ICTP, Italy

16 March 2017

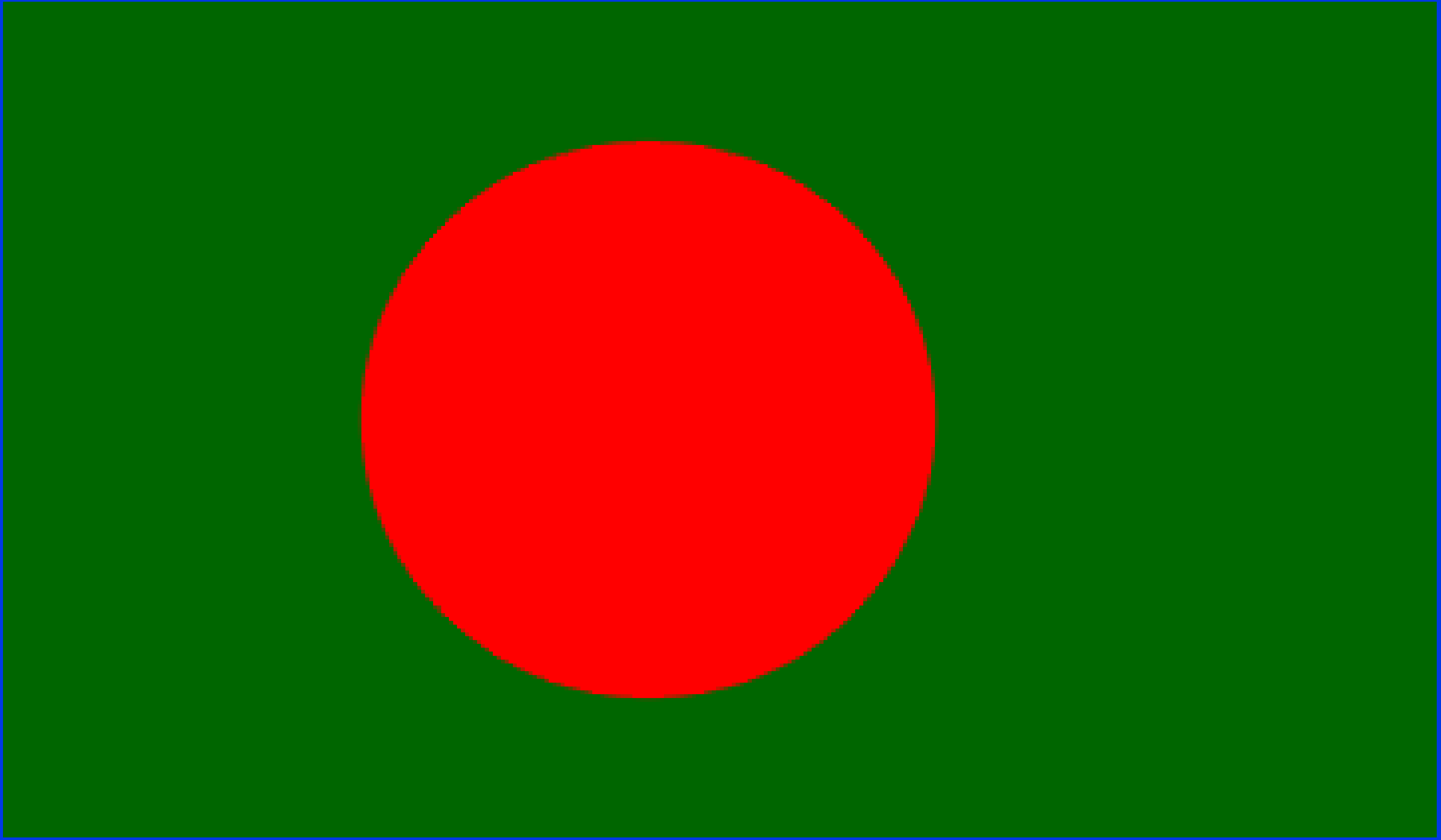
Bangladesh!
&
Action Recognition: Few Points

Md. Atiqur Rahman Ahad

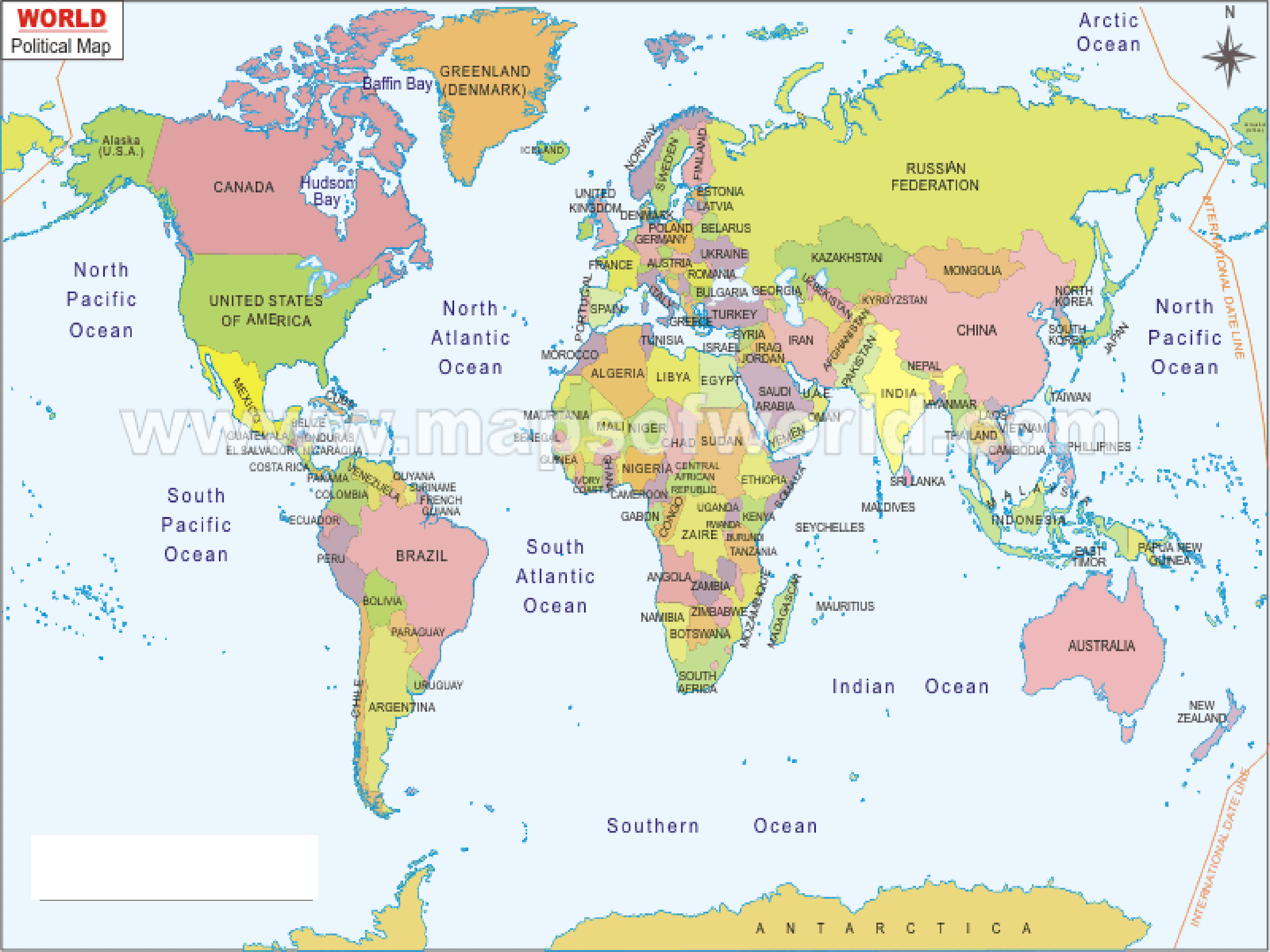
University of Dhaka, Bangladesh

Web: http://aa.binbd.com

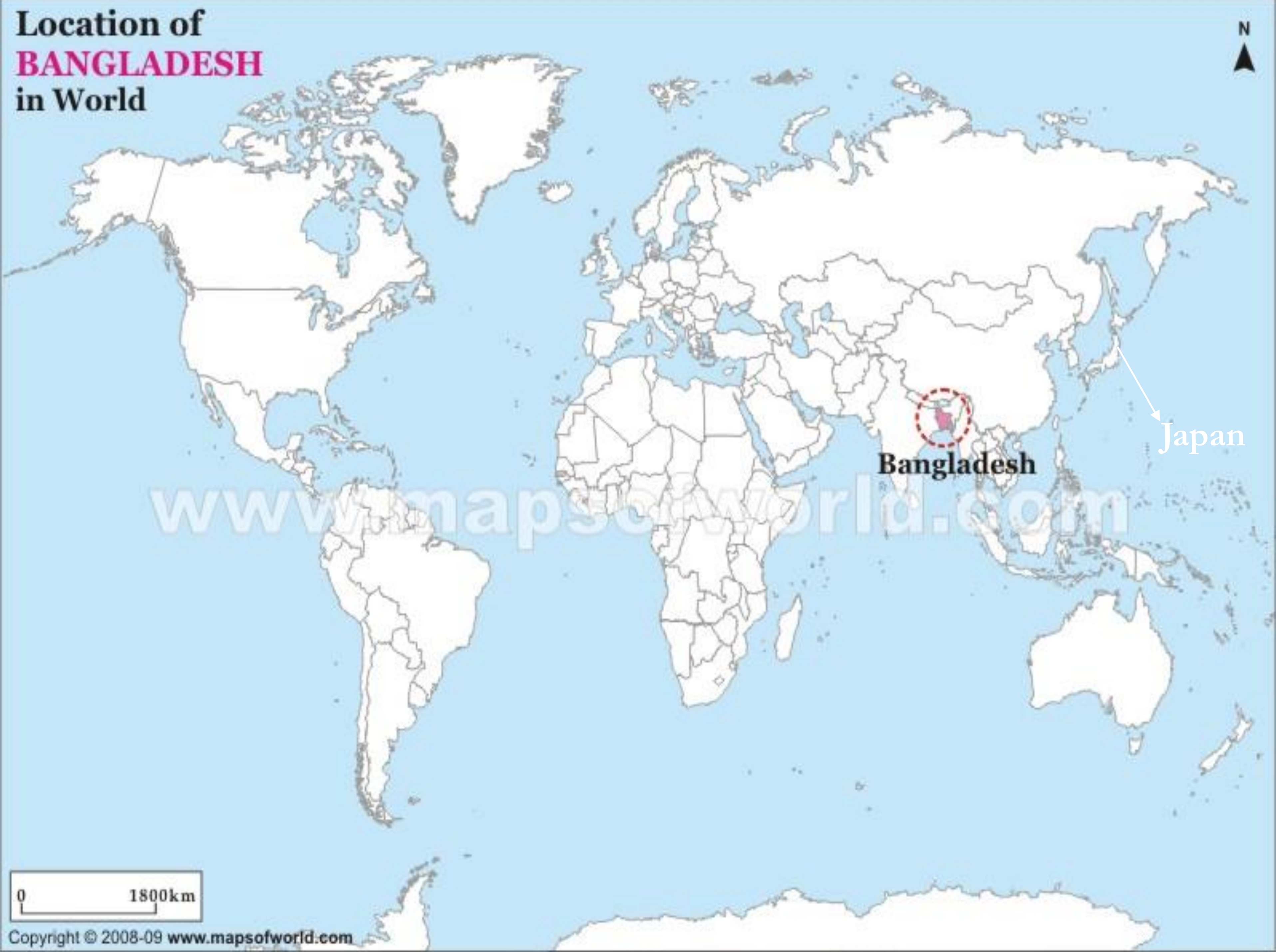
Email: atiqahad@univdhaka.edu

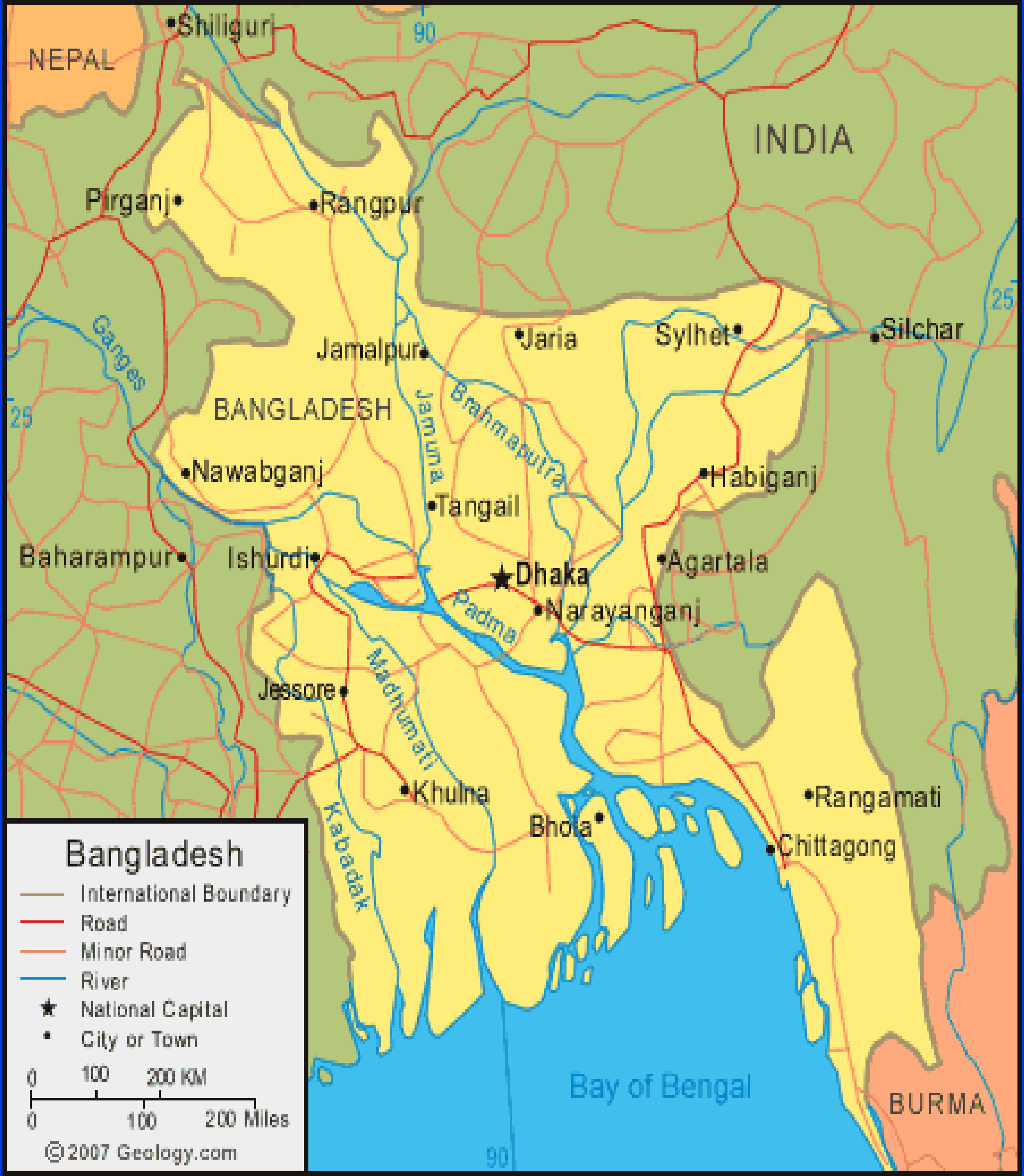


বাংলাদেশ
BANGLADESH



WORLD
Political Map


Location of
BANGLADESH
in World
N
Bangladesh
Japan
0
1800km


Shiliguri
90
NEPAL
INDIA
Pirganj
Rangpur
Jamalpur
Jaria
Sylhet
Silchar
25
Ganges
25
BANGLADESH
Jamuna
Brahmaputra
Nawabganj
Habiganj
Tangail
Baharampur
Ishurdi
Dhaka
Agartala
Padma
Narayanganj
Jessore
Madhumati
Khulna
Rangamati
Bhola
Chittagong
Kabadak
Bay of Bengal
BURMA
90
Bangladesh
International Boundary
Road
Minor Road
River
National Capital
City or Town
0
100
200 KM
0
100
200 Miles
© 2007 Geology.com

