Introduction to computer vision

Tracking people by learning their apparences

Yoann Bourse

2010-2011 : Semestre 1

Presentation plan

1. Introduction
2. General model
   - Hidden Markov Model
   - Global scheme
   - Human body model
3. Building targets' models
   - Bottom-up approach
   - Top-down approach
4. Tracking learnt models
5. Performances

Tracking people

A difficult task

- Fast and unpredictable movements
- Variety of poses and clothes
- Complex environment

Existing methods

- Multiple cameras
- Manual initialization
- Simplified/controlled background

Tracking people

A difficult task

- Fast and unpredictable movements
- Variety of poses and clothes
- Complex environment

Existing methods

- Multiple cameras
- Manual initialization
- Simplified/controlled background

Model learning method

Based on low-level and precise information.

2007 :
Deva Ramanan
David A. Forsyth
Andrew Zisserman

Presentation plan

Hidden Markov Model

X_t hidden states : individual(s) positions
I_t observable parameters : images

$$\mathbb{P}(X_T, \ldots, X_1, I_T, \ldots, I_1) = \prod \mathbb{P}(X_t \mid X_{t-1})\mathbb{P}(I_t \mid X_t)$$

- **Inference** :
  predicting the next step and refining it with data
- **Data association** : predicts regions of interest
  background suppression, skin detection...

Hidden Markov Model

X_t hidden states : individual(s) positions
I_t observable parameters : images

$$\mathbb{P}(X_T, \ldots, X_1, I_T, \ldots, I_1) = \prod \mathbb{P}(X_t \mid X_{t-1})\mathbb{P}(I_t \mid X_t)$$

- **Inference** :
  predicting the next step and refining it with data
- **Data association** : predicts regions of interest
  background suppression, skin detection...

- $\mathbb{P}(X_t \mid X_{t-1})$: **Dynamic model** :
  fixed type of movement ?
- $\mathbb{P}(I_t \mid X_t)$: **Likelihood model** :
  generic but specific, updated online ?

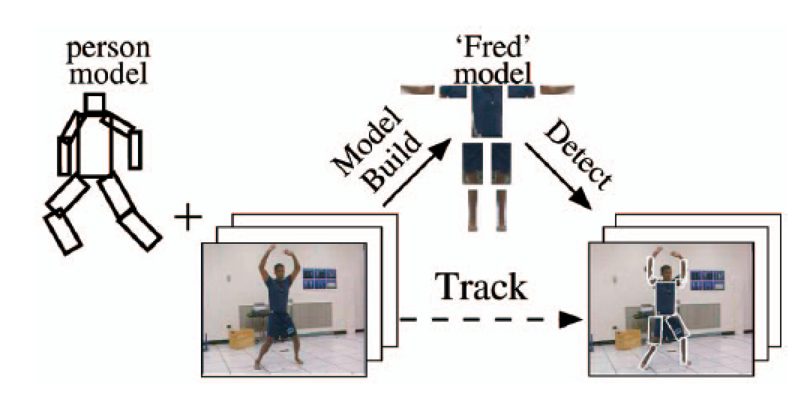
  $\Longrightarrow$ **Template** (edges) :
  Pre-determined = too generic to be useful
  $\Rightarrow$ built from source video

- $\mathbb{P}(X_t \mid X_{t-1})$: **Dynamic model** :
  fixed type of movement ?

- $\mathbb{P}(I_t \mid X_t)$: **Likelihood model** :
  generic but specific, updated online ?

