IST- 2001- 37540
CAVIAR

WP1 – Active Image Acquisition and description

plus …. quite a bit of WP4 and WP5.

Pedro Ribeiro, Alexandre Bernardino, Jonas Hornstein, Plini Moreno, José Santos-Victor

Instituto Superior Técnico,
Institute for Systems and Robotics,
Lisbon, Portugal
{pribeiro, jasv}@isr.ist.utl.pt

WORKPACKAGE 1

T1.1 Integrate the Vision System with Foveated Cameras

T1.2 Develop the feature extraction algorithms

T1.3 Implement Spatial and Temporal Grouping strategies

…..

Feature evaluation/selection for short –term activity recognition. (WP1 + WP4 + WP5)

Scientific questions

Q1: What are good features for video-based human behaviour understanding? (WP1+WP4)

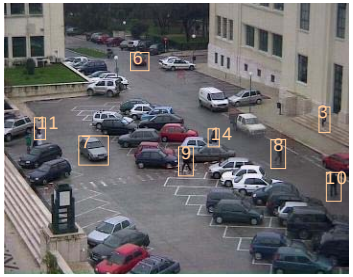
Q4: The effect of the temporal window length for describing movement (WP1+WP4)

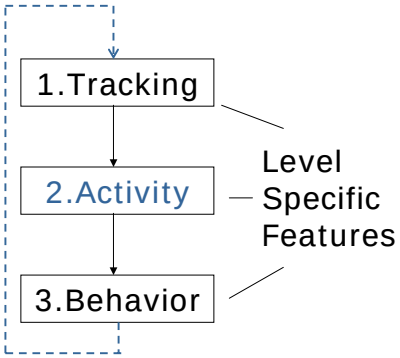
Q2: Is there a benefit to using Gabor versus Gaussian derivative filters for feature extraction? (WP1)

Outline

1. problem statement
   a. scenario
   b. framework
2. low-level activities & features
   a. activities definition
   b_1. trajectory based features
   b_2. blob based features
3. feature selection & classification
   . exhaustive search Vs suboptimal search
   . classifier
4. classifier structure
5. conclusions & future work

1. problem statement





1. Find ROIs at low resolution. (Tracking)
2. Process ROIs at full resolution. (Activity recognition)
3. Model ROIs interactions. (Behavior interpretation)

1.a problem statement

scenario

Attention on isolated people

walking
...
fighting
...
running
...

1.a problem statement

scenario

Tracking

Ground truth Bounding Box ("the perfect tracking")

Target reference point

Bounding box centroid, backprojected on the floor (x,y)

Target detected pixels

From background subtraction (within the Bound. Box)

Activities

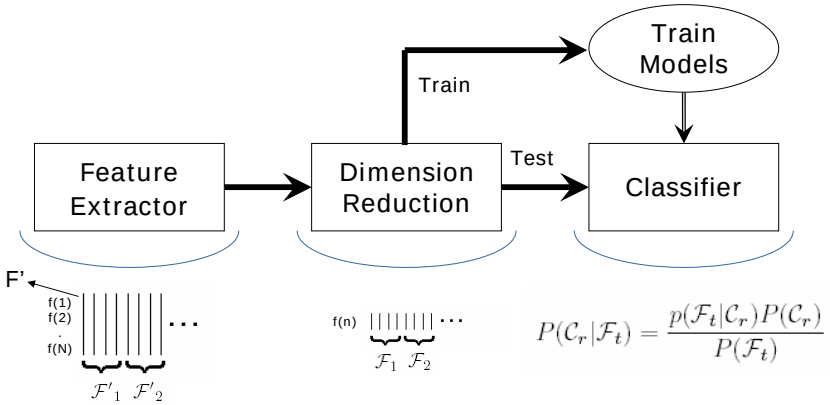
Caviar ground truth (revised)

From CAVIAR project:

Used a total of 15907 targets (persons)

1.b problem statement

framework



Features set

Selection
Compression

Bayesian
framework

2. low-level activities & features

2.a low-level activities & features

Low-level activities:

#	# frames	Activity	Description
1	3211	inactive	a static person/object
2	1974	active	person making movements but without translating in the image.
3	9831	Walking	there are movements and overall image translation
4	297	Running	as in *walking* but with larger translation.
5	594	Fighting	large quantities of movement with few translation.

