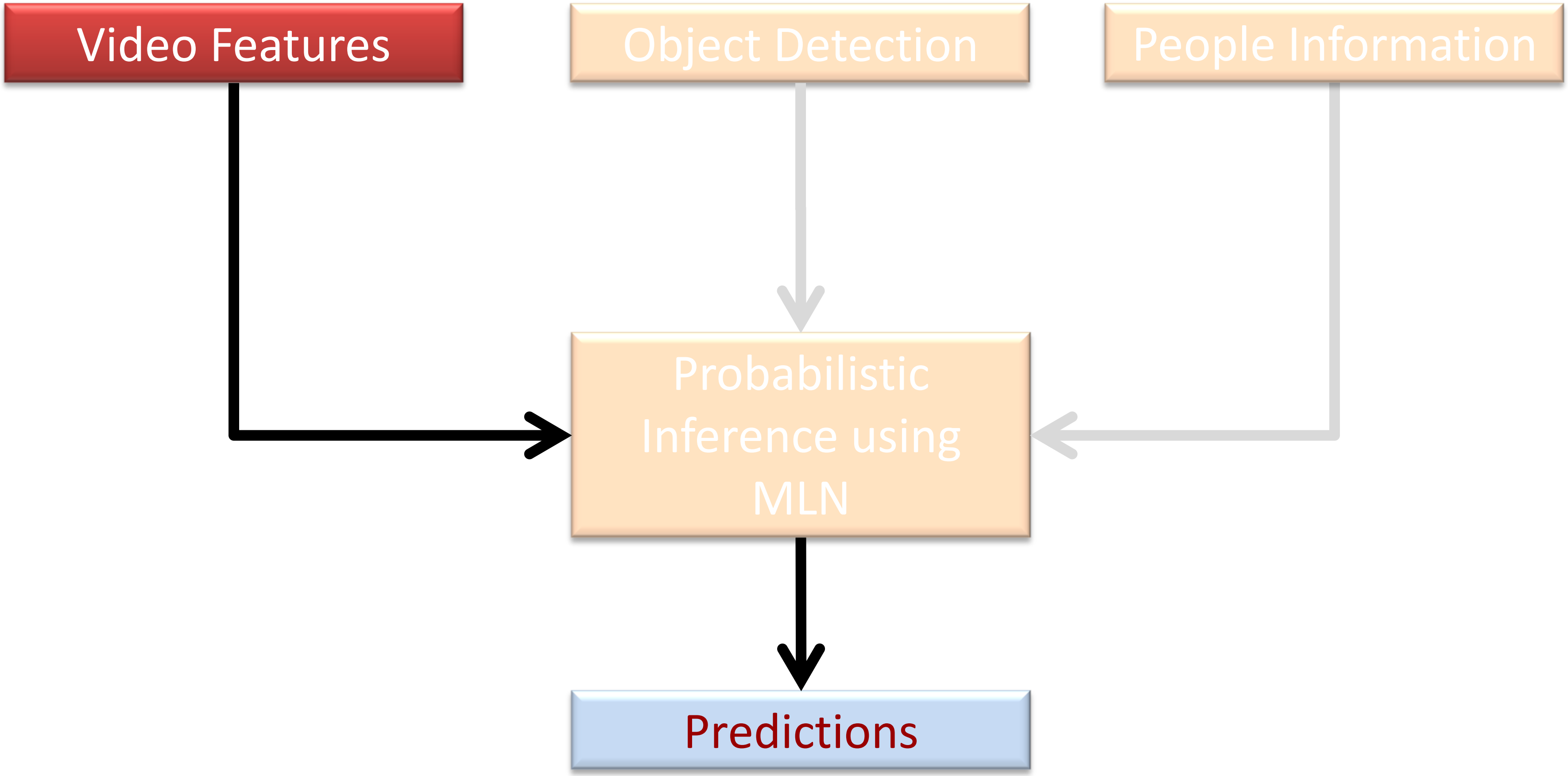
Background - schematic



Background – Data set

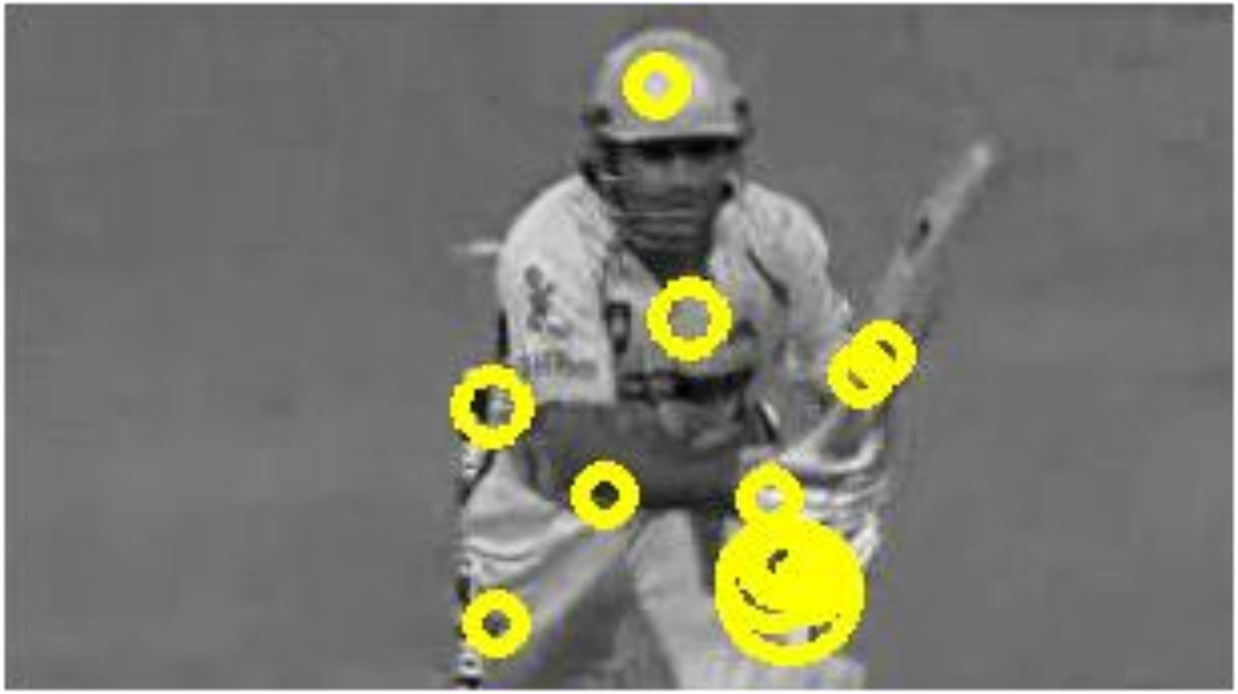
- Hollywood2 data set
- 823 Training and 884 Testing video clips
- 12 Activity Classes
- Labeled data