  $\Longrightarrow$ **Template** (edges) :
  Pre-determined = too generic to be useful
  $\Rightarrow$ built from source video

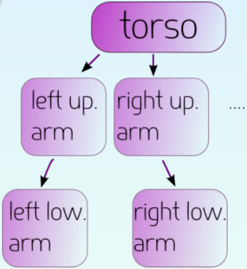
Global scheme



Human body model

Pictorial structure

body = puppet of rectangles, ordered as a tree



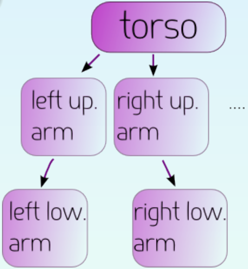
The model will rely on :

- $\mathbb{P}(P_t^{arm} \mid P_t^{torso})$:
  consistency of the body (threshold)
- $\mathbb{P}(P_t^{arm} \mid P_{t-1}^{arm})$:
  consistency of the movement (threshold)
- $\mathbb{P}(I_t \mid P_t^{arm})$:
  consistency with observation (gaussian)

Human body model

Pictorial structure

body = puppet of rectangles, ordered as a tree



The model will rely on :

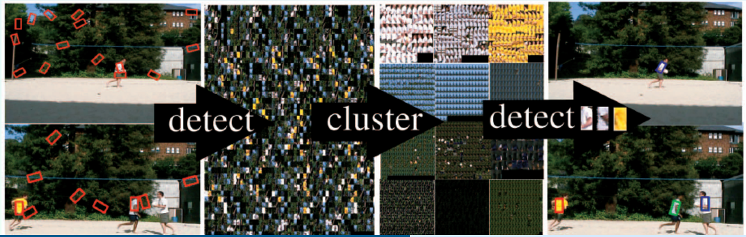
- $\mathbb{P}(P_t^{arm} \mid P_t^{torso})$:
  consistency of the body (threshold)
- $\mathbb{P}(P_t^{arm} \mid P_{t-1}^{arm})$:
  consistency of the movement (threshold)
- $\mathbb{P}(I_t \mid P_t^{arm})$:
  consistency with observation (gaussian)

Building models

Bottom-up approach

Principle

- **Detection** with a rough detector (edges), on several frames
- **Clustering** to regroup the detected objects corresponding to the same thing
- **Eliminating** some clusters



Rough detection

- Look for rectangle projections of cylinders
- Scale-sensitive
- Convolution
- Separate in two parts and consider the minimum score to avoid false positive
- Hardest part of bottom-up method

Clustering

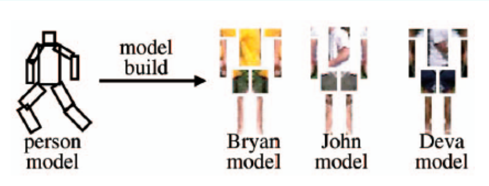
- Unknown number of clusters : **mean-shift algorithm**
- 512 dimensions RGB histogram
- Radius sensitive
- Overmerging (two arms)

Treatment

- Sticking to position
- Sticking to motion
- Eliminate parasites and stillness

Online evolution

- **Inference** to preview torso's next position and look for limbs around it.
- Recover and learn body structure
  $\Rightarrow$ refining into person-specific detectors
  (50-100 frames suffice)



Top-down approach

Difficulties :

- Variability (shape, pose, clothes...)
- Building model : need for precision, limb localization...

$\Rightarrow$ **Opportunistic detection** frame-wise on easy positions
Then upgrade model, and discriminate false positives.

Top-down approach

Difficulties :

- Variability (shape, pose, clothes...)
- Building model : need for precision, limb localization...

⇒ **Opportunistic detection** frame-wise on easy positions
Then upgrade model, and discriminate false positives.

Position

- Walking laterally : kinematic constraints
- Fixed scale
- Scissor-leg and occluded arm
- Looking for mirror position (second sense)

- Ignore vertical and horizontal rectangle (too many false positives)
- Consistency constraints : similarity of limbs
- Region-sensitive

Based on the average distance between template edge and closest image edge. (likelihood threshold)

Classify

Simply classify limbs in RGB space, ignoring low-information pixels :
Lowest number of misclassified pixels and threshold

$\Rightarrow$ Get limbs masks

Note : illumination insensitive and discriminative

Performances

A lot of false positives (threshold if too many people), but still good precision-recall.
Sometimes handles atypical poses.

