Page :
Date :

→ Rec 1

* medium = technical means to store, Transfer & convey information
  Physical channel = air / lights.

* Modality = human sensory channel.
  visual channel = see
  auditory channel = hear
  olfactorial = smell
  haptic = touch
  Guctatory = taste
  Vestibular = Balance

→ temp sensor

* Senses	organs	Modality	sensor
vision	eyes	visual	camera
Hearing	ears	auditory	microphone
Touch	skin	Haptic	Touch screen, gloves
olfact.	nose	olfactorial	}
Taste	Tongue	Guctatory	}
Balance	equilibrium organ	vestibular	}

human perception — for system

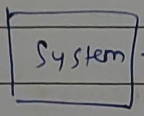
Appln
Gesture, Face, Hand.
Speech
Posture / motion recog.

Modality

= capability of system to mimic such human sensory channel.

multimedia vs. multimodality

* multimedia = multiple media, system with more than one way to transfer info from system to user.



* multimodality = system that can stimulate more than 1 human sensory channel.

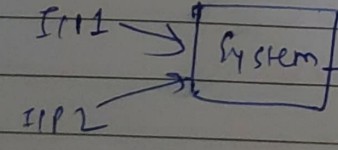
Ex: TV → screen (visual) } multimedia
TV → loudspeaker (auditory) } multimodality

It is a multimedia

A system can't be both multimedia & multimodality

* Alternate def

multimodal system = A system that provides more than 1 user IP to interact with it.



TV = no user IP

Smartphone = multimodal. (touchscreen + microphone)

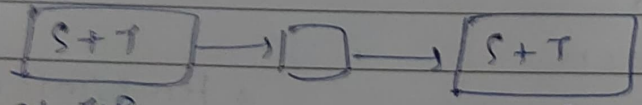
focused on o/p side

2nd modified def

* multimedia system = interactive system that provides information via several o/p channels (sound, graphics etc.)

focused on I/P side

* multimodal system = processes 2 or more combined user I/P modes (such as speech, touch, gestures, hand, head etc.--) & o/p gives that many no. of modes



at I/P side

at o/p side

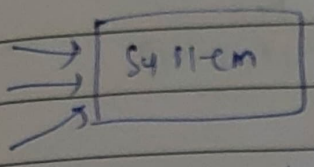
Sharon oviatt (2012)

* for a truly multimodal system, info we provide at the input must be related & coordinated to give o/p

* multimedia = o/p of the system

3rd def (Benoid)

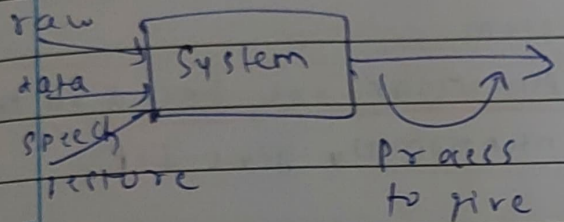
* A multimodal system represent & manipulate information from diff human communication channels. These system automatically extracts meaning from multimodal raw data & gives desired information



IIP = human sensory channel

OIP = desired OIP

↳ human sensory channel | at IIP then it is multimodal



* Domains of multimodal system :

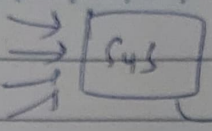
i) Active IIP mode
ii) Passive IIP mode
iii) virtual environment

* Put that there = 1979 MIT

↳ speech + gesture (voice)

Page :
Date :

raw data



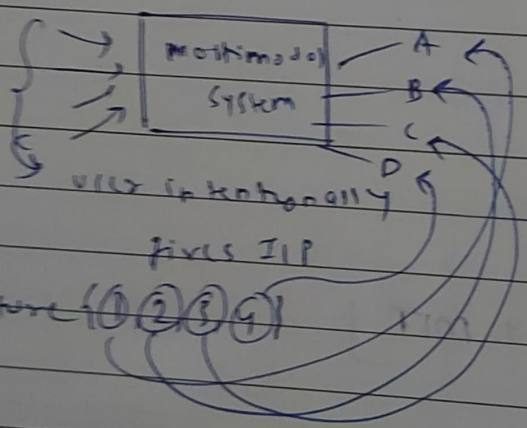
Lec 2

* multimodal systems are subset of multimedia systems.