Area: 147, 570 km^2

Capital: **Dhaka**

Population: **170 million ☺**

Mostly flat plain, with hills in the northeast and southeast

University of Dhaka
http://www.du.ac.bd/

- From 1921 ~
- 13 Faculties
- 77+ departments
- 11 institutes
- 51+ research centers
- 38,000+ students
- ~2000 teachers

Faculty of Engineering & Technology

- Dept. of Electrical & Electronic Engineering

F.Zaman

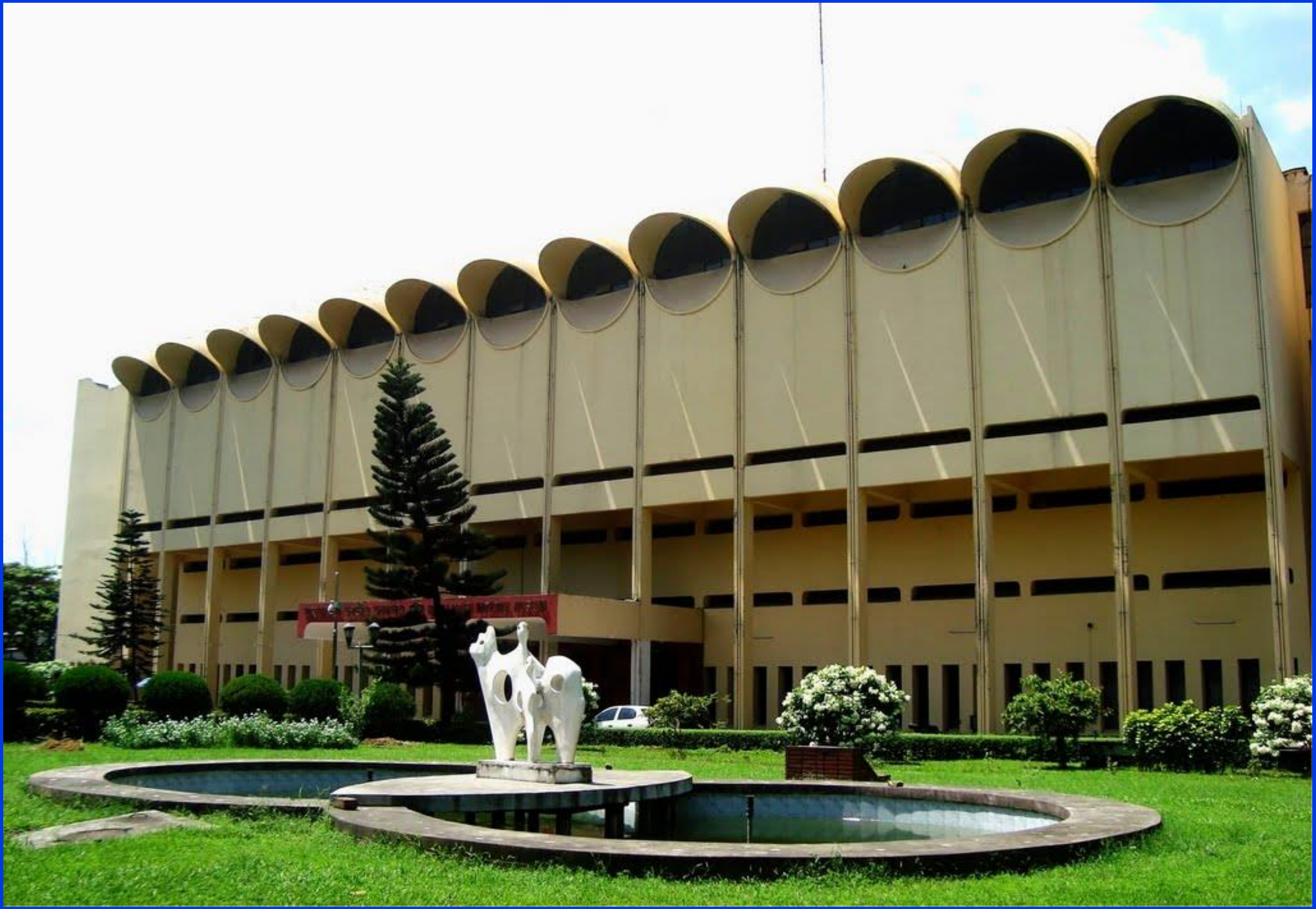
DU

My home!

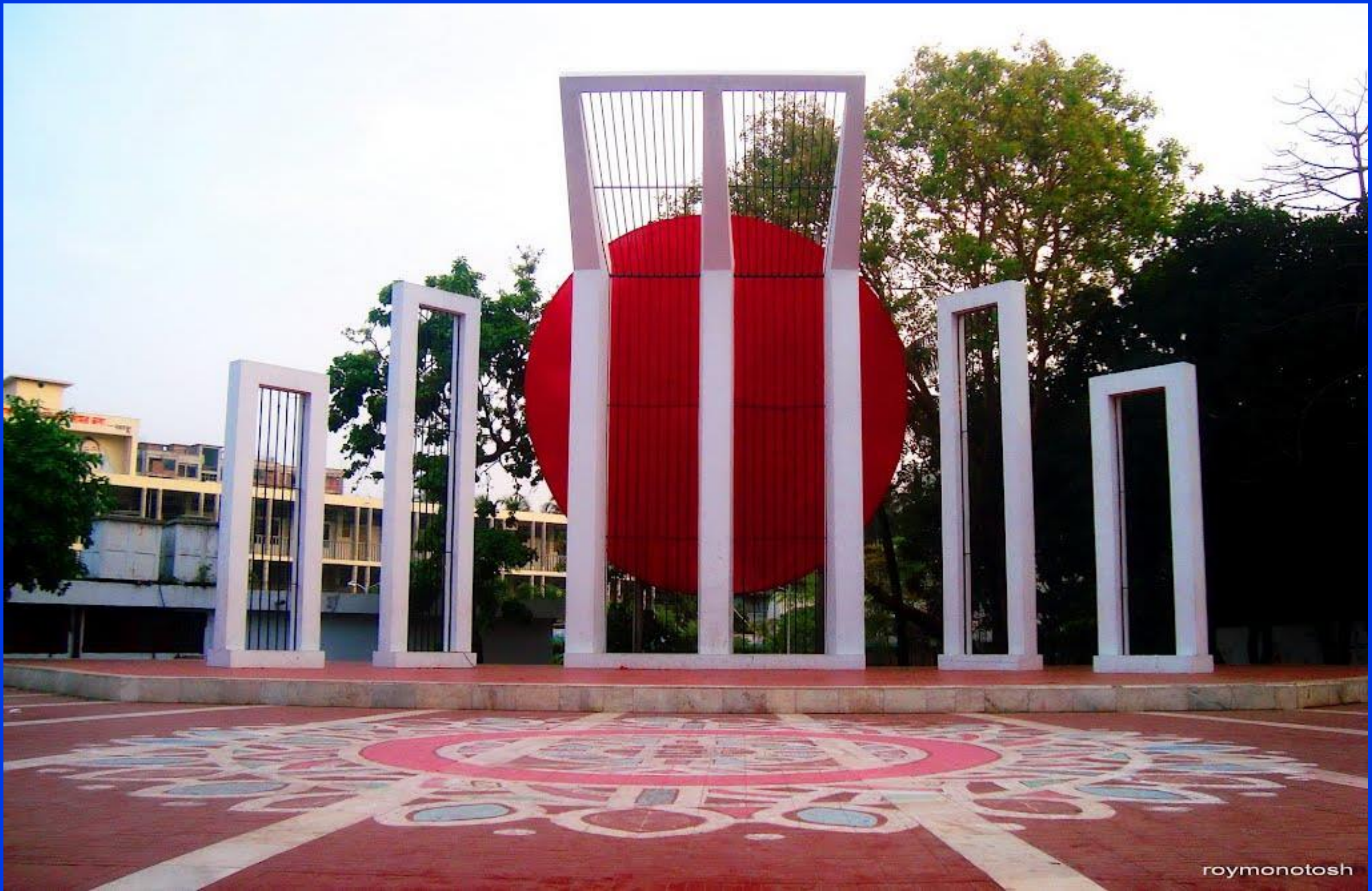
DU

2nd International Conference on Informatics, Electronics & Vision
17-18 May, 2013
Dhaka, Bangladesh
http://ICIEV.org
Sponsor
OSA
The Optical Society
Technical Co-Sponsor
IEEE computer society
PAMI TC
ICROS
Organized by
SER
Engineering Research