Model building

Bottom-up :
efficient on short time and various poses
depends on background

Top-down :
more robust, time-requiring
depends on poses

Tracking learnt models

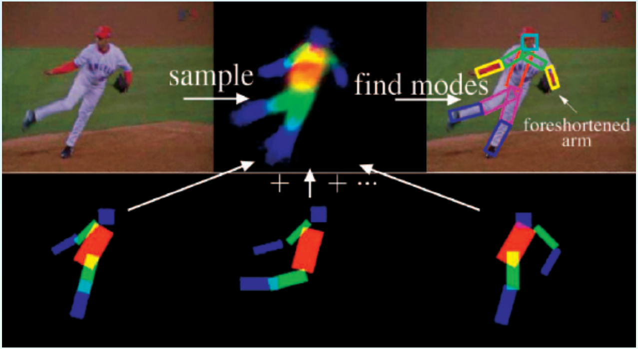
Generative or discriminative representation of instance-specific models.

$\Rightarrow$ **Generative** for multiple people

Compares candidate patches to learnt models.
Handles multiple scale by pyramid search.
Handles many different activities.
Frame by frame, but smoothed temporally for coherence.

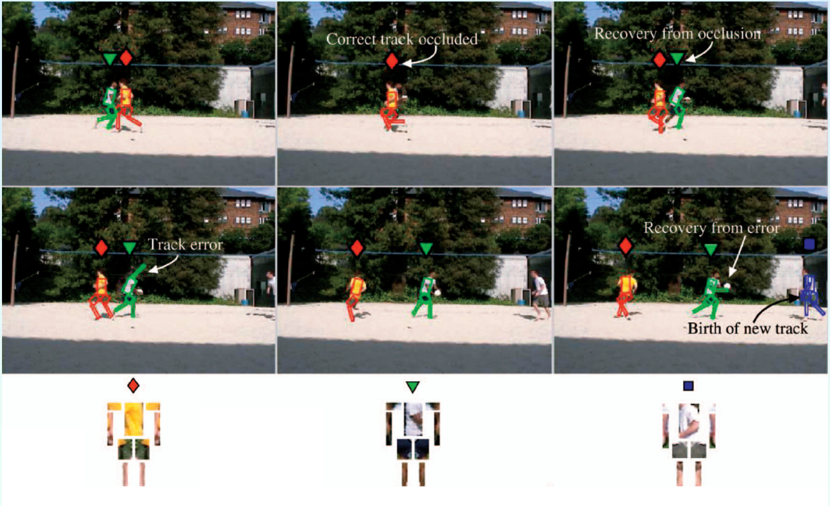
Dealing with occlusion

Search for one arm and one leg.
$\Rightarrow$ Meanshift when no overlapping.



Also result in spatial smoothing : stable and adaptative

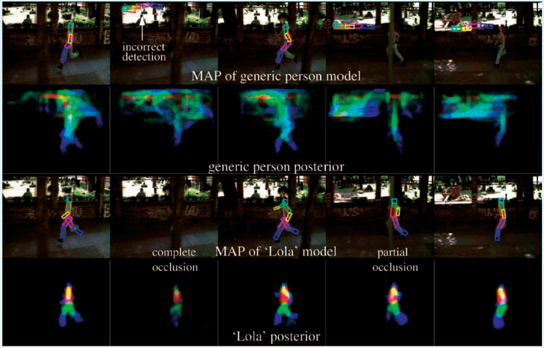
Example


Correct track occluded
Recovery from occlusion
Track error
Recovery from error
Birth of new track

How to measure ?

Measure detection rate and not time to fail.
Correct localisation when the majority of pixels are correctly labelled.
Manually specified ground truth
Biased by distinctive clothing in movies.
But still good for many people alike.

Instance-specific models



% frames correctly localized	Torso	Arm	Leg
Generic detector	31.4	13.0	22.1
Specific detector	98.1	94.3	100

Tracking initialization

Build models on a subset of frames :
build easily torso but not other limbs.

Efficient generalization :
No need for a lot of frames : parasite models.

Conclusion

- Auto-initialized
- Instance-specific
- Unbounded
- Dynamic