* Domains of multimodal System
  → max. Applications are based on this

1. Active I/P mode = User intentionally give I/P
2. Passive I/P mode
3. virtual environment (Gaming Appln).

1. User can intentionally give information to the system.
   forex: fixed set of commands, gesture



"We know which type of I/P to provide to give a particular o/p"

2. A dialog system with diff I/P modes. Chatbot can be of 1 example."

chatbot

We input some text & we get fixed set of answers.

* fixed I/P to get desired o/p

P Q R

6 4 6

8 1010
1000
1011
6+17+10

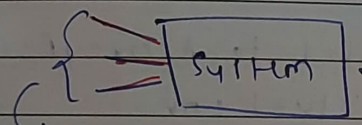
8 10
1100
110
1010
8 2

8 4 16

1000
1000 0100
10 0000

(1) Passive I/P mode = All kinds of inp that user do not produce intentionally.

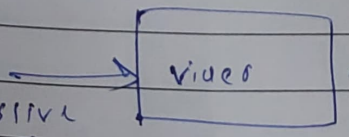
Ex: emotions, gazing at something
≡ ↳ sad, happy etc.



↳ not intentionally provided by the user.
(no fixed pattern at I/P end)

Unlock a phone with happy face = intentional

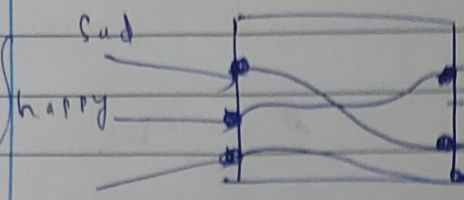
Ex: Depending upon emotions of classroom facial exp of students
(how ⇒ he captures video of class recordings)



→ real time feedback of classroom (is it boring/interesting)

↳ how can classroom be made interesting / whether to continue class or not / improve skill of professor

"fixed Pattern"



① switch on TV
② ON AC
③ open DOOR
④ ON DOOR

Intentionally to get o/p (so this is not Passive i/p but active I/p.

(3) virtual env = User can behave naturally to perform a task.

OR

Users can use all kind of information, behave naturally as he/she is in natural environment & the system would be able to analyse all this information & act accordingly ".

* 1 multimodal appln. related to Active/Passive/ virtual env.

1.4.3

* Active i/p mode = i/p is intentionally provided.

I/p → [system] → o/p

intentionally provided.

(command to comp.) directly

speech
pointing to something
gesture
writing

i/p

* Passive mode

naturally triggered → System → o/p

(not give command to computer directly)

(perform this task X)

gaze, brain wave patterns, lip movement, facial exp

* virtual env (can take up both active & passive i/p)

i/p → [ ] → virtual exp at o/p end. (real life experience kind of thing we get)

may be anything (Active / Passive)

* Recognition process will be done for both active or passive input.

* Process recognition is common for any system.

modality relations

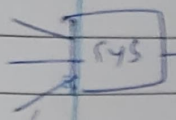
CARE = Complementarity, Assignment Redundancy, equivalence

"Properties of multimodal systems"

Page:
Date:

Complementarity

I/P



multiple modalities are used together at a time to reach o/p

Ex: Speech + gesture

both will work at same time

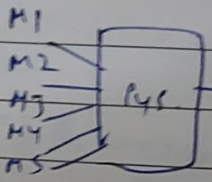
M1 + M2 → O/P (at same time used)

Ex: 1979 MIT "Put that there" (Pointing hand + voice) at same time

All I/P work together

Assignment

only 1 modality will be selected for a certain piece of information



I/P

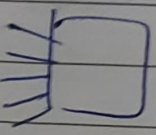
"one modality will be selected at a time to give o/p"

"when we operate 1 modality, other modalities will be switched off"

1 of modality work at a time

Redundancy

only a part of information will be used.



Ex: MIT pointing hand + voice

PUT that red circle in blue if only 1 triangle & circle, then gesture (Pointing hand) is redundant) voice is enough

equivalence

Any available modality can be used with combination of anything

"Sequential" OR "parallel" OR "using Assignment property"

any kind of system

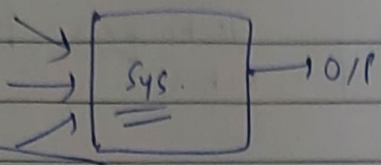
System (system will work with any combination of inputs (whether sequential, parallel or Assignment)

"A system can be of multiple types"

Types of multimodal systems:

Fusion

① Fusion based mm / Blended mm system / interface:



Atleast
1 active I/P
+
1 passive I/P

Active Passive

Ex: Hand Gesture + gaze

Active	Passive
Hand gesture +	gaze
+	

major application based on this

LipNet ⇒ works on CNN

↳ Lip movement = Passive

↳ Speech = Active.