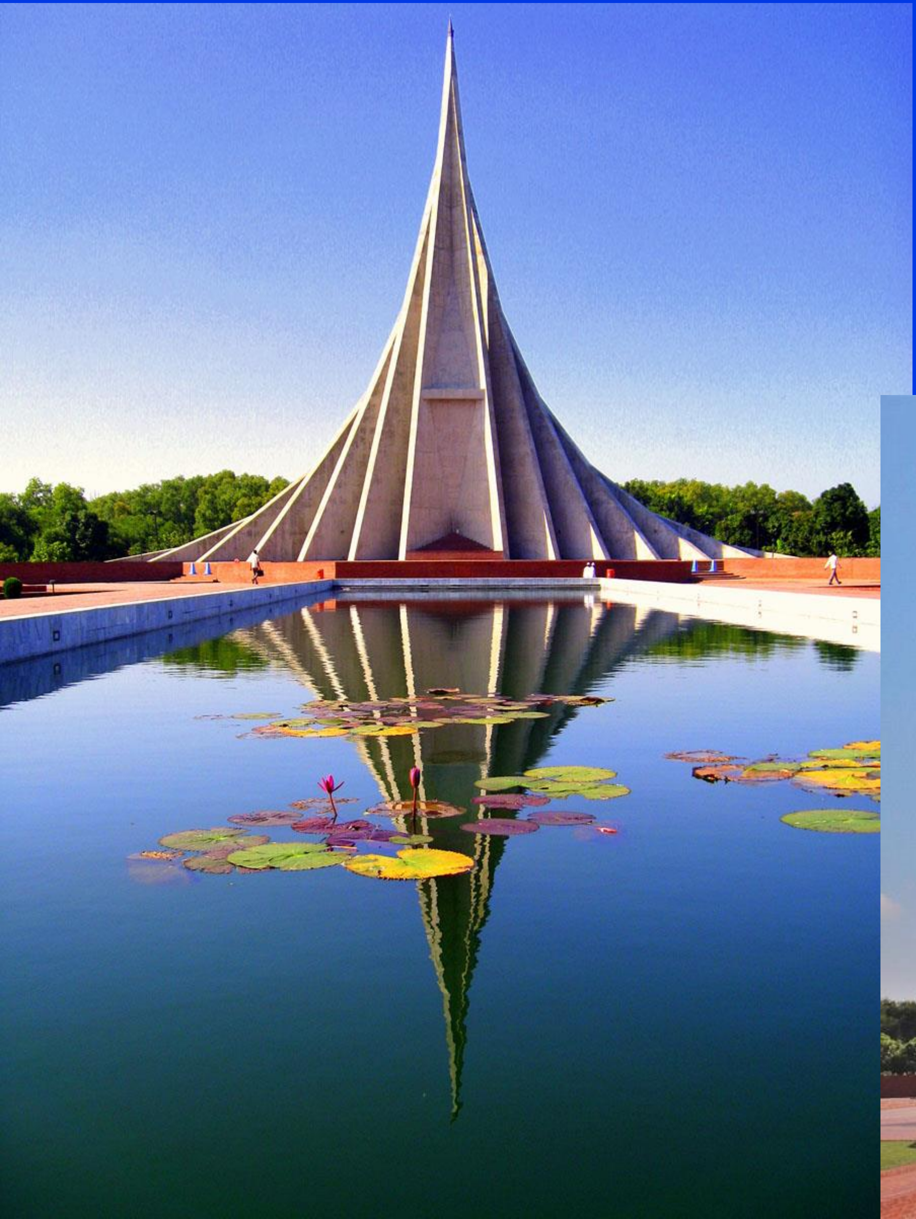
National Museum



Shaheed Minar – Int’l Mother Language day Monument



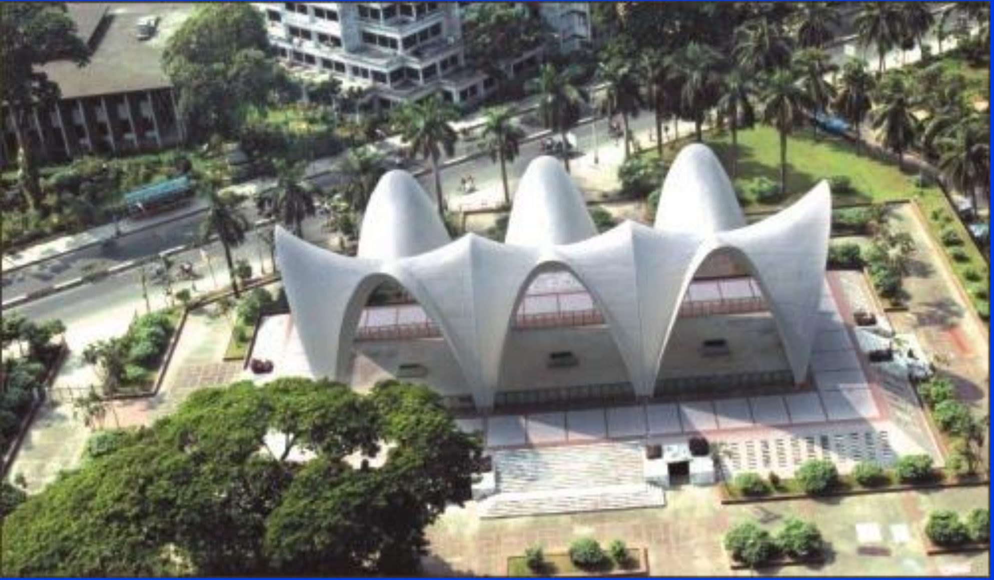
National Memorial



Lalbagh fort

Sonargaon

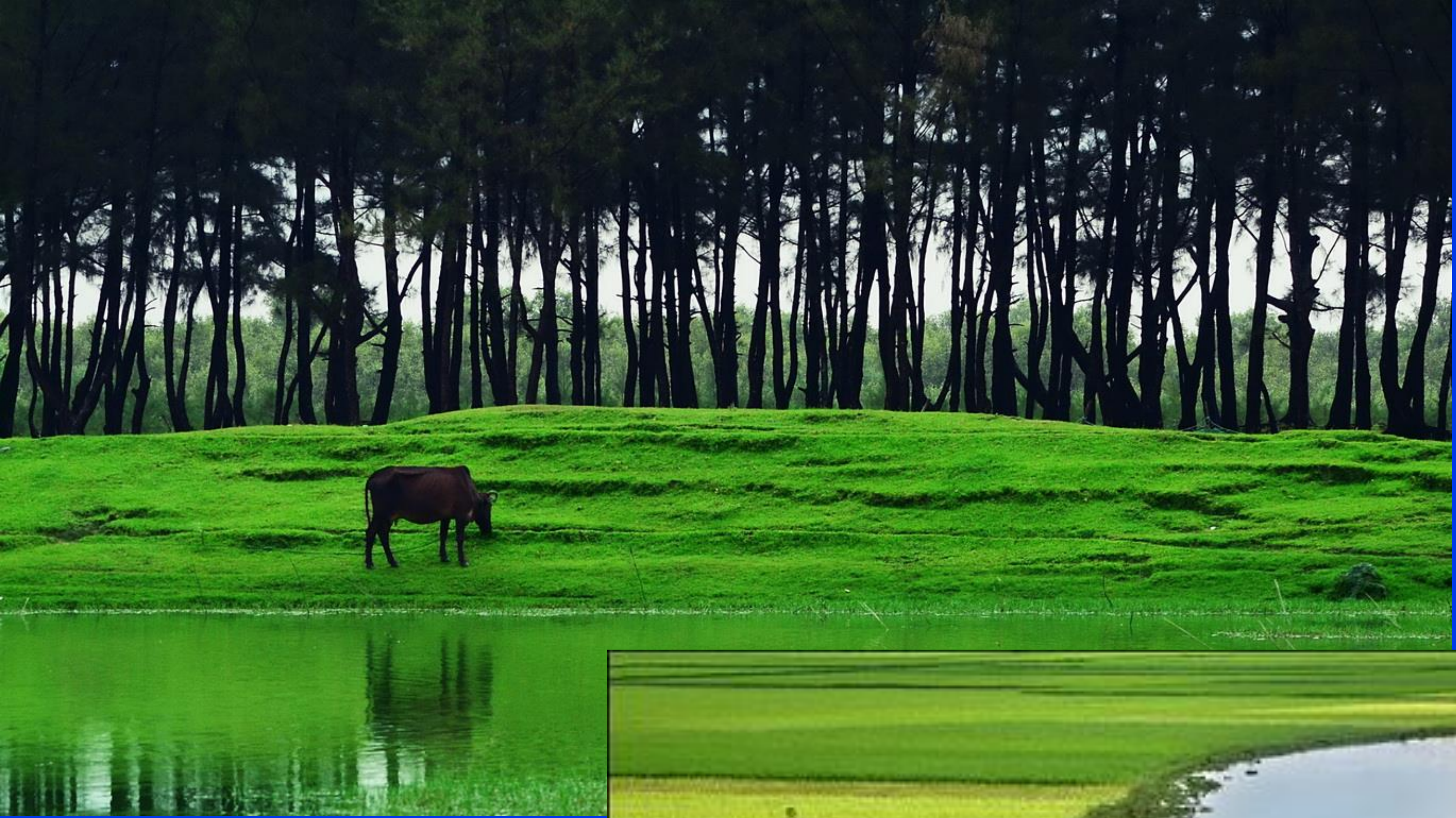
Parliament // Around DU



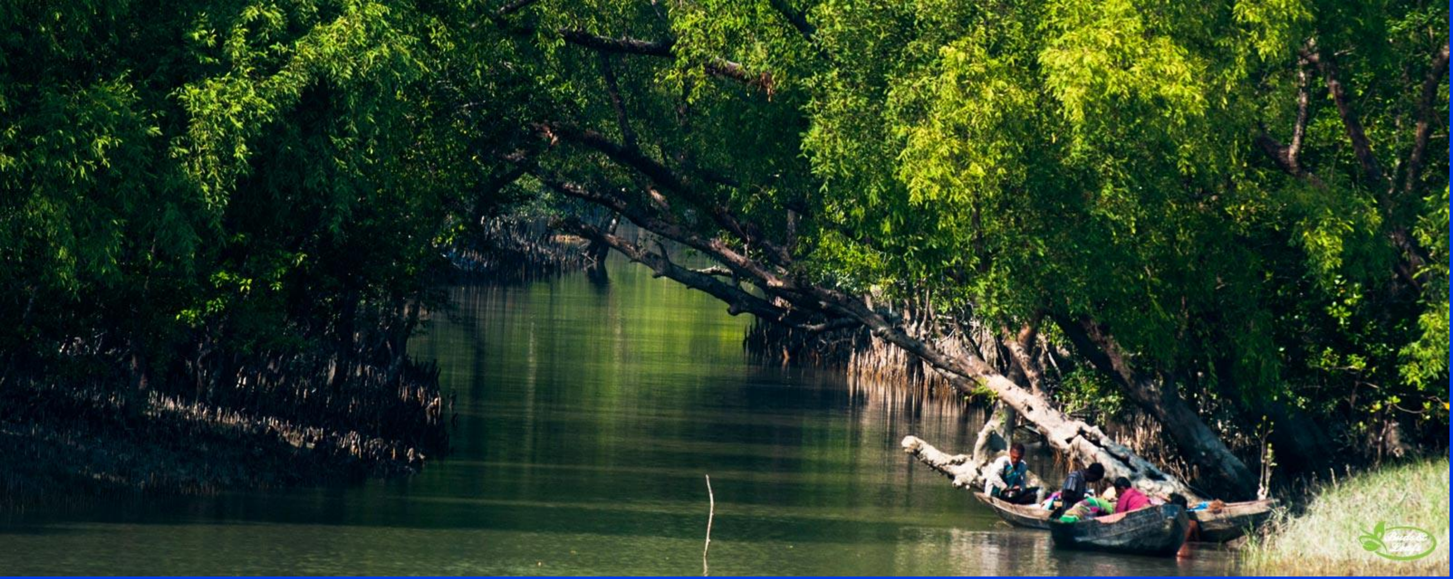
Ahsan Manjil – next to DU

Green BD

Green BD



Green BD



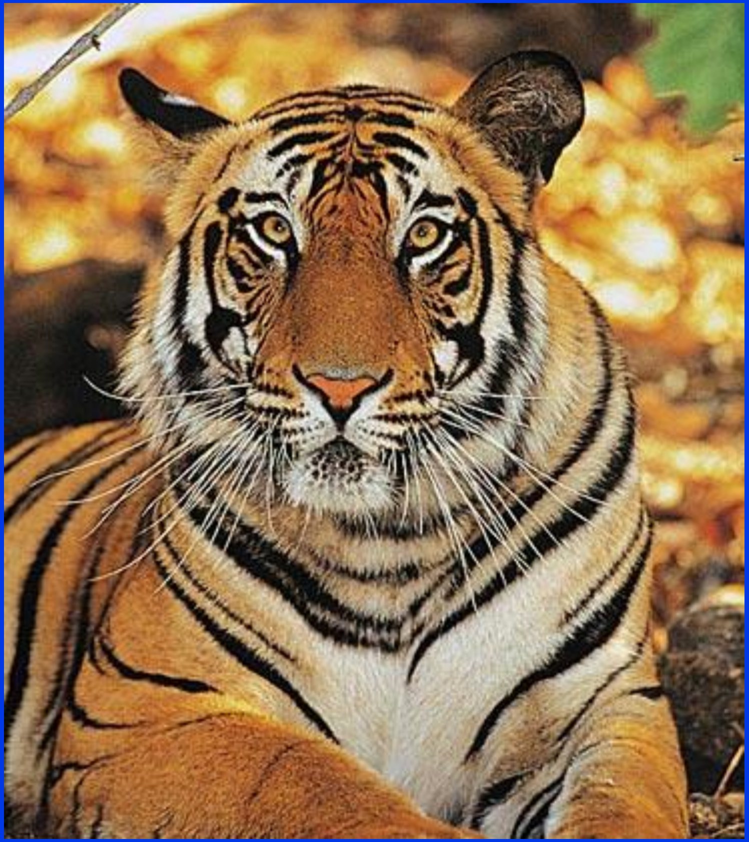
UNESCO World’s Heritage:

The Sundarbans – World’s largest Mangrove forest

© Subarashi Tours

© Nik

In Sundarbans



Royal Bengal Tiger - Our National Animal

UNESCO world's Heritage -

Ruins of the Buddhist Vihara at Paharpur

UNESCO World's Heritage:

Historic Mosque City of Bagerhat

Cox’s Bazar – World’s longest sandy beach

Saint Martin’s Island

Our National Bird

Doel Bird (Magpie Robin)

Our National Fruit

Jackfruit (*Kathal*)

Summer fruits!

©J.M.Garg 2007

Summer fruit – Palm tree!

Our National Flower

Water Lily (*Shaapla*)

Summer Flowers

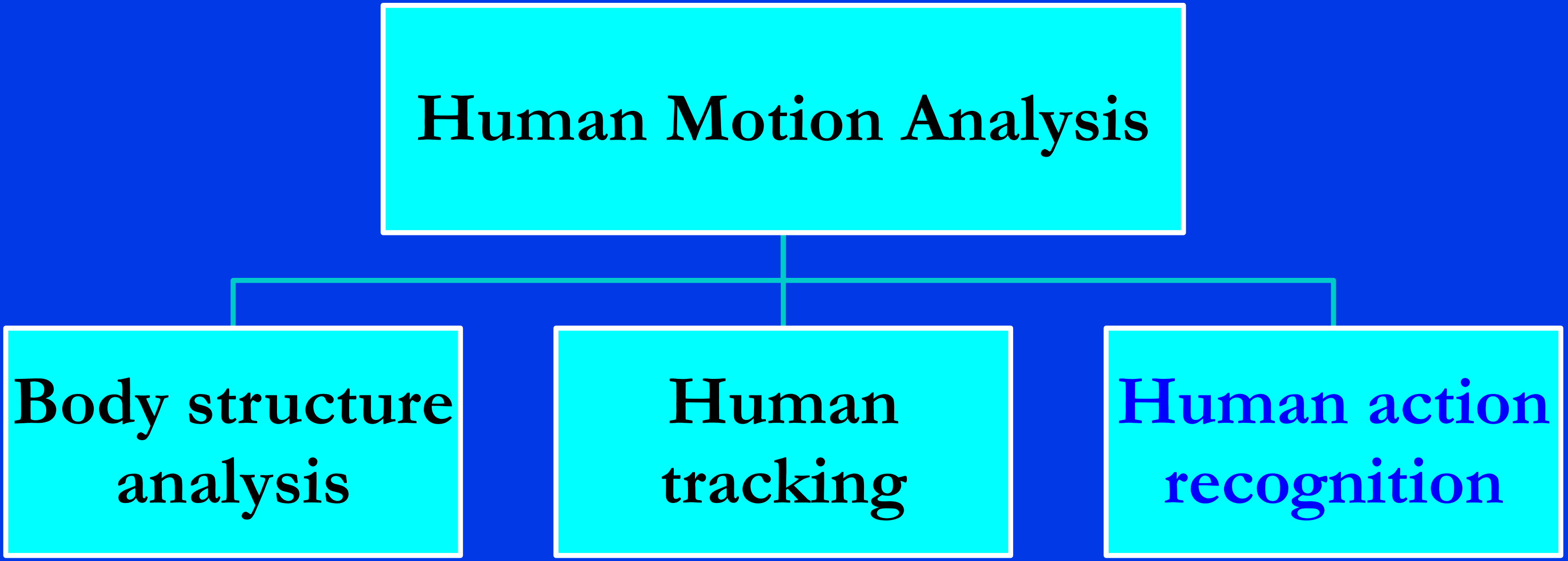
Thanks a lot!