Background - schematic



Background - Video Features

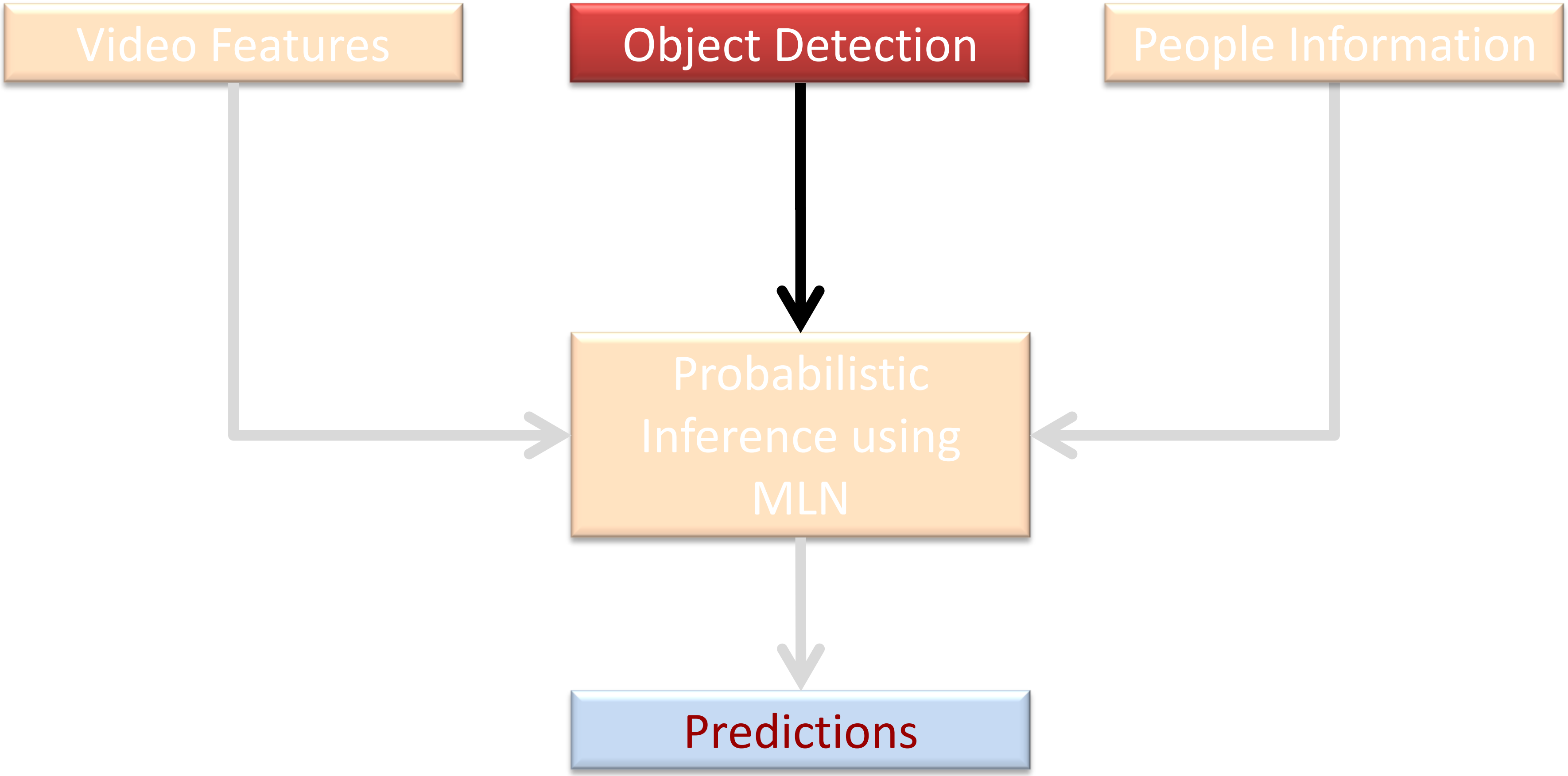
- HoGHoF features at STIPs.



Background - Video Features

- HoGHoF features at STIPs.
- Describe a video clip as bag-of-features.
  - Cluster all HoGHoF feature descriptors using k-means.
  - Represent clip as a histogram over these clusters
- Train a SVM classifier
  - Supervised - Dataset is pre labeled
  - Output – For each clip, confidence value for all 12 activities.

Background - schematic



Background – Object Detection

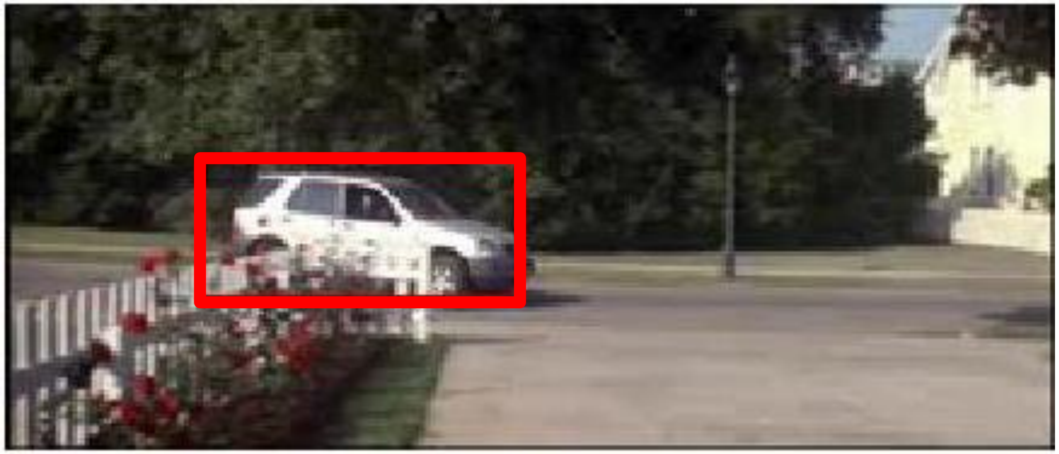
- Using Discriminatively Trained Deformable Part Models (Felzenszwalb - PAMI'10)
  - Pre-trained object detector for 20 objects
  - aeroplane, **bicycle**, bird, boat, **bottle**, **bus**, **car**, cat, **chair**, cow, **dining table**, dog, horse, **motorbike**, potted plant, **person**, sheep, **sofa**, train, **tv monitor**