Table 1: Low-level activities and data distribution.

Detection:

from short video sequences (few seconds: 3-5)

2. low-level activities & features
(recap form last year…)

Perspective normalization $[XYw]^T = H\,[xy1]^T$

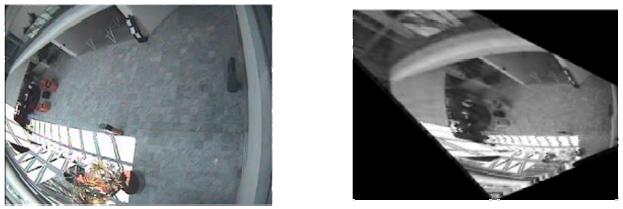


Figure 1: Original (left) and resulting transformed (right) images of the INRIA - Caviar scenario. Perspective distortion is removed by mapping the ground plane using the estimated homography.

2.b$_1$ low-level activities & features

Trajectory based features

Reference:

target reference (X,Y)

Features extracted coding:

velocity (norm)
speed
trajectory regularity
temporal energy …

A total of 16 features

2.b$_1$ low-level activities & features

Trajectory based features:

Instantaneous

Averaged

Energies

#	NAME	DEFINITION
		Instantaneous
-	position	$\vec{p}(t) = (X(t), Y(t))$
-	velocity	$\vec{v}(t) = \frac{d\vec{p}}{dt} = (v_x(t), v_y(t)) \simeq (X(t) - X(t-1), Y(t) - Y(t-1))$
1	speed	$v(t) = \|\vec{v}(t)\| = \sqrt{v_x^2(t) + v_y^2(t)}$
		Time averaged over T frames
2	mean speed	$\bar{v}_T(t) = \frac{1}{T}\sum_{i=t-T+1}^{t} v(i)$
-	mean velocity norm	$\|\bar{\vec{v}}_T(t)\| \Rightarrow$ Use one of 4 different methods:
3	- averaging vectors	$\|\bar{\vec{v}}_T(t)\|_1 = \|\frac{1}{T}\sum_{i=t-T}^{t-1} \frac{\vec{p}(t) - \vec{p}(i)}{t-i}\|$
4	- mean vector	$\|\bar{\vec{v}}_T(t)\|_2 = \|\frac{1}{T}\sum_{i=t-T+1}^{t} \vec{v}(i) = \frac{\vec{p}(t) - \vec{p}(t-T+1)}{T}\|$
5	- linear fitting	$\|\bar{\vec{v}}_T(t)\|_3 =$ Linear Least Squares Fitting
6	- quadratic fitting	$\|\bar{\vec{v}}_T(t)\|_4 =$ Quadratic Least Squares Fitting
7..10	speed/velocity ratio	$R_{\bar{v}_T}(t)_i = \frac{\bar{v}_T(t)}{\|\bar{\vec{v}}_T(t)\|_i}, i = 1,2,3,4$
		Temporal energy/2nd order moments
11	speed	$\sigma_v^2(t)_1 = \frac{1}{T-1}\sum_{i=t-T+1}^{t} v^2(i)$
12	speed (centered)	$\sigma_v^2(t)_2 = \frac{1}{T-1}\sum_{i=t-T+1}^{t} (v(i) - \bar{v}_T(t))^2$
-	velocity	$\Sigma_{\vec{v}}(t)_1 = \frac{1}{T-1}\sum_{i=t-T+1}^{t} \vec{v}(i)\vec{v}(i)^T$
-	velocity (centered)	$\Sigma_{\vec{v}}(t)_2 = \frac{1}{T-1}\sum_{i=t-T+1}^{t} (\vec{v}(i) - \bar{\vec{v}}_T(t))(\vec{v}(i) - \bar{\vec{v}}_T(t))^T$
13,14	trace	$tr_{\Sigma_{\vec{v}}}(t)_i = trace(\Sigma_{\vec{v}}(t)_i), i = 1,2.$
15,16	eigenvalues ratio	$R_{\Sigma_{\vec{v}}}(t)_i = \frac{\lambda_{min}}{\lambda_{max}}(\Sigma_{\vec{v}}(t)_i), i = 1,2.$