(recognize lip movements even if noise in background)

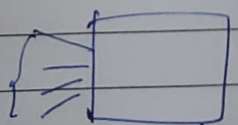
(MIT media lab)

detect lip movement & speech

Autonomous (used in vehicles)

② Temporally cascaded MM system "much less explored"

Process 2 or more modalities sequenced in particular temporal order.



Sequenced in particular temporal order

M1 → M2 → M3 ⟶

Ex: First gaze then gesture, speech.

Camlin

Page :
Date :

Ex: A cascaded multimodal natural user interface to reduce driver distraction (IEEE paper)

- ↳ Speech + button
- ↳ Speech + touch
- ↳ gaze + button

} different cascading techniques for autonomous vehicles.

Ex: (Speech + gaze = "Turn on music" + (increase volume by moving hand up or down))

Lec 4

used to capture diff information (gesture are the knobs) => for recognition purposes

Ex: A cascaded multimodal natural user interface to reduce driver distraction (IEEE paper)

- ↳ speech + button
- ↳ speech + touch
- ↳ gaze + button

} different cascading techniques for autonomous vehicles.

Ex: (Speech + gaze = "Turn on music" + (increase volume by moving hand up or down))

Recap

if bad lighting conditions = problem may be there

facial recog => 3D model based techniques, variation of angle of hand is used & detection is used.

Hand Gesture => video is taken & then frames of video taken to capture hand

* 1 Gesture can be provided in several ways.

* Application of facial recog

↳ Viola-Jones algorithm

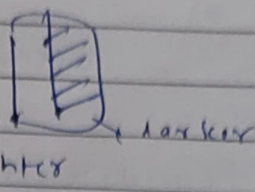
Viola Jones = to extract facial features using haar-like features.

↳ can be used to train diff. objects/body parts.

Challenges if face detection ⇒ identification
recognition.

Page :
Date :

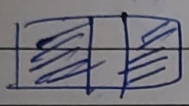
* Like Haar like features



8-10 haar like features
(combination of black & white patches)

these patches made to detect features from face

to extract information about nose



* Different haar like features for eyes, nose, lips, mouth (to extract features from facial image)
* By default, it select 4 facial features:
  1. eyes
  2. nose
  3. mouth
  4. mouth + nose

(II) Creating an integral image = facial features are combined to extract information.

(III) Adaboost training

(IV) Cascading

→ can also detect other part of body

→ If 8/10 features matched, then it is face of that particular person A

Camlin

Page :
Date :

Lec 6

occluded = overlapping

* Challenges of face detection / things that affect face detection also
1. Pose
2. Presence or absence of structual components
3. facial expression
4. occlusion (many faces are there)
5. Image orientation
6. imaging (lighting) conditions

To deal with challenges

Knowledge based method = rule based method

mathematical based method

(dist b/w eyes & nose, ~~face~~ nose & lip)

"most / much used"

Problems: if we take mathematical calculations, then these calculations may change with orientation change of face (if orientation variation, change of camera)

" if face is static (face is not moving much) "

* Feature - invariant approaches → invariant to expression faced

→ Some features generated from face which will be invarient to lightning conditions, pose, angle & so on

→ diff methods to find those feature points.

→ used because of its invariant prop.

→ widely used

→ edge detection, texture of face.

→ All features combined to detect the face

Camlin

Page :
Date :

Template matching → not much used

→ match template with current image
→ correlation based method

Cons = template & image must comprise of almost similar image otherwise detection may not work correctly (if not much diff. b/w input image and template)

→ correlate with templates stored in folder
→ if whole image is only rotated, then template will work but if face is rotated by θ, then will not work.

Appearance based method = widely used

→ training based method.
→ folder ⇒ lot of illumination variation, diff faces, angle variation.
→ train
→ for detection.

* SVM, NN

→ Train dataset & rule information to detect face.

* Diff types of features :

1. SIFT (Scale invariant feature transform)
2. HOG (Histogram of gradient) ⇒ colour info
3. LBP (Local binary patterns) ⇒ edges & how allocated.
4. SURF (Speeded up robust features)
5. DAISY features.
6. ICF (integral channel feature)

* HOG => used for colour info & info about edges / lines.

lec 7

→ Knowledge based methods

↳ ex: Attendance marker (Person has to come at camera & mark attendance)

→ Feature invariant

→ Template matching = used for recognition purpose (whether particular student present in a group of students picture)

feature comb = nowadays many used.