– Application to videos

– **Output** – Confidence of object being present in selected frames from the video

Output of Object Detector

- FRAME 1
  car -0.181786



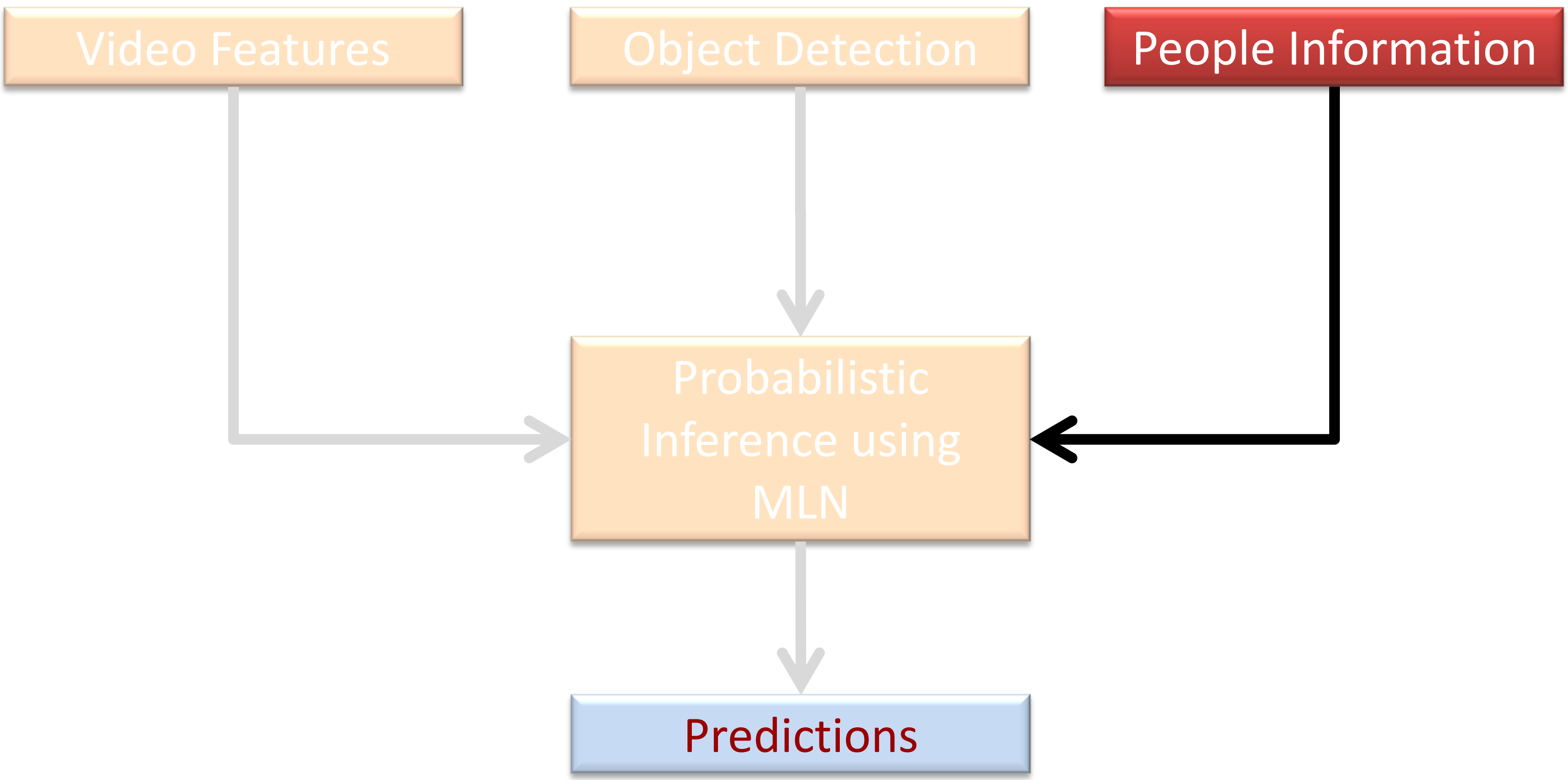
.

.

.

- FRAME 151
  person 0.579786
  person -0.593087

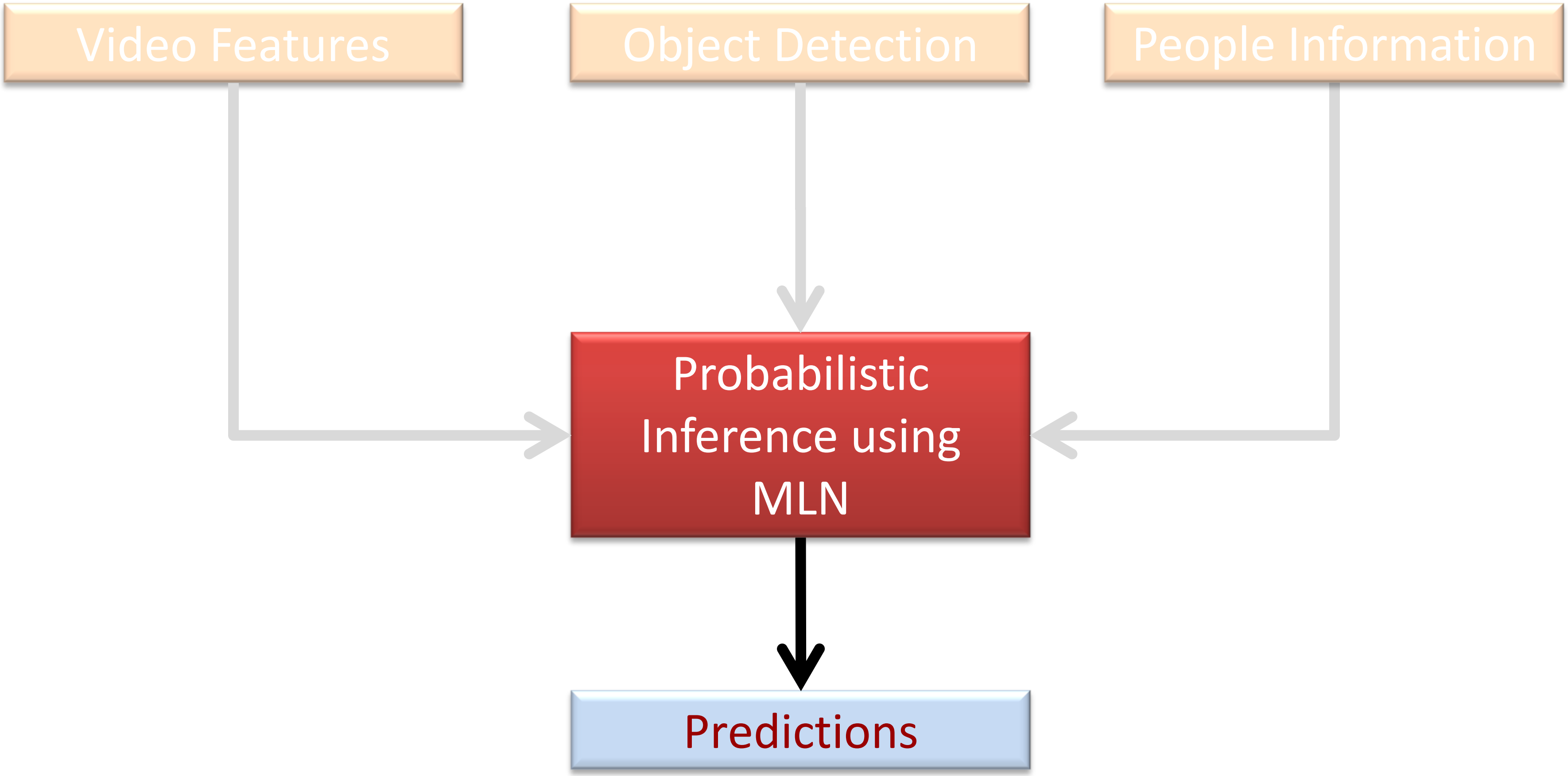
Background - schematic



Background – People Information

- Object “Person”
- Average number of persons per frame

Background - schematic



Background - Inference Using MLN

- Undirected Graphical models to represent the joint distribution of a set of random variables.
- Graph has a node for each variable, and the model has a potential function for each clique in the graph.

$$P(X = x) = \frac{1}{Z} \prod_k \phi_k(x_{\{k\}})$$

$$Z = \sum_{x \in X} \prod_k \phi_k(x_{\{k\}})$$

$$P(X = x) = \frac{1}{Z} exp \left(\sum_i w_i f_i(x) \right)$$

Background - Inference Using MLN

- Framework to form Markov networks
- MLN :- A set of pairs (F_i, w_i) and set of constants C.
- F_i :- formula in first-order logic
- w_i :- real valued weight
- Contains binary node for each possible grounding.

Background - Inference Using MLN

- Probability calculated as :-

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{F} w_i n_i(x)\right)$$

- F : - number of formulas in the MLN.
- $n_i(x)$:- is the number of true groundings.

Approach

- MLN Evidence
  - Partitioning confidence into bins
  - Threshold for object detection confidence
  - True labels from data set

HasActivity("actioncliptrain00001","SitUp")

ActivityConf_N2_TO_N15("actioncliptrain00001","AnswerPhone")

ActivityConf_N15_TO_N1("actioncliptrain00001","DriveCar")

ObjPresent("actioncliptrain00001","person")

NumPersons_1_TO_15("actioncliptrain00001")

- MLN Query made on “HasActivity”

Approach

- MLN Rules
  - Positive and Negative

ActivityConf_P1_TO_P15(c,a) => HasActivity(c,a)

ActivityConf_P15_TO_P2(c,a) => HasActivity(c,a)

ObjPresent(c,“chair") => HasActivity(c,"Eat")

ObjPresent(c,“car") => HasActivity(c,“DriveCar")

ObjPresent(c,"bus") => HasActivity(c,“StandUp")

ObjPresent(c,"car") => HasActivity(c,“HandShake")

NumPersons_1_TO_15(c) => HasActivity(c,+a)

NumPersons_15_TO_2(c) => HasActivity(c,+a)

Approach

- TF-IDF features
  - Appending 10 tf-idf features – one per object
  - And 1 feature corresponding to number of people

Results

- MLN Experiments

Activity Class	SVM	MLN			
		Only Action	Action & Object	Action & People	Action Object & People
AnswerPhone	11.36%	10.64%	11.11%	11.67%	12.73%
DriveCar	66.96%	66.06%	66.67%	71.57%	68.18%
Eat	45.45%	32.50%	40.00%	35.00%	40.00%
FightPerson	57.63%	56.90%	54.84%	61.54%	62.26%
GetOutCar	17.86%	8.00%	13.79%	17.39%	14.29%
HandShake	25.93%	21.43%	25.00%	30.77%	41.67%
HugPerson	15.15%	15.79%	13.79%	14.29%	16.13%
Kiss	18.18%	18.07%	19.78%	19.79%	20.65%
Run	38.78%	36.42%	41.48%	40.32%	42.15%
SitDown	40.96%	38.10%	35.56%	34.78%	39.56%
SitUp	5.26%	0.00%	5.26%	0.00%	12.50%
StandUp	35.20%	38.46%	36.29%	38.26%	36.24%
AAP	31.56%	28.53%	30.30%	31.28%	33.86%

Results

- TF-IDF Experiments

Activity Class	AP - Basic SVM	AP - Object and People
AnswerPhone	11.36%	16.67%
DriveCar	66.96%	70.09%
Eat	45.45%	50.00%
FightPerson	57.63%	66.04%
GetOutCar	17.86%	12.12%
HandShake	25.93%	31.82%
HugPerson	15.15%	17.86%
Kiss	18.18%	20.69%
Run	38.78%	39.35%
SitDown	40.96%	42.05%
SitUp	5.26%	8.70%
StandUp	35.20%	37.88%
Average AP	31.56%	34.44%

References

- *Actions in Context* by M. Marszalek, I. Laptev and C. Schmid; in Proc. CVPR-2009
- *Improving Video Activity Recognition using Object Recognition and Text Mining* by Tanvi Motwani and Raymond J. Mooney, ECAI-2012
- *Markov Logic* by Pedro Domingos, Parag Singla, et.al., Probabilistic Inductive Logic Programming (pp. 92-117), 2008. New York: Springer.
- *Learning realistic human actions from movies* by Laptev et.al., Conference on Computer Vision & Pattern Recognition, Jun 2008.
- *Object Detection with Discriminatively Trained Part-Based Models* by Pedro F. Felzenszwalb, et.al., PAMI-2010