Table 2: Features derived from the target global position provided by a tracker.

2.b$_2$ low-level activities & features

Blob based features

Reference:

optical flow of target detected pixels

Extracted Features coding:

target motion

relative motion inside the blob

directions of movement

directions of energy...

A total of 13 features => grand total of 29 features

2.b$_2$ low-level activities & features

Blob based features:

Instantaneous

Averaged

Energies

.spatial

.temporal

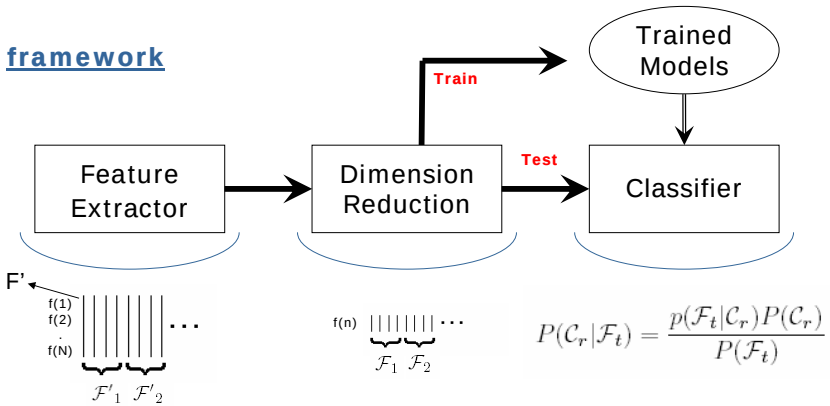
#	NAME	DEFINITION
		Reference set of pixels
-	target pixels	$P(t) = \{(x,y) \in \mathbb{R}^2 \mid (x,y) \in target(t)\}$
		Instantaneous
-	optical flow	$\vec{F}(t,x,y) = (f_x(t,x,y), f_y(t,x,y)), (x,y) \in P(t)$
-	mean flow	$\vec{F}(t)_1 = \frac{1}{N}\sum_{(x,y)\in P(t)} \vec{F}(t,x,y)$
-	mean flow (centered)	$\vec{F}(t)_2 = \vec{F}(t) - \vec{v}(t)$
-	flow norm	$f(t,x,y)_1 = \|\vec{F}(t,x,y)\| = \sqrt{f_x^2(t,x,y) + f_y^2(t,x,y)}, (x,y) \in P(t)$
-	flow norm (centered)	$f(t,x,y)_2 = \|\vec{F}(t,x,y) - \bar{\vec{v}}_T\|$
		Spatial energy/2nd order moments
17,18	motion energy	$\bar{f}(t)_i = \frac{1}{N(t)}\sum_{(x,y)\in P(t)} f^2(t,x,y)_i, i = 1,2.$
-	flow covariance	$\Sigma_{\vec{F}}(t)_1 = \frac{1}{N(t)}\sum_{(x,y)\in P(t)} \vec{F}(t,x,y)\vec{F}^T(t,x,y)$
-	flow covariance (centered)	$\Sigma_{\vec{F}}(t)_2 = \frac{1}{N(t)}\sum_{(x,y)\in P(t)} (\vec{F}(t,x,y) - \bar{\vec{F}}(t))(\vec{F}(t,x,y) - \bar{\vec{F}}(t))^T$
19,20	eigenvalues ratio	$R_{\Sigma_{\vec{F}}}(t) = \frac{\lambda_{min}}{\lambda_{max}}(\Sigma_{\vec{F}}(t)_i), i = 1,2.$
		Time averaged over T frames
-	mean flow	$\bar{\vec{F}}(t)_{i+2} = \bar{\vec{F}}(t)_i - \frac{1}{T}\sum_{j=t-T+1}^{t} \vec{F}(j)_i, i = 1,2.$
21	motion energy	$\bar{f}_T(t) = \frac{1}{T}\sum_{i=t-T+1}^{t} f(i)$
		Temporal energy/2nd order moments
-	mean flow	$\Sigma_{\bar{\vec{F}}}(t)_i = \frac{1}{T}\sum_{j=t-T+1}^{t} \vec{F}(j)_i\vec{F}(j)_i^T, i = 1,2,3,4.$
22..25	trace	$tr_{\Sigma_{\bar{\vec{F}}}}(t)_i = trace(\Sigma_{\bar{\vec{F}}}(t)_i), i = 1,2,3,4.$
25..29	eigenvalues ratio	$R_{\Sigma_{\bar{\vec{F}}}}(t)_i = \frac{\lambda_{min}}{\lambda_{max}}(\Sigma_{\bar{\vec{F}}}(t)_i), i = 1,2,3,4.$