Join 6th ICIEV, 1~3 Sept. 2017

University of Hyogo, Japan!

http://cennser.org/ICIEV

Few points on action recognition



Human Motion Analysis

Body structure analysis

Human tracking

Human action recognition

more

Application Arenas

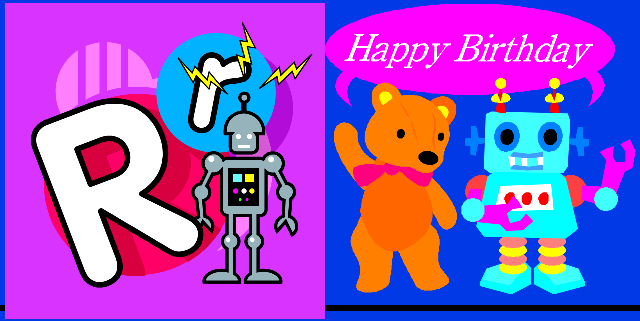
Surveillance

Parks, streets, venues, etc. → Security

Hospital, rehabilitation center, smart-house

Entertainment

Sports video analysis



Action understanding by robot

Monitoring crowded scenes

http://mha.cs.umn.edu/proj_recognition.html

Action Recognition in Surveillance Video

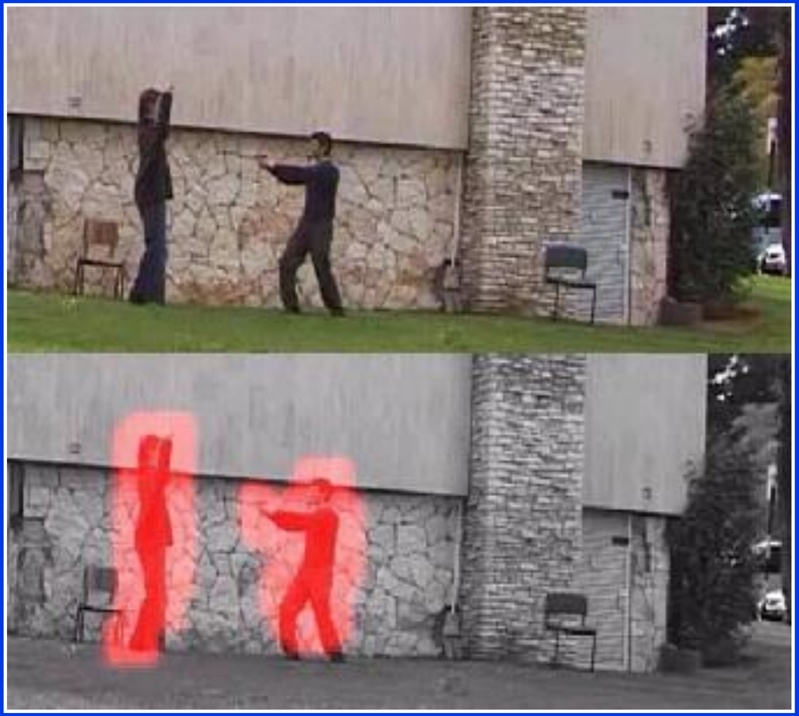
Detecting people fighting

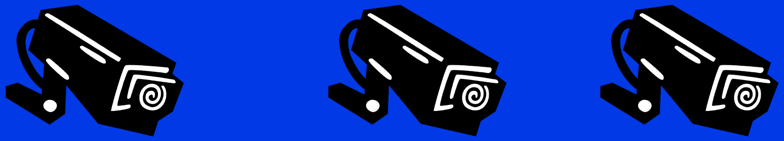
Falling person detection

Detecting Suspicious Behavior

Fence Climbing

Shooting

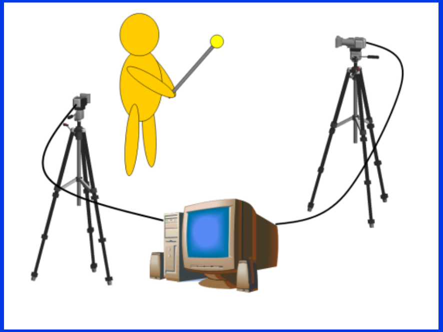




Many cameras → Lots of input sequences

→ Difficult for man-controlled surveillance

Hence, automated action recognition, behavior analysis, motion segmentation, etc. are crucial tasks to handle



SOME ASSUMPTIONS ON ACTION RECOGNITION

Some Assumptions…

a) Assumptions related to **movements**

- Subject (human/car) remains **inside** the workspace
- None or constant **camera motion**
- Only one person in the workspace at the time
- The subject faces the camera at all time
- Movements parallel to the camera-plane
- No occlusion
- Slow and continuous movements
- Only move one or a few limbs
- The motion pattern of the subject is known
- Subject moves on a flat ground plane

Some Assumptions …

b) Assumptions related to **appearance**

Environment –

1. Constant lighting - indoor
2. Static background
3. Uniform background
4. Known camera parameters
5. Special hardware (FPGA, etc.)

Subject -

1. Known part pose
2. Known subject – gender, size, height, race, etc.
3. Markers placed on the subject
4. Special cloths – color, no texture...
5. Tight-fitting cloths

Action Analysis …

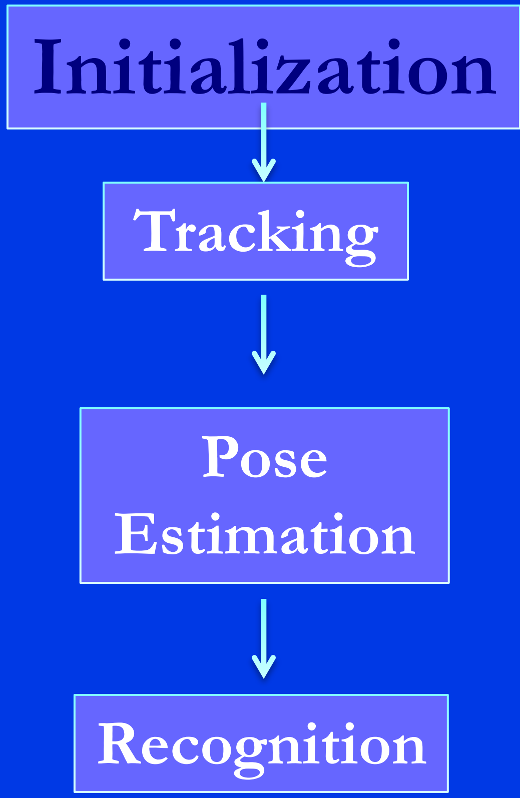
1. Initialization:

Ensuring that a system starts its operation with a correct interpretation of current scene.

→ processing of video/image –

- camera calibration,
- adaption with scene conditions,
- filtering, normalization,
- scene identification.

→ Model-based – in virtual reality



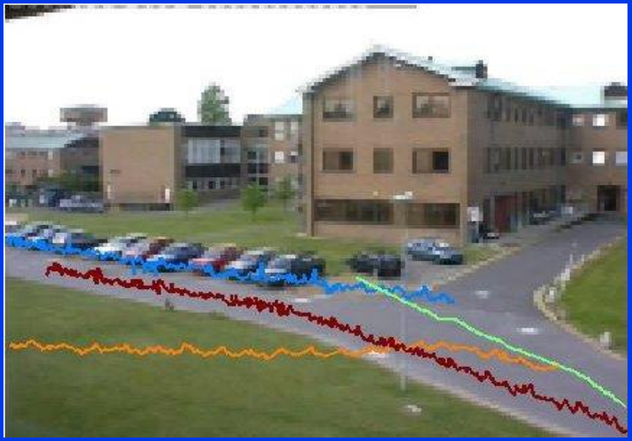
Model Initialization

- **Need prior info.** - e.g., kinematic structure (limb, skeleton); 3D shape; color appearance; pose; motion type.
- Initialization of appearance models for monocular tracking and pose estimation remains an open problem.
    - e.g., initialization of appearance based on image patch exemplars or color mixture models (e.g., color-based particle filter).
- Fully automatic initialization – future task!

2. Tracking – human/moving objects, between limbs

- **Tracking!**

- outdoor tracking,

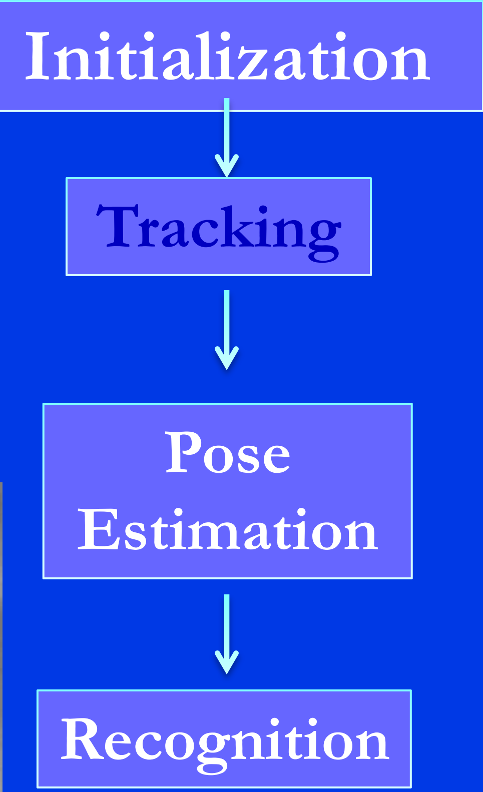


- tracking through occlusion, &
- detection of humans in still images.

e.g.,

Robotic line tracking,

Tracking vehicles, persons



Initialization

Tracking

Pose Estimation

Recognition

2. Tracking – Segmentation…

2.1 Initial step for many – **Background Subtraction**

→ **divided into** →

Background representation (color space – RGB, HSV; mixture of Gaussian),

Classification (shadow problem, false positive, etc. – classifiers based on color, gradients, flow info),

Background updating (outdoor – change of light, dynamic), **&**

Background initialization.

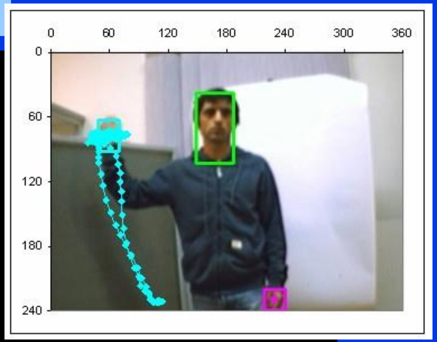
2.2 Motion-based segmentation

- motion gradient, optical flow, frame subtraction

Data Representations

Object-based	Image-based
point	Spatial - x,y
box	Spatio-temporal - x,y,t
silhouette	edge
blob	features

directly on the pixels

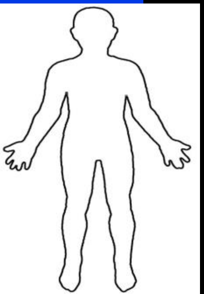


Point representations:

- Active/passive markers.
- Multi-camera system → 3D

Box:

- Set of boundary boxes – region-of-interest (ROI)
- track the box, process, …



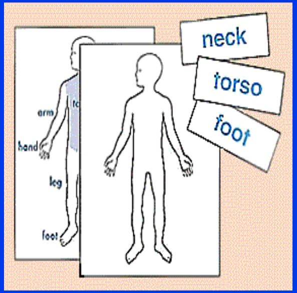
Silhouette:

- by threshold / subtracting
- find active contour or ROI

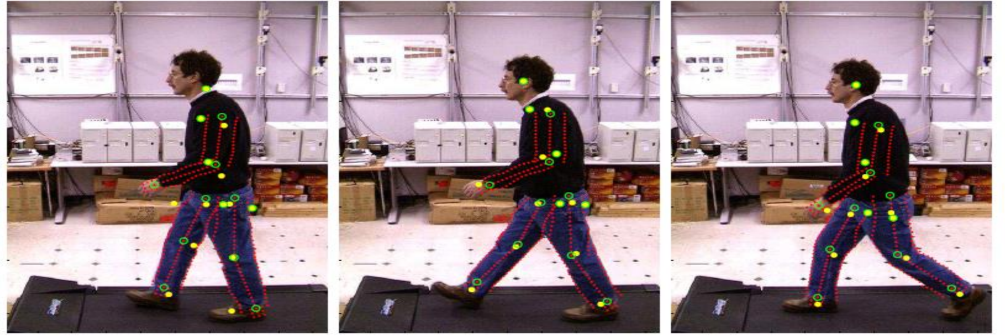
Blobs:

- grouping similar info/interest points
- based on correlation, flow, color-similarity, hybrid

3. Pose estimation – for surveillance



- Process of estimating the configuration of the underlying kinematic (or skeletal) articulation structure of a person → hand/head/body's center



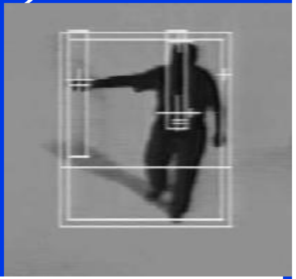
- It can be a post-processing step in a tracking algorithm
- It can be an active part of the tracking process

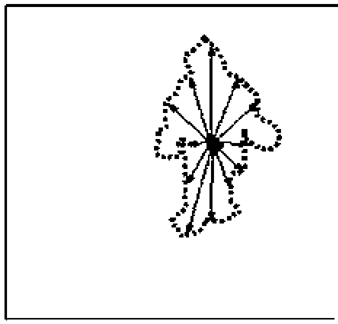
3. Pose estimation – human MODEL

Geometric model *or,* Human model

Category: based on human model's use –

a) Model-free (individual body parts are first detected and then assembled to estimate the 2D pose) – points, simple shape/box, stick-figures.



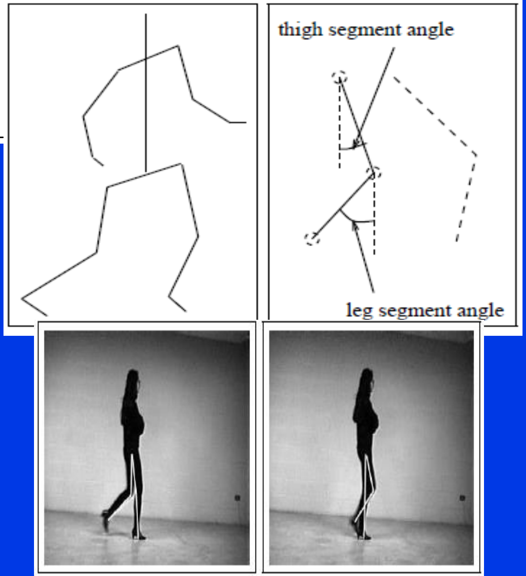


→ with **markers** – easy!

→ **no markers** –

- use hands & head (3 points!)

- mouth/center of body...



3. Pose estimation – human MODEL…

b) Indirect model use – use model as a reference/ look-up table (positions of body parts, aspect ratios of limbs, etc.)

c) Direct model use (Kalman filter, particle filter) – model is continuously updated by observations.

→ model type: cylinders, stick-figures, patches, cones, boxes, ellipse

→ model parts: body, leg, upper body, arm...

→ abstraction levels: edges, joints, motion, silhouette, sticks/anatomy, contours, texture, blobs...

→ dimensionality: 2D, 3D, 2.5D [estimating 3D pose data *based on* 2D processing // testing a 3D pose estimating framework *on* pseudo-3D data]

4. Recognition – what a person is doing!

Action Hierarchy

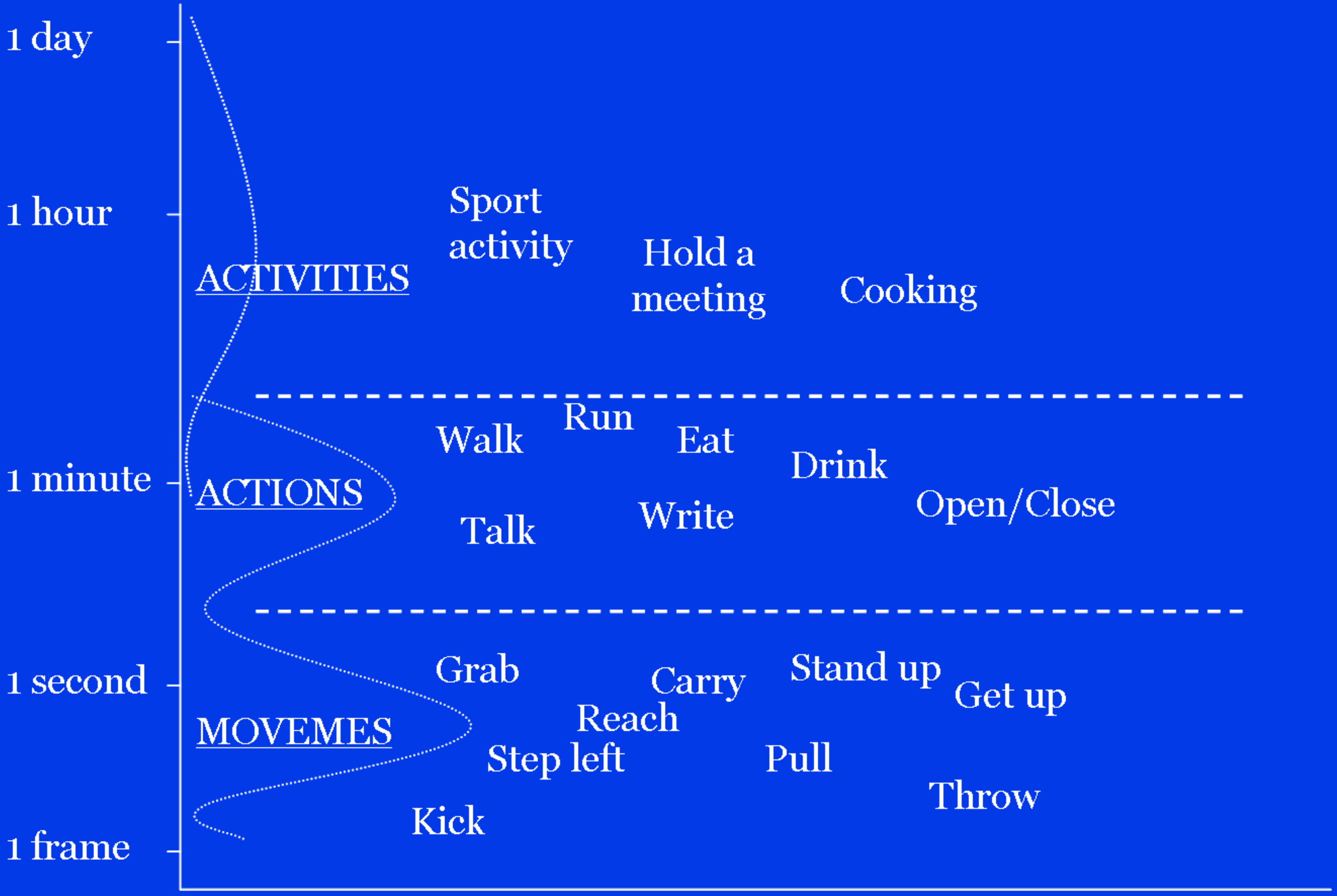
- *action primitives / basic action* (atomic entities out of which *actions* are built. Tennis: e.g., forehand, backhand, run left, & run right)

- *actions* (sequence of *action primitives* needed to return a ball)
- *activities* (playing tennis!)

actions, activities, simple actions, complex actions, behaviors, movements, etc.

→ interchangeably by different researchers.

Action Hierarchy…



1 day

1 hour

1 minute

1 second

1 frame

ACTIVITIES

Sport activity

Hold a meeting

Cooking

ACTIONS

Walk

Run

Eat

Drink

Open/Close

Talk

Write

MOVEMES

Grab

Carry

Stand up

Get up

Reach

Step left

Pull

Throw

Kick

What are Actions?

Actions Come in Many Flavors

No Motion

Prolonged

Motion

Multi-tasking!

Whole body

Local

4. Recognition *(cont.)*

- **Scene interpretation –**

  Entire image is interpreted without identifying particular objects or humans (*detecting unusual situation, surveillance*)

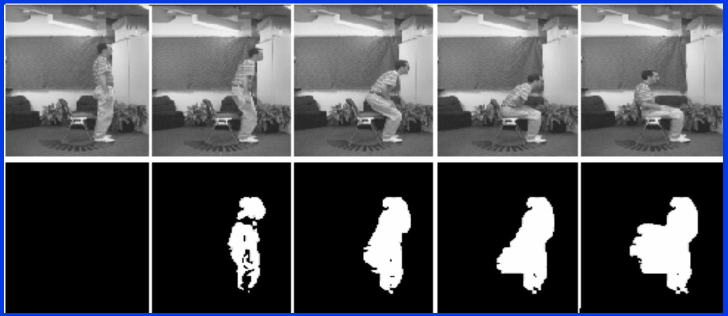
- **Holistic recognition –**

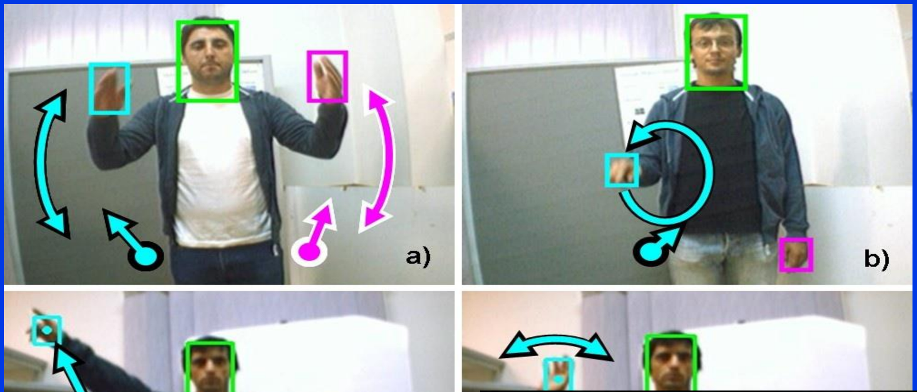
  Either the entire human body or individual body parts are applied for recognition (*human gait, actions;* mostly silhouette-/contour-based – full body!)

- **Action primitives & grammars –**

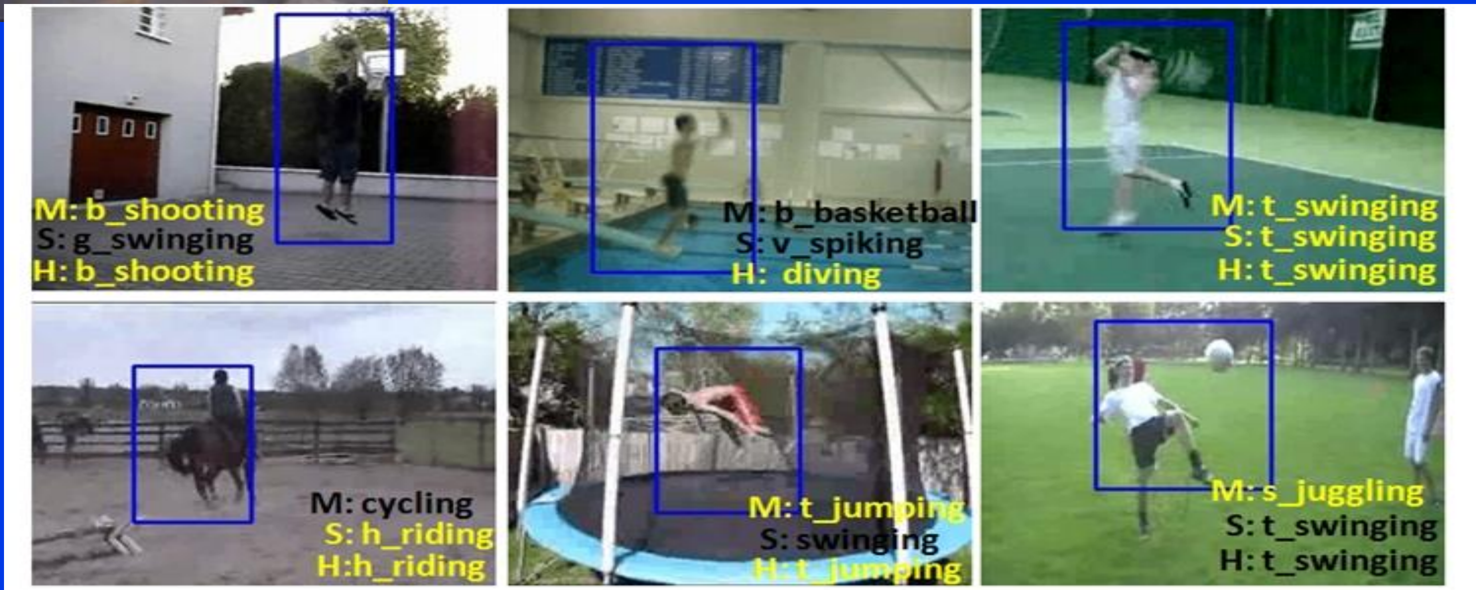
  where an action hierarchy gives rise to a semantic description (parts, limbs, objects) of a scene.

4. Recognition *(cont.)*





4. Recognition *(cont.)*



VARIOUS APPROACHES

←

View-*based* vs. view-*invariant* recognition

- View-invariant methods are difficult
- XYZT approaches try with multi-camera system
- Most of the methods are view-based – mainly from single camera

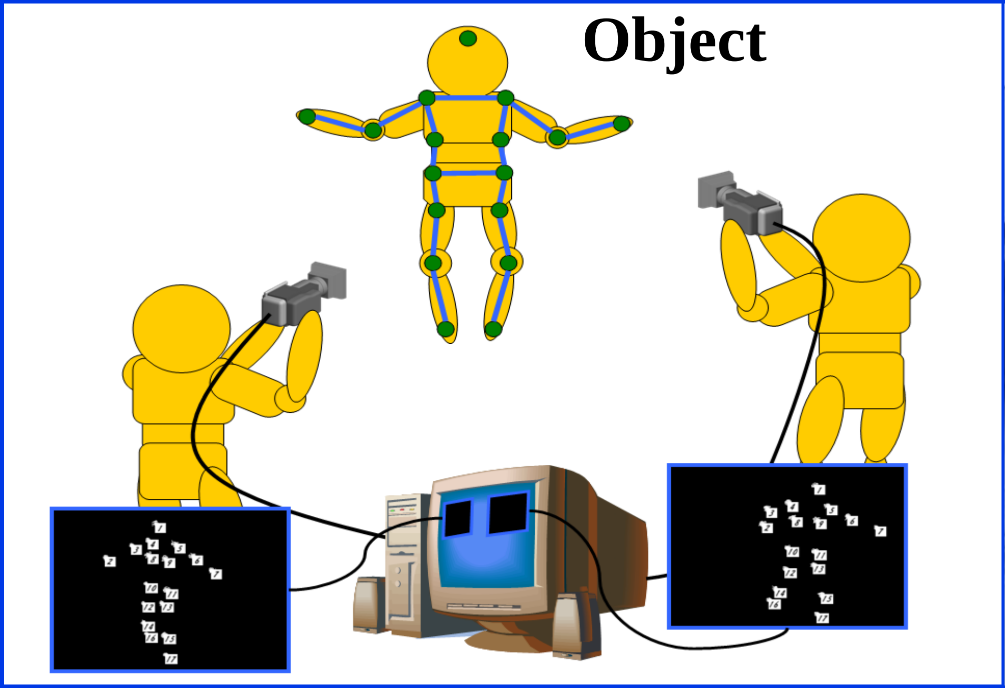
Intrusive/Interfering-based technique

Two techniques to recognize human posture:

- *Intrusive*: track body markers

- *Non-intrusive:* observe a person with cameras & use vision algorithms.

Employing feature points



- Difficult to track feature points.

- Self-occlusion or missing points create constraints.

'Good features to track!'

Spatiotemporal (XYT) features

Spatio(*x,y*)-temporal(*time*) features – can avoid some limitations of traditional approaches →

of intensities, gradients, optical flow, other local features

Spatiotemporal (XYT) features (cont.)

- Space(*X*,*Y*)-time(*T*) descriptors may strongly depend on the relative motion between the object & camera.

- Some corner points in time, called *space-time interest points* can automatically adapt the features to the local velocity of the image pattern.

But these space-time points are often found on highlights & shadows

□ So, sensitive to lighting conditions and reduce recognition accuracy.

Space-time Interest Points

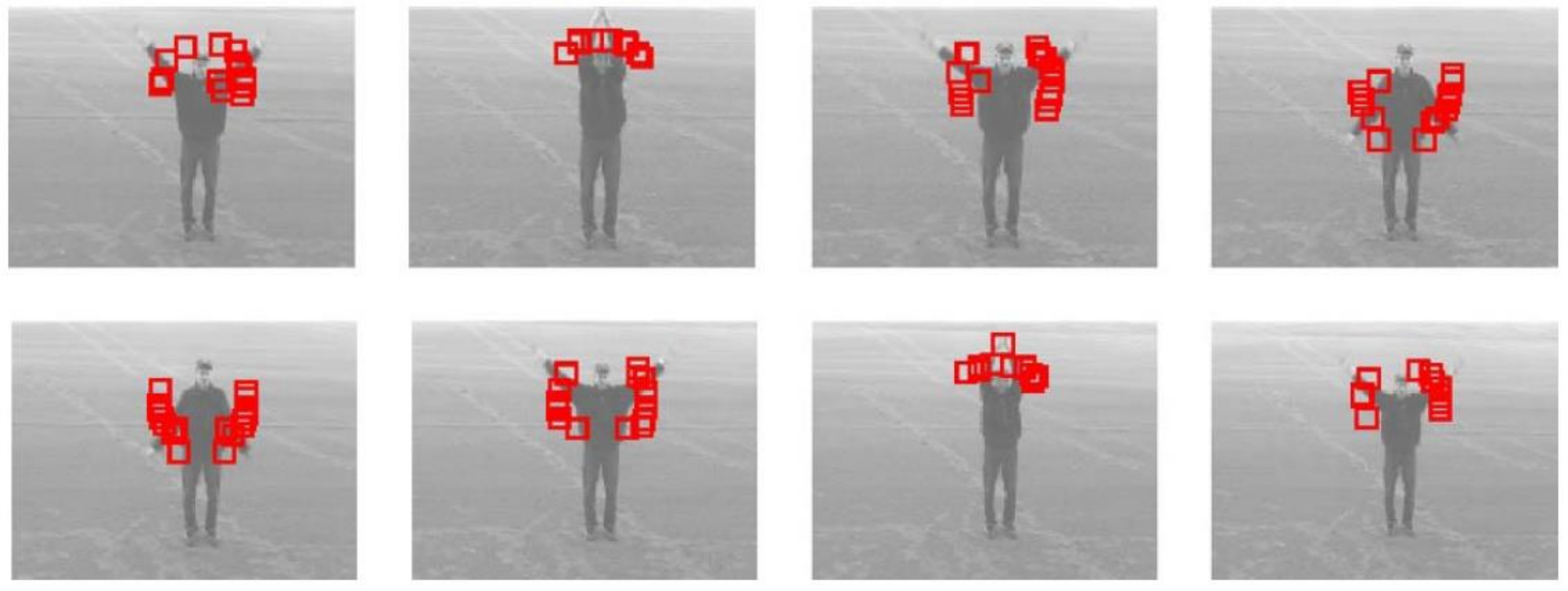


Figure from Niebles et al.

Local Space-time Features

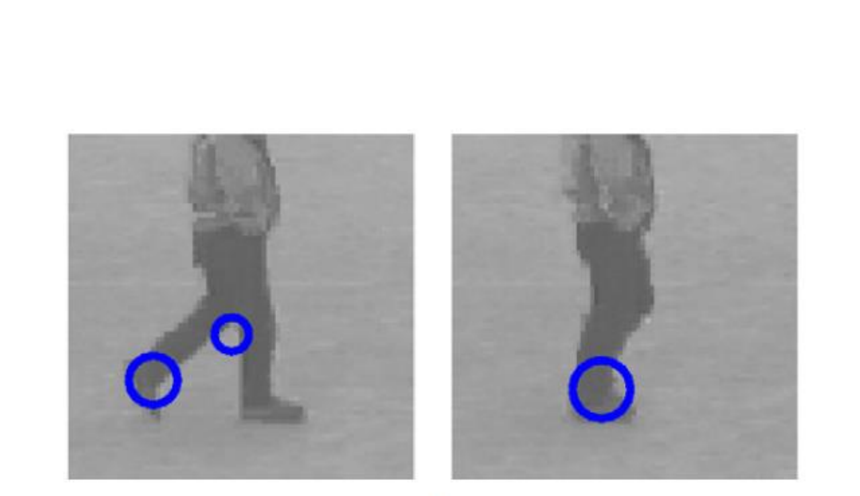


Figure from Schuldt et al.

DATABASES

Weizmann dataset

Run

Side

Skip

Jump

PJump

Bend

Jack

Walk

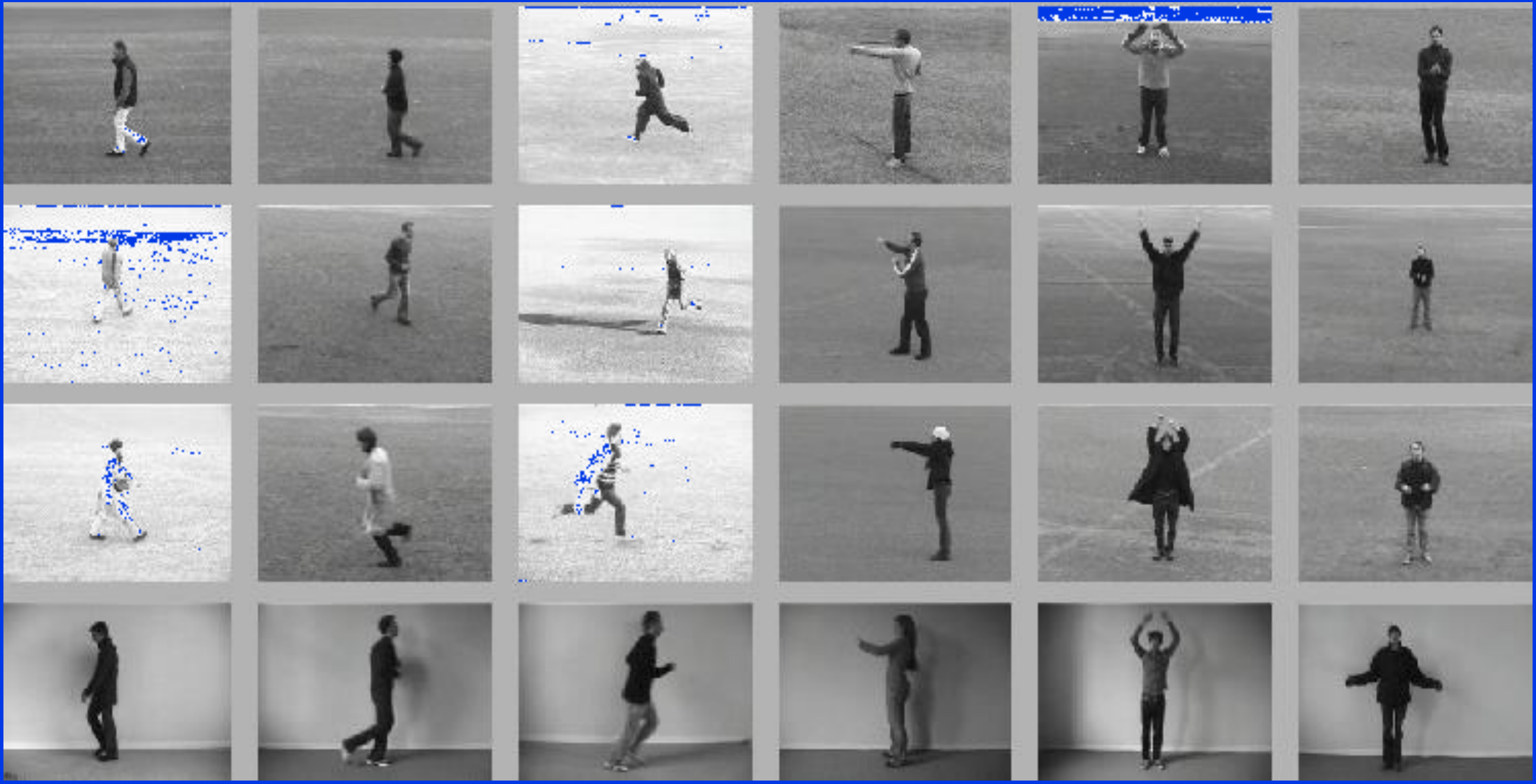
Wave1

Wave2

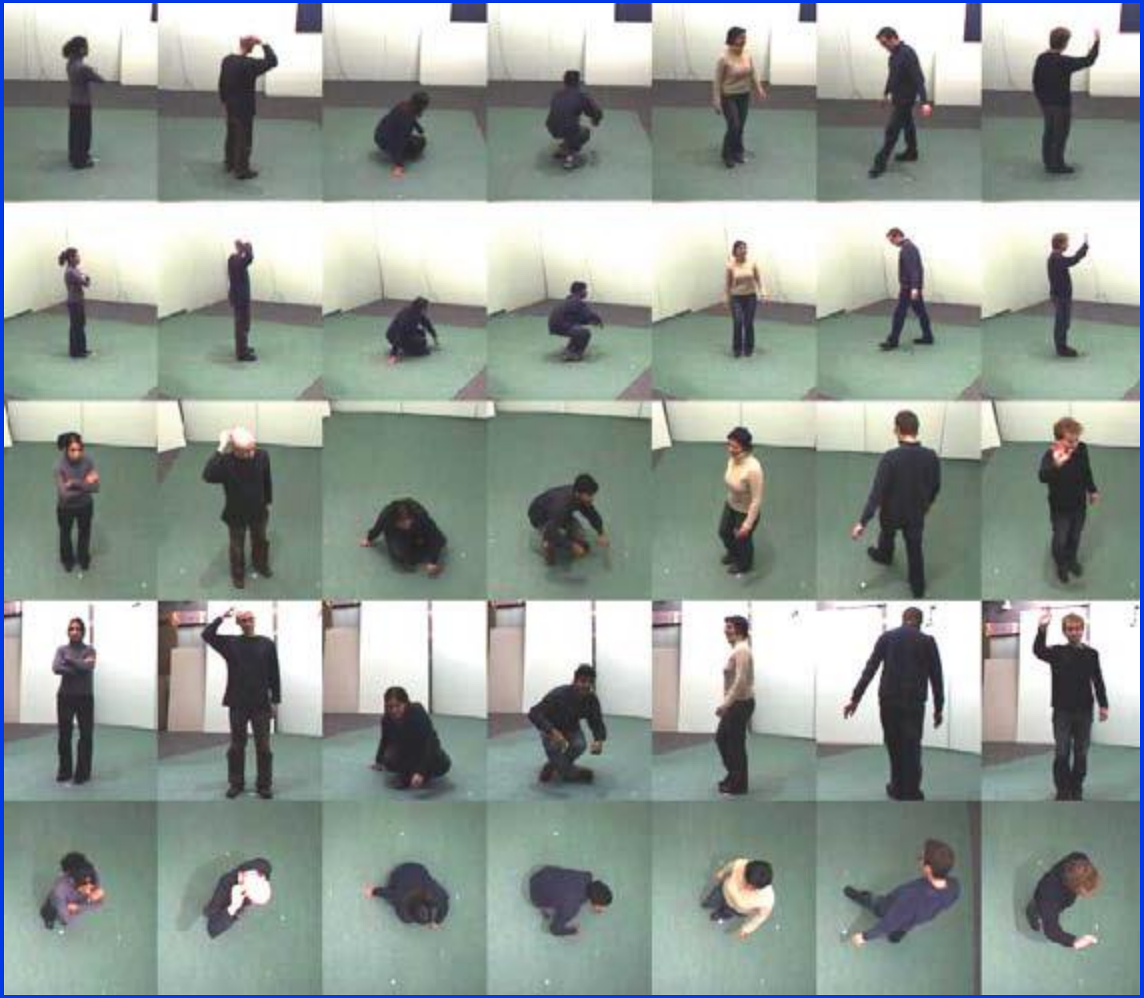
Weizmann dataset –
easiest!

KTH db

Walking Jogging Running Boxing HandWaving Clapping



IXMAS database



Wide-area activity db – UTexas

UT db from Tower

Pointing

Standing

Digging

Walking

Carrying

Running

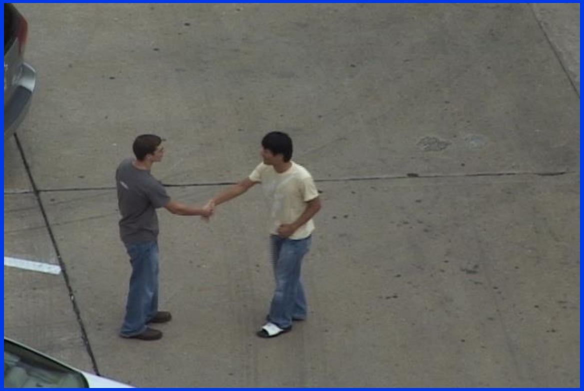
Wave1

Wave2

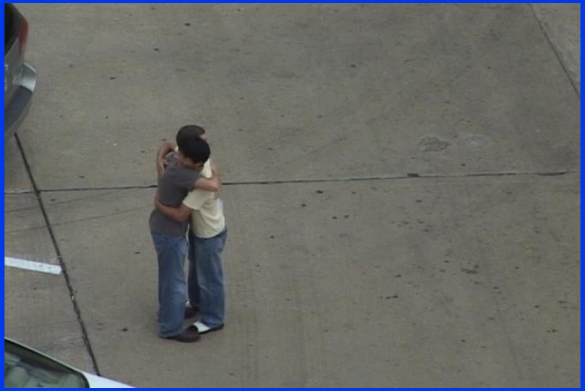
Jumping

2-persons interaction - UTexas

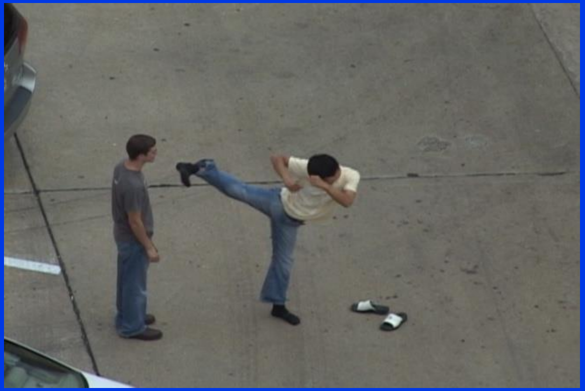
Hand shake



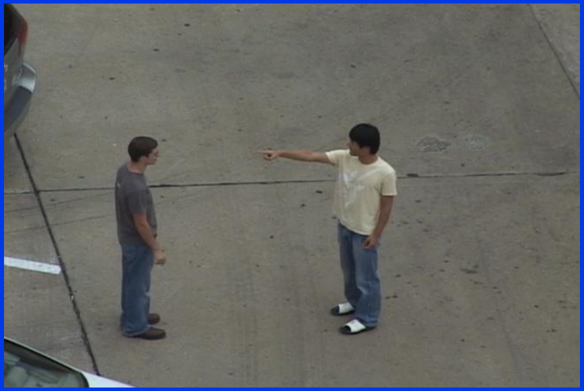
Hugging



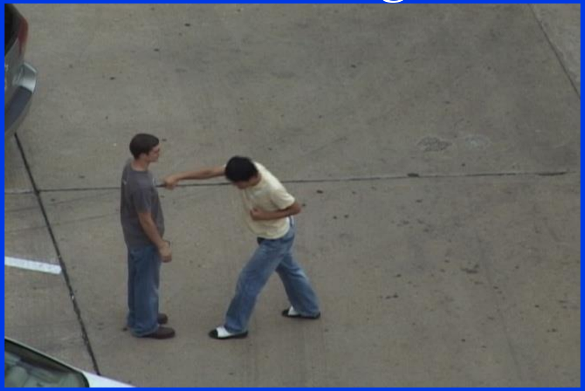
Kicking



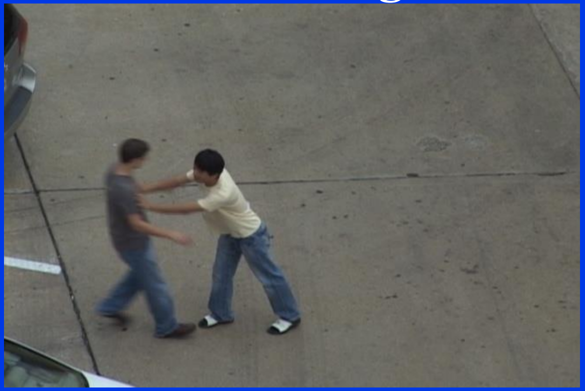
Pointing



Boxing



Pushing



Dataset Employed in PRL special issue

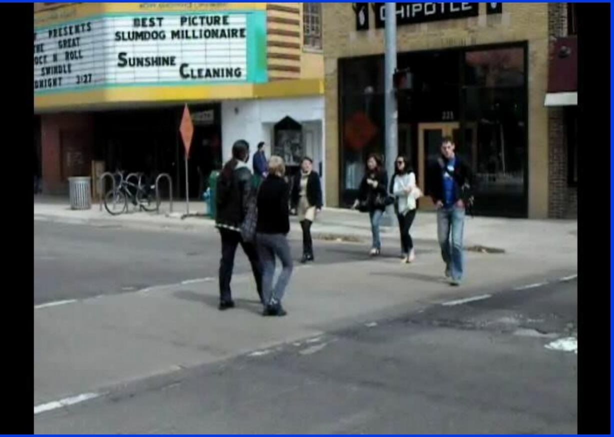
- TUD-MotionPairs dataset
- University of Texas (UT) interactions dataset
- i3DPost database
- AIIA-MOBISERV database
- HMDB51 dataset
- Weizmann database – used by many as it is relatively an easy dataset
- KTH database – the most-widely used dataset
- UCF Sports dataset
- UCF YouTube dataset
- Ballet datasets
- TUM dataset
- IXMAS dataset
- MuHAVi dataset
- Hollywood dataset
- Hollywood-2 Dataset (TV Human Interactions)
- TRECVID2006 dataset
- PAINFUL database

Dataset Employed in PRL special issue …

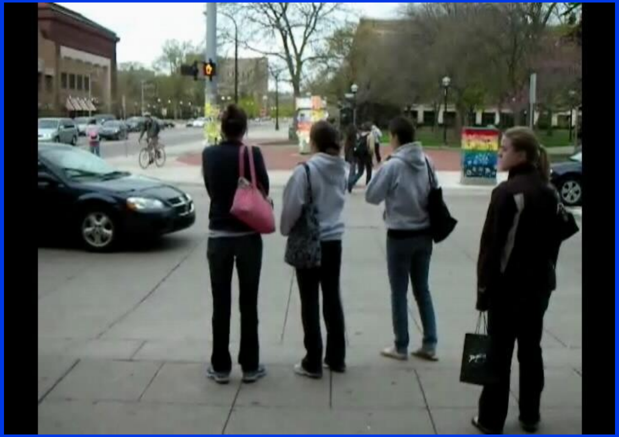
- ChaLearn Gesture Dataset (CGD2011)
- 48 actions from visint.org dataset
- One artificially generated dataset (the first dataset corresponds to a car manufacturing scenario)
- Opportunity dataset, which comprises sensory data of different modalities in a breakfast scenario
- Recordings in laboratory (ShopLab) captured with a fish-eye camera
- Two affective movement datasets (hand movements, full-body movements)
- One unconstrained (in-the-wild) YouTube action dataset
- Database with audio-visual recordings of unwanted behavior in trains, which include aggression in various degrees and normal, neutral situations
- Synthetic data that are obtained from the CMU Graphics Lab Motion Capture Database
- New - Waiting Room dataset 'WaRo11'
- New - the ISI Atomic Pair Actions Dataset
- New - video-tag YouTube dataset
- New - the MMU GASPFA (Gait-Speech-Face) multimodal biometric database that contains audio, video and accelerometer data for 82 subjects

CHALLENGES AHEAD

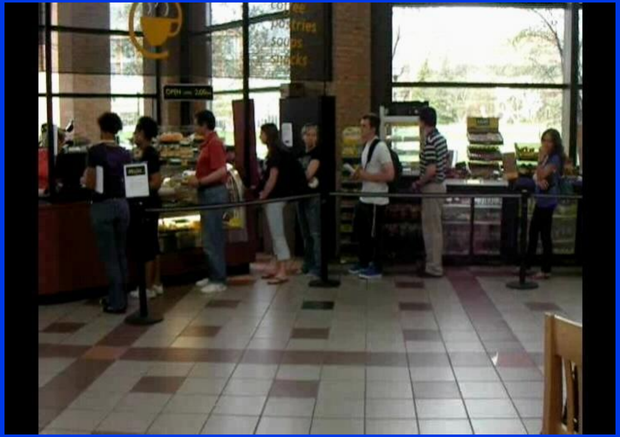
Understanding Collective Activities



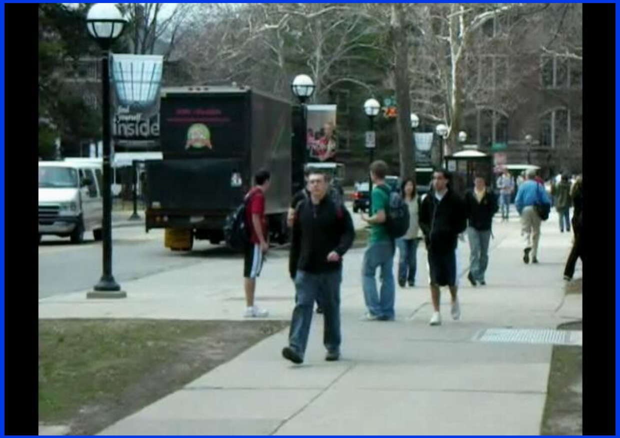
Crossing



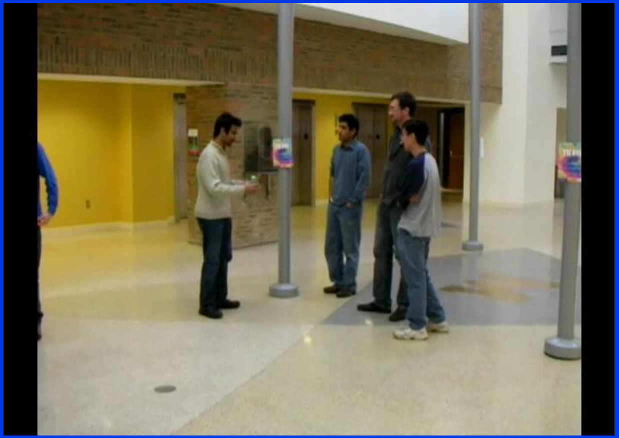
Waiting



Queuing



Walking



Talking

Mass crowd – normal vs. abnormal activities

Escape panic, clash, fight

Difficult to recognize localized activities

→ *that vary from person to person*

Hug

Kiss

Answering Phone

Opening Door

Number of actions or types and variations

→ are hugely varied

→ So difficult!

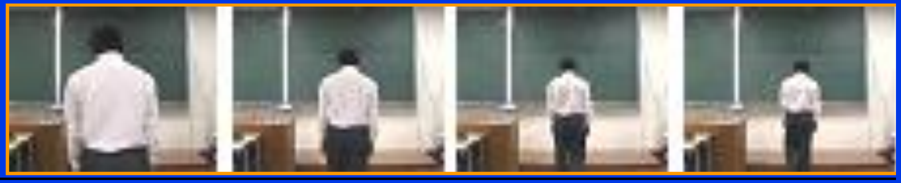
Challenges ahead!!!

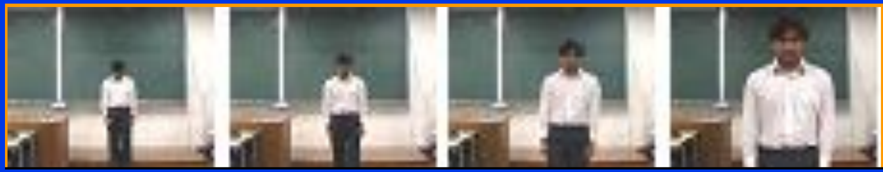
- Human action or activities recognition is difficult due to the presence of various dimensions of motion and the environments.

- 3 important sources of variability are:
  - View-invariance issue,
  - Execution rate, &
  - Anthropometry [size, height, dress effect, gender] of actors.

Challenges ahead - system as view-invariant

- To develop a system as view-invariant will incur time complexity.
- View-dependent methods may fail when the motion is coming towards the optical axis of the camera.





- Motion (e.g., run) are from different directions, diagonal…
- Speed or pace of actions vary

[slow, fast; e.g., jogging vs. running]

Challenges ahead – real-time

- **Real-time** motion recognition is difficult

→ May need prior information, modeling, database or feature vectors to calculate

- **No. of classes:** more classes → slower
- It hinders the performances in real-time.

Challenges ahead – illumination-variation

- Another important constraint is illumination change.
- Most of the works are indoor.
- Outdoor scenes may have → light change, cluttered environment, presence of edges, etc.

- Illumination variations [morning vs. noon vs. afternoon, night, cloudy vs. sunny, etc.] cause recognition problem in most of the approaches.

Challenges ahead – varieties of DB, poor-video

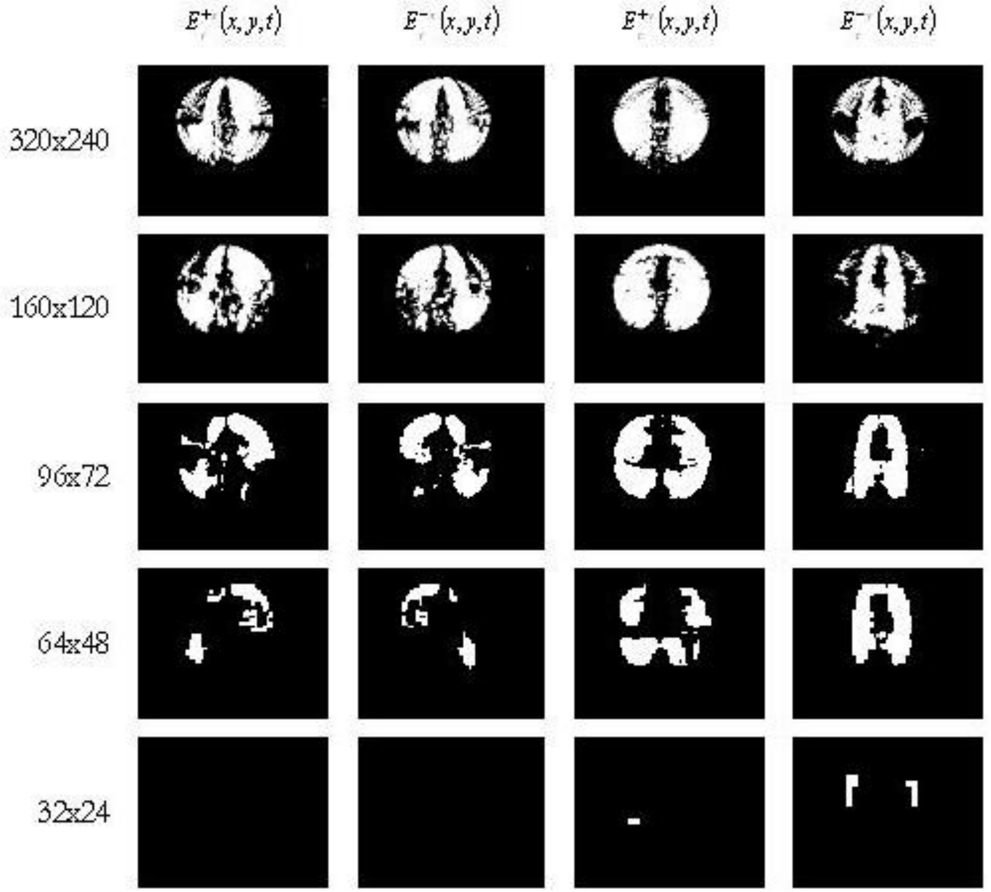
- Issue of dataset: As various methods are analyzed with various datasets, it is very difficult to rationalize the methods & their performances.

- Low resolution and poor-quality video recognition is another challenge in computer vision community.

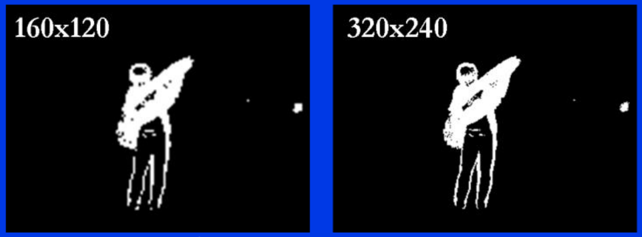
Low-resolution action recognition

Low-resolution image → *Less pixels*

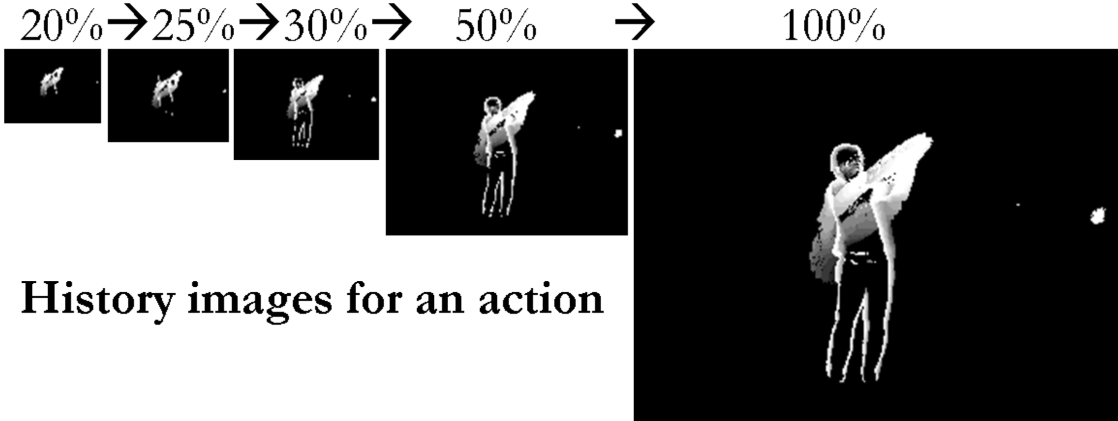
So its processing, recognition → Very difficult.



