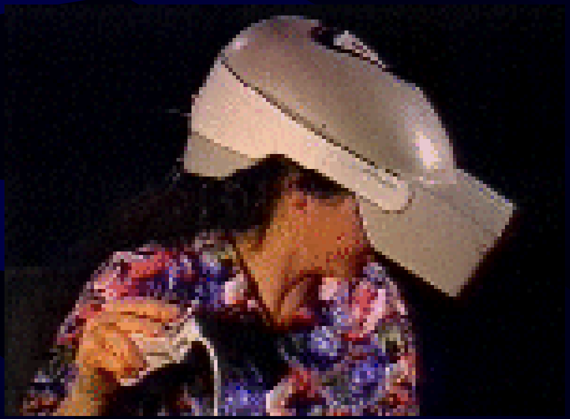
Stereo-based Hand Gesture Tracking and Recognition in Immersive Stereoscopic Displays

Habib Abi-Rached

Thursday 17 February 2005.

Objective

- Mission: Facilitate communication:
  - Bandwidth.
  - Intuitiveness.
  - Efficiency.

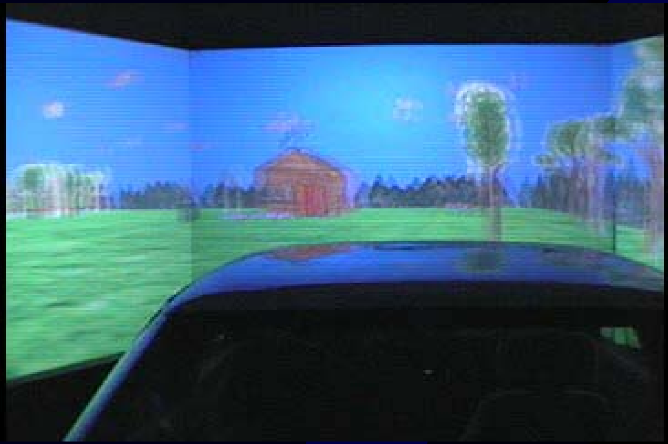


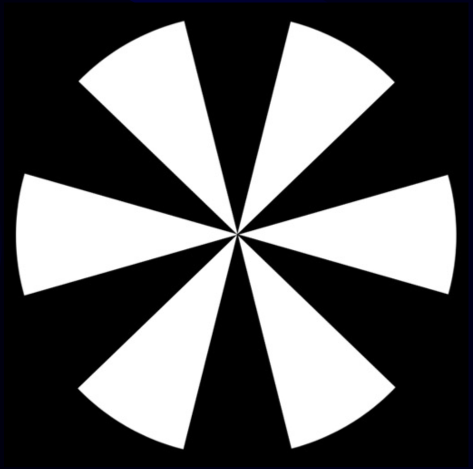
- Means:
  - Visual (Displays, HMD …).
  - Gestural.

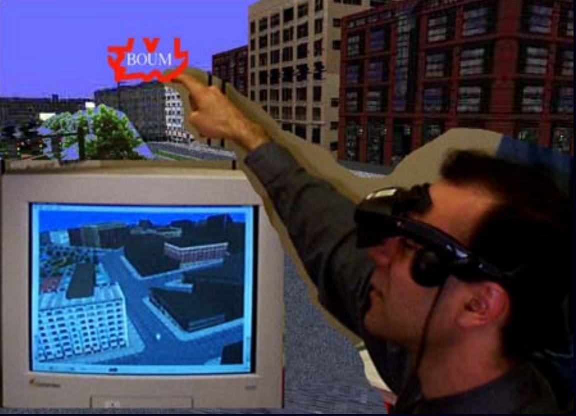
Initial Exploration. (Kodak).

- Domes.
- Driving simulators.
- Cave like environments.

➢ *Simulator sickness.*

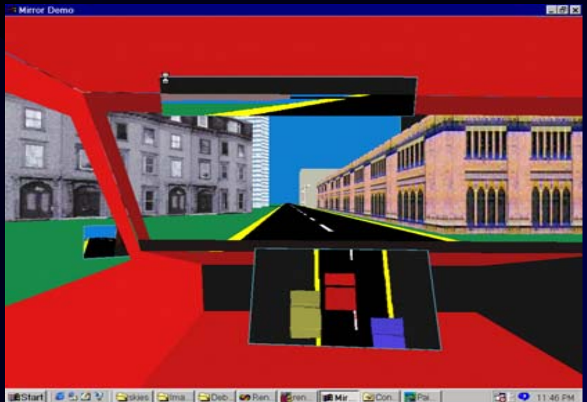






Initial Exploration. (Ford).

- Accuracy of the user's mental models based on visual displays.
- Usefulness of stereo displays.



Limitation of Current Technology.

- Limited efficiency.
  - Mouse Keyboard…
- No 3D. (Monitors).
- Small FOV. (Monitors).
- Few Degrees of Freedom. (Joysticks, Mice).
- Limited intuitiveness.
- Physical connection.
  - (Gloves, Mice, HMD, phantom, polhemus).
- Precision depends on distance.



Hand Gestures

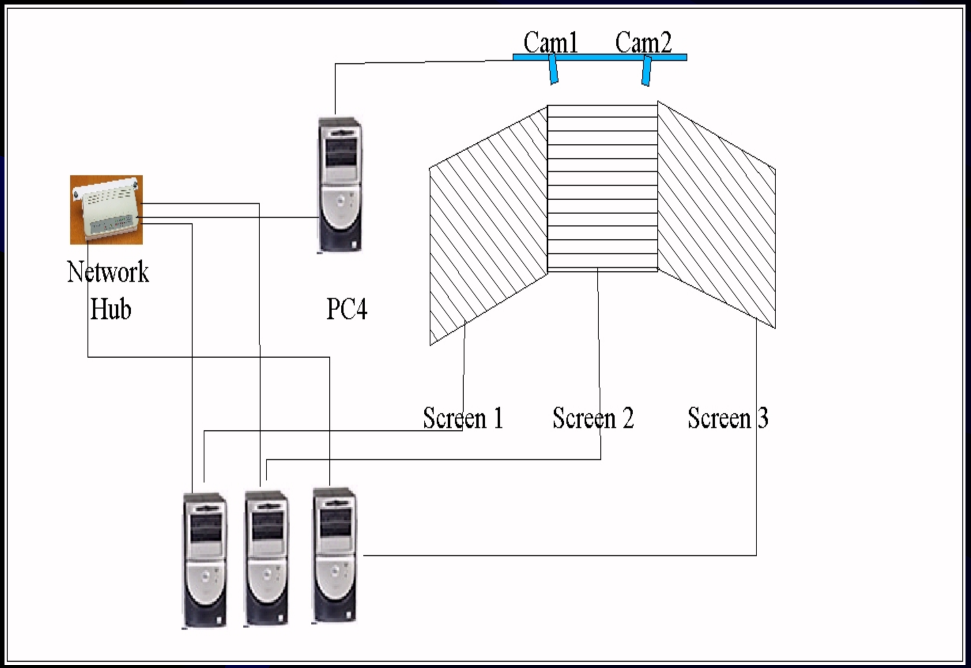
- Human-computer interaction (HCI) has become an increasingly important part of our daily lives.
- Keyboards and mice are the most popular mode of HCI.
- Virtual Reality and Wearable Computing require novel interaction modalities with following characteristics:
  - in a way that humans communicate with each other.
- Hand gesture is a natural and intuitive communication mode.
- Other applications: Sign Language Recognition, video transmission, and so on.

Introduction

- Vision-based recognition of dynamic hand gestures is a challenging interdisciplinary project.
  - hand gestures are rich in diversities, multi-meanings, and space-time variation.
  - human hand is a complex non-rigid object.
  - computer vision itself is a ill-pose problem.

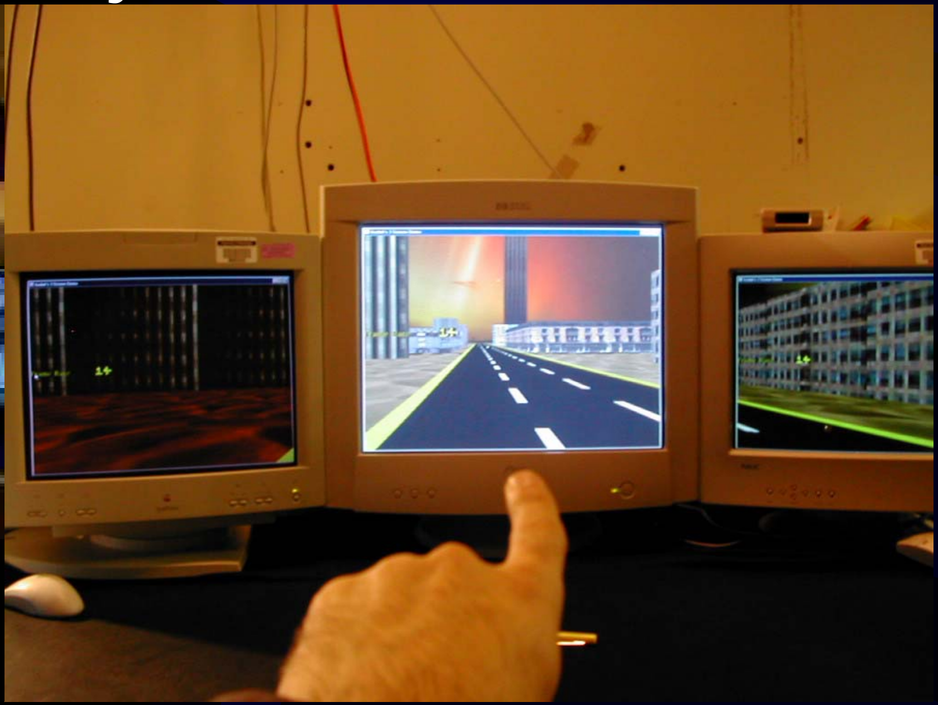
Our Approach.

- Inexpensive immersive PC-based gesture tracking / recognition System.



Gesture-based Interaction With 3D Displays.

- Intuitive interaction, easy to learn.

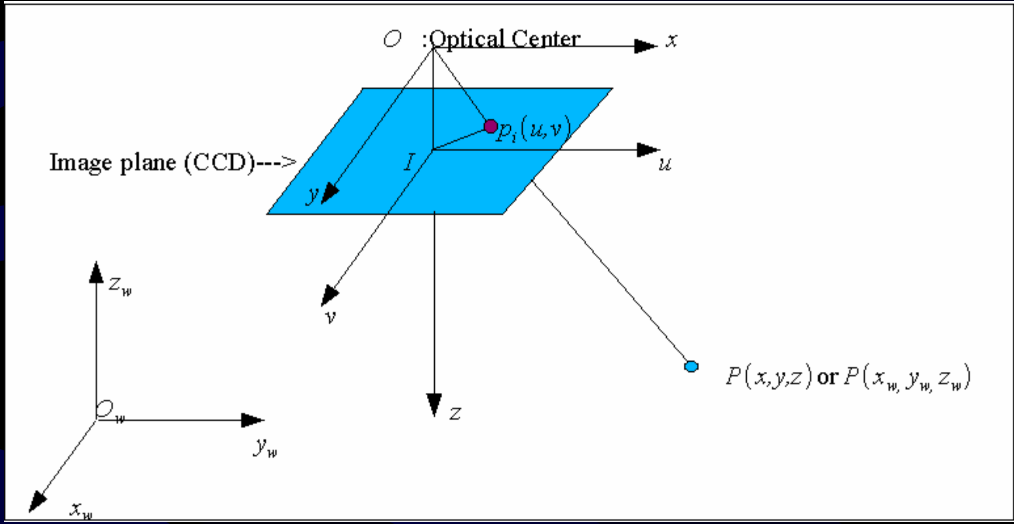
Previous Gesture tracking and recognition methods.

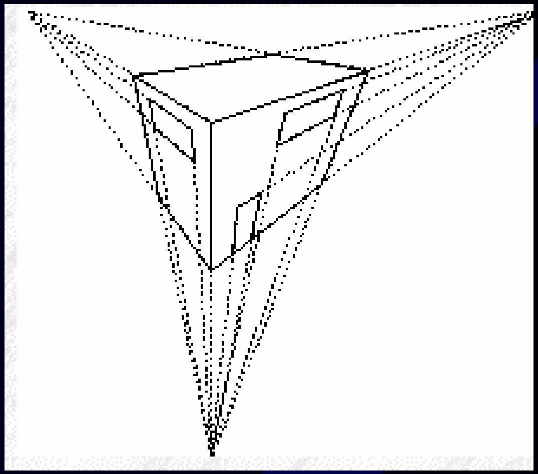
- Temporal modeling and recognition: *(Kendon-MIT)*.
- Spatial modeling and recognition:
  - Appearance-based approach:
    - Predefined static image templates. *(Freeman)*.
    - Deformable 2D templates. *(Taylor)*
  - 3D hand model
    - Volumetric models.
    - Physical models.
    - Skeletal models.

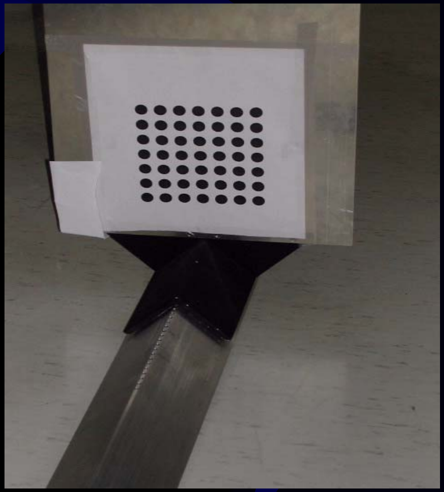
- Feature detection and recognition.
  - Huang (silhouette).
  - Darell (whole image).
  - Essa (spatio-temporal motion).

Calibration methods.

- Tsai method.
- Stringa method.
- Faugeras method.
- Caprile method.

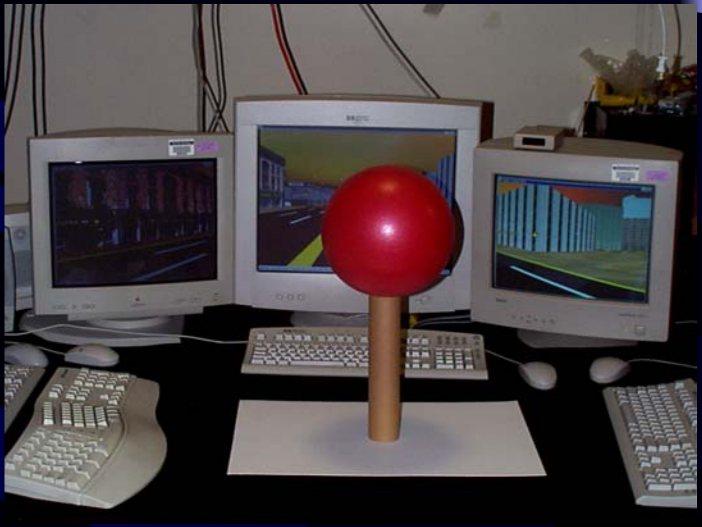


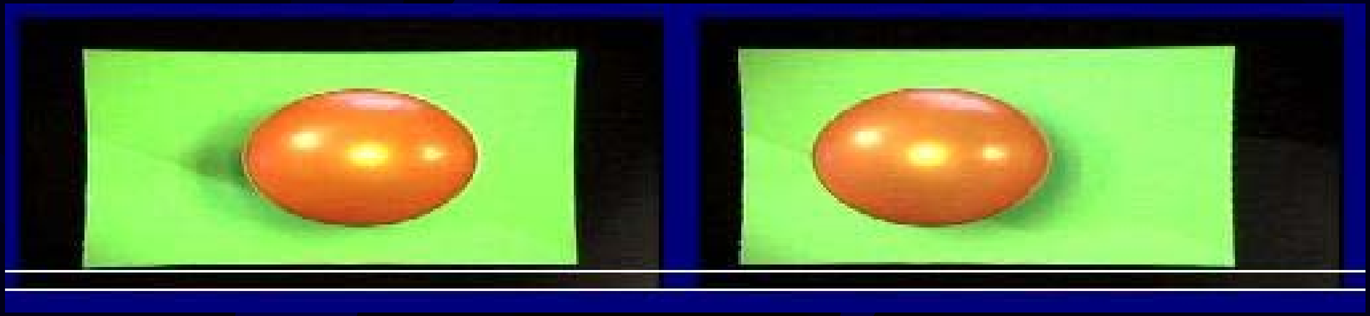




Why develop our own calibration.

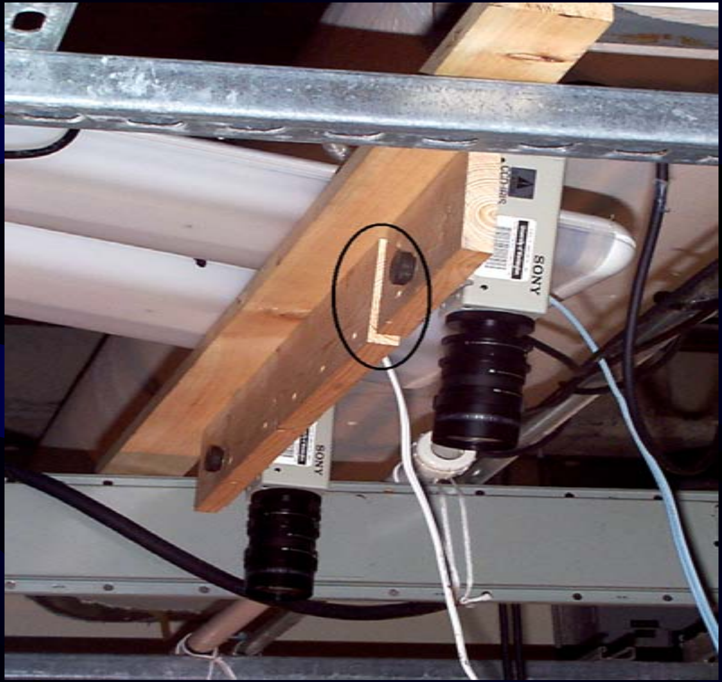
- Simple, inexpensive calibration tools.
- One iteration.
- Orthographic cameras.
- Vertical cameras.

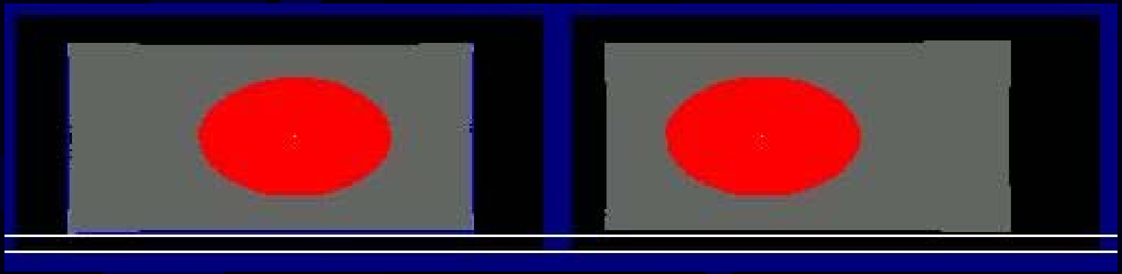




Why develop our own calibration.

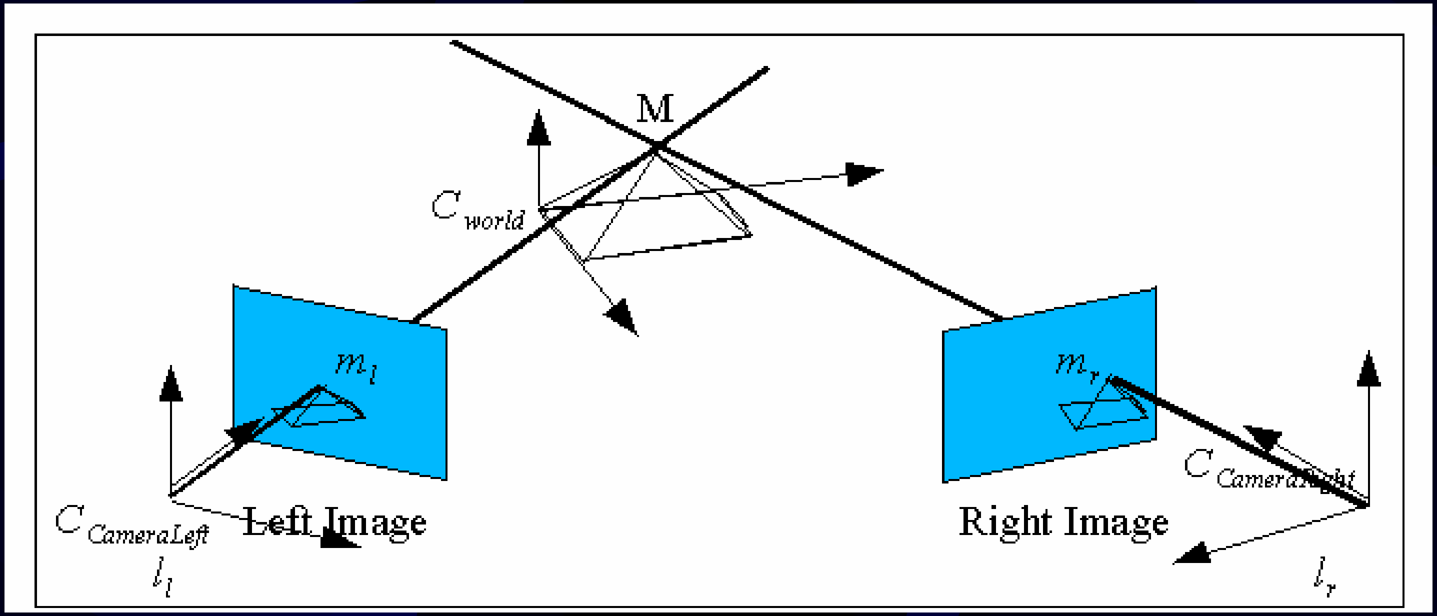
- Faster stereo reconstruction
  – Orthographic projection.
  – Simple complexity.
  – No rectification phase.