Table 3: Features derived from the detected target optical flow.

3. feature selection & classification

3. feature selection & classification

framework



3. feature selection & classification

Classification:

Bayesian classifier
classes modeled with *Gaussian Mixtures*
Likelihood function learned with *Expectation-Maximization*

Feature selection:

by the *classification error (ground truth exists)*
start by evaluating all *features individually*

3. feature selection & classification

Exhaustive search

<brute search> all combination of 2-features, 3-features, etc.
(for 3 features → 3654 trials!)

Suboptimal search:

<lite search> search one dimension
2 features: start with the best individual feature and search, amongst all the others, for the best pair (57 trials)
3 features: from the best pair search, amongst the remaining, for the best triplet. (84 trials)

<lite-lite search> combine only the best individual features
2 features: 2 best individual features (30 trials)
3 features: 3 best individual features (30 trials)

RELIEF algorithm

3. feature selection & classification

Results with T=50 frames (2 sec.):

	Brute Search		Lite Search		Lite-lite Search		RELIEF	
# Features	Feat.	R. rate	Feat.	R. rate	Feat.	R. rate	Feat.	R. rate
1	5	86%	5	86%	5	86%	19	29%
2	5 23	92%	5 23	92%	5 8	74%	19 20	50%
3	5 19 23	95%	5 23 19	95%	5 8 17	79%	19 20 5	68%

Computational cost:

Classification error (choosing 3 features)
<brute search> 3654 trials ~ 60 hours
<lite search> 84 trials ~ 1,5 hours
<lite-lite search> 30 trials ~ 20 min.
Data distribution
<relief> ~ 30 sec.

3. feature selection & classification

Confusion matrix (for the best result – 3 features)

Class	1	2	3	4	5
1	95,2	4,8	0	0	0
2	5,8	84,7	0	0	9,5
3	0	0,2	98,1	0,8	0,9
4	0	0	0	100	0
5	0	2,7	5,3	0	92,0

F5 - Speed w/ constant velocity model
F19 – ratio eigenvalues flow spatial covariance matrix
F23 – trace flow temporal covariance matrix