64x48
80x60
96x72

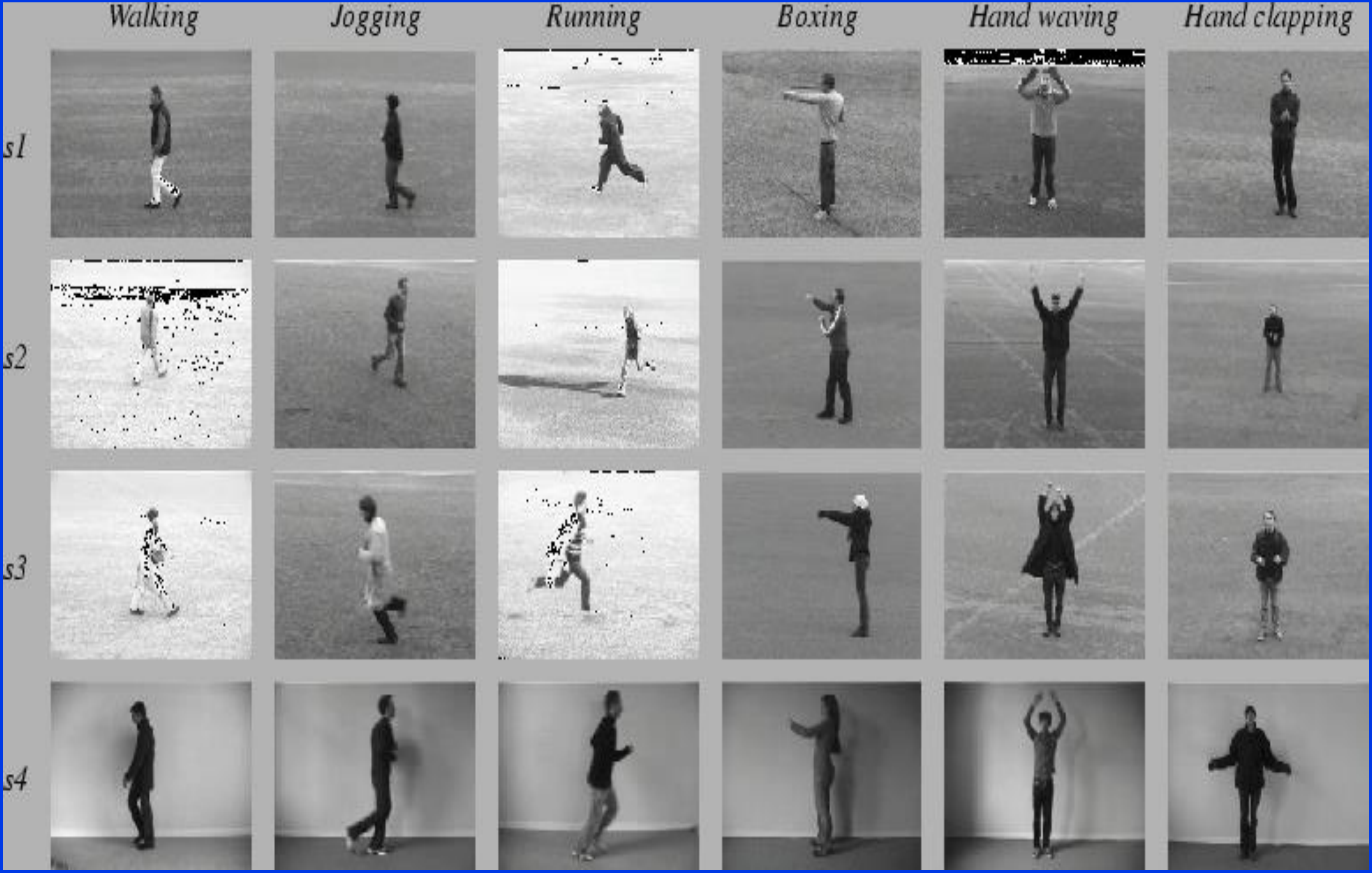


Energy images



History images for an action

Poor-quality video… http://www.nada.kth.se/cvap/actions/



Partial Occluded Video…

http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

- Following actions are ‘walking’ but having varieties – note only 1 person!

Walk with a dog

Occluded feet

Occluded by a "pole"

Swinging a bag

Challenges ahead – applications

- Biometrics issues are incorporating through gait analysis, gesture analysis, emotion analysis through facial expression, etc.
- Robust action recognition → assist human beings.

- Rehabilitation centers as aged people are increasing with less people to support and 'smart-house' concept is important.

Country	**Aged Population**
Japan	65yr+: 20% in 2007 25% in 2030
China	60yr+: 33% in 2050
Korea, some EU countries	…

Challenges ahead – applications

- For Intelligent Transport System (ITS), safety driving, video surveillance, etc. are other demanding areas for smart recognition and behavior analysis -- under --

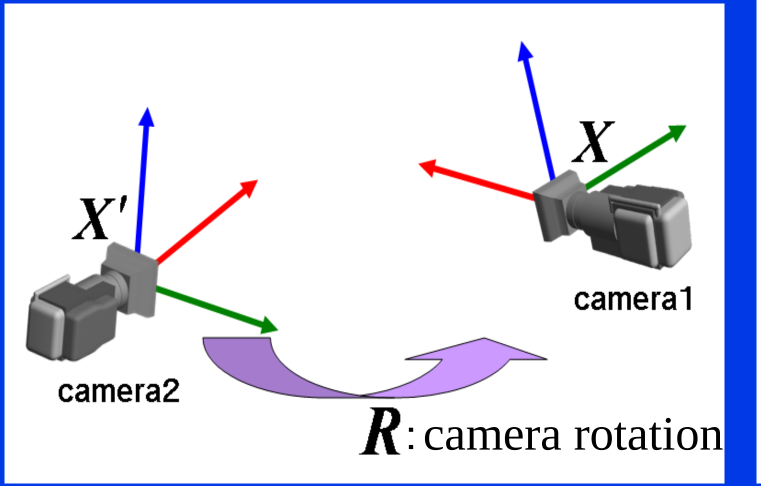
  - multiple objects,

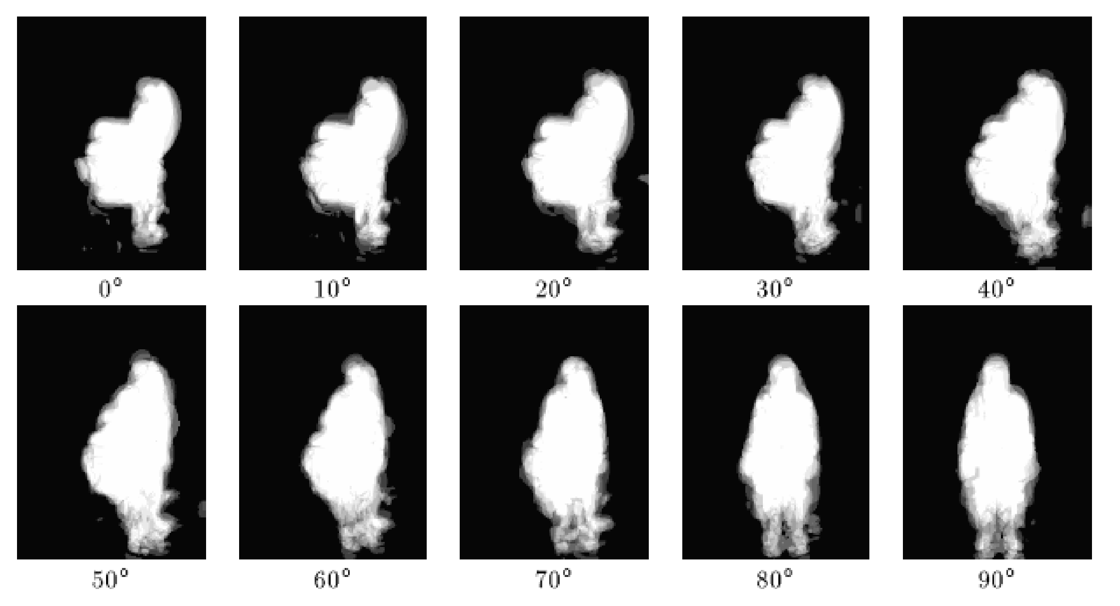
  - image depth,

  - illumination changes, etc.

Challenges ahead – camera motion, multi-cams

- Need camera motion compensation
- Changes in view – same actions may look like a different action from different view





Motion Energy Images for an action from 10 different angles

Challenges ahead – occlusion, etc.

- **Occlusions:** Action may not be fully visible

• **Action variation:** Different people perform different actions in different ways.

• **Background "clutter":** Other objects/humans present in the video frame.

Challenges ahead – emotion

Need good dataset. Getting actors to generate data means

- Intentions are known
- Conditions are controlled
- Sample is balanced

But

- Performances vary massively &
- Transfer to real trials is poor

Need: "rich, *spontaneous* human behaviour"

- Strong interpretation:
  - to detect emotion in a given context,
  - we need training data from that context

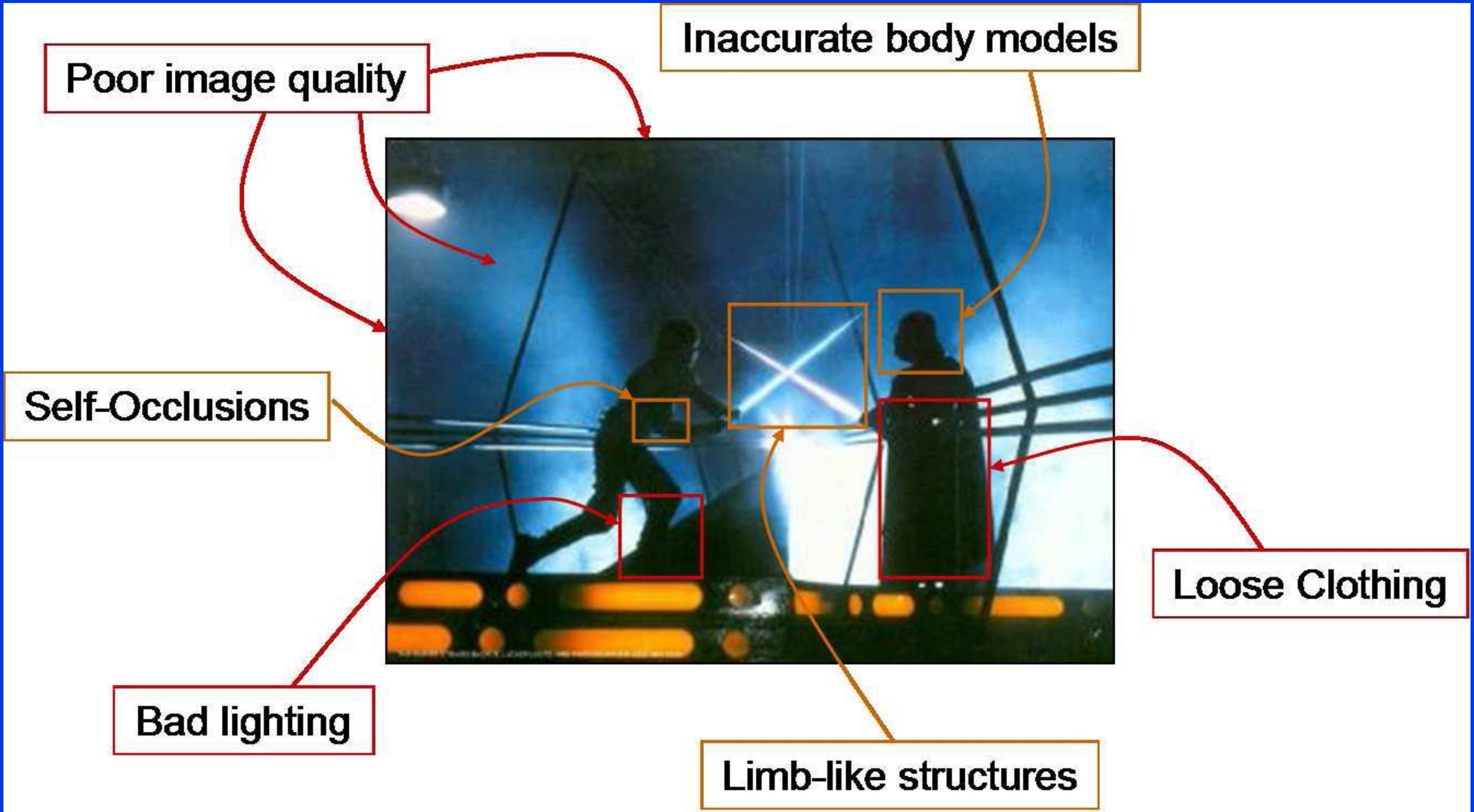
  e.g., HUMAINE database, etc.

Challenges ahead – multi-modality

- Now, most papers consider only video or visual info

- Need to include → multi-modality
  - ✓ text,
  - ✓ audio,
  - ✓ object recognition,
  - ✓ facial action units (FACS/AU),
  - ✓ emotions/psychology,
  - ✓ context,
  - ✓ background, etc.

Problem of Human Motion Estimation



Problems of Human Motion Estimation…

- **Poor image quality:** Grainy images result in noisy measurements, and motion blur obscures limb edges.

- **Self-Occlusion:** Even when a subject is in plain-view, limbs are often obscured by other parts of the body.

- **Inaccurate body model:** At a certain level of detail, any model of the human body will be inaccurate. People come in varying proportions, and a good model must be robust to wide variation in human appearance.

Problems of Human Motion Estimation…

- **Loose clothing:** Even with an accurate body model, loose clothing disturbs limb location & muddles appearance.

- **Limb-like structures:** Without constraints on scene background characteristics for a capture sequence, it is easy to misidentify miscellaneous scene elements as subject substructure.

- **Bad lighting:** Excessively dim or excessively bright lighting conditions make feature detection more challenging.

Conclusion

- Action or activity recognition & analysis – very important
- From video or image to understand
- Global scene vs. localized
- Various challenges – especially in real-life applications
- Applications are based on assumptions & limited action sets.

Sources:

1. Md. Atiqur Rahman Ahad, Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding, *Atlantic Press*, available in *Springer*, 2011.

2. Md. Atiqur Rahman Ahad, Motion History Images for Action Recognition and Understanding, *Springer*, 2012.

3. Md. Atiqur Rahman Ahad, Computer Vision – Datasets for Action & Behavior Analysis, *Springer*, 2013 (to appear).

4. Special Issue, SAHAR, *Pattern Recognition Letters, Elsevier*, 2013.

5. Various other papers.

Thanks a lot!

Join 6th ICIEV, 1~3 Sept. 2017

University of Hyogo, Japan!

http://cennser.org/ICIEV