3. feature selection & classification

Activity recognition example

4. classifier structure

4. classifier structure

Change the structure of the classifier

group "similar" activities

different features **for different groups**

model the likelihood **easily**

different classes <-> different features

Hierarchical Classification

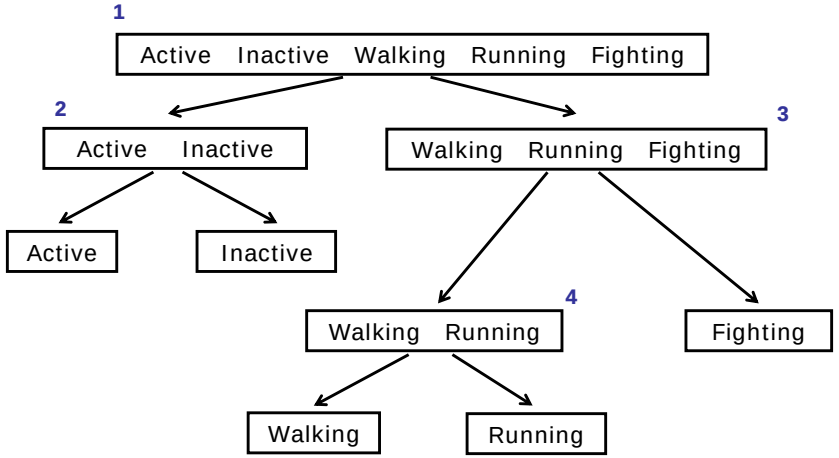
break **the problem into 4 simpler steps**

smaller **number of classes**

each classifier produces a *binary decision* **(2 classes)**

4. classifier structure

Hierarchical classifier:



4. classifier structure

Results:

# Feat.	Class.	Brute Search Feat.	Brute Search Reco. rate		Lite Search Feat.	Lite Search Reco. rate		Lite-lite Search Feat.	Lite-lite Search Reco. rate		RELIEF Feat.	RELIEF Reco. rate	
1	1	17	97%	95,1%	17	97%	95,1%	17	97%	95,1%	19	77,3%	95,1%
	2	3	96,8%		3	96,8%		3	96,8%		20	74,8%	
	3	24	99,2%		24	99,2%		24	99,2%		2	31,4%	
	4	18	97%		18	97%		18	97%		15	49,2%	
2	1	1 4	98,2%	97,2%	17 22	96,7%	94,9%	17 4	94,5%	91,5%	19 20	91,7%	71,8%
	2	8 10	98,1%		3 24	97,2%		3 8	91,3%		19 20	85,8%	
	3	20 24	99,3%		24 20	99,3%		24 25	99%		2 3	89,8%	
	4	22 23	99,1%		18 7	97%		18 5	94,6%		15 5	68%	
3	1	1 4 28	99%	98,8%	17 22 24	96,6%	94,9%	17 4 11	94,1%	92,5%	19 20 2	92,2%	60,7%
	2	8 23 28	99,1%		3 24 20	97,8%		3 8 15	88,1%		20 19 28	95,3%	
	3	4 7 18	99,8%		24 20 25	99,1%		24 25 7	98,7%		2 3 18	74,5%	
	4	4 18 24	100%		18 7 5	97,9%		18 5 22	98,9%		15 5 16	71%	

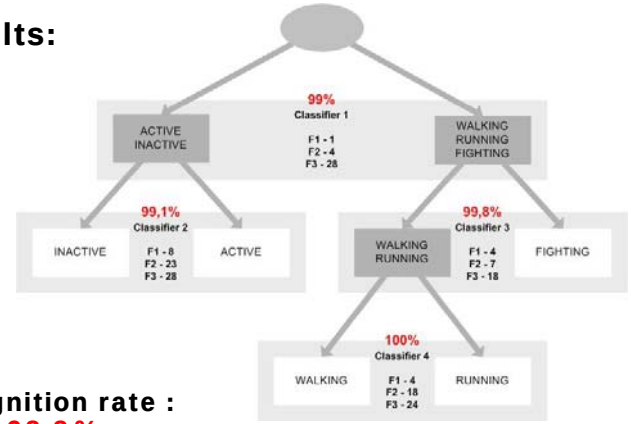
Number of Gaussians in the mixture:

<brute search> **1 feat ~ 11** **2 feat ~ 17** **3 feat ~ 32**

<relief> **1 feat ~ 14** **2 feat ~ 24** **3 feat ~ 42**

4. classifier structure

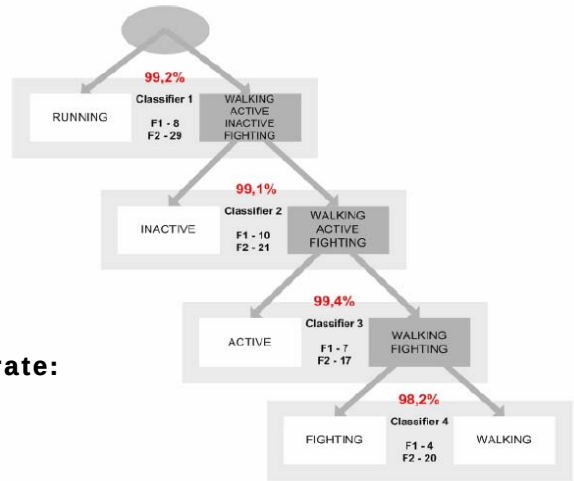
Results:



Recognition rate :
98,8 %

4. classifier structure (automatic)

Results:

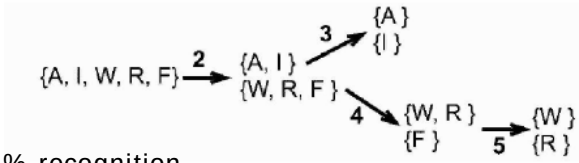


Recognition rate:
97,4 %

Scientific questions

Q1: What are good features for video-based human behaviour understanding? (WP1+WP4)

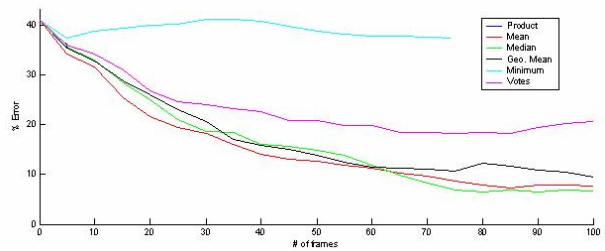
Bayesian hierarchical classifier, GMM, 29 features, {Active, Inactive, Walk, Run, Fight)



98.8% recognition

Scientific questions

Q4: The effect of the temporal window length for describing movement (WP1+WP4)



Error rate w/ different temporal fusion methods
50 frames used.

Scientific questions

Q2: Is there a benefit to using Gabor versus Gaussian derivative filters for feature extraction? (WP1)

Looking for eyes in faces:

Detection rate:

Gabor filter	Gaussian der	Directional der
47.55%	37.8%	37.8%

5. Conclusions & future work

Key aspects discussed
- Importance of feature selection
- data modeling (Gaussian mixtures + EM)
- classifier design (hierarchy)

Future
- Learning from partially unlabeled data (co-training)
- Contextual information
- Behavior interpretation (longer time horizon)

Integration in the CAVIAR architecture

4. classifier structure

Learn automatically **the classification structure**

feature selection <-> classification structure

Optimal solution (3 features, 5 classes): 438 480 trials

First level:

chose the class that is *best separable from all* **others**

evaluating the recognition error <brute search>

Next levels:

Continue the process to find which class is now easier to separate ...

5. conclusions & future work

recap question answers...

Questions

- REVIEW WP GOALS/DESCRIPTION, ETC....
- RECAP last year

- Q1 (WP4)
- Q2 (WP1)
- Q4 (WP4)

- Q3 (WP3)
- Q5 (WP3)
- Q6 (WP4)
- Q7 (WP4)
- Q8 (WP2)
- Q9 (WP5)
- Q10(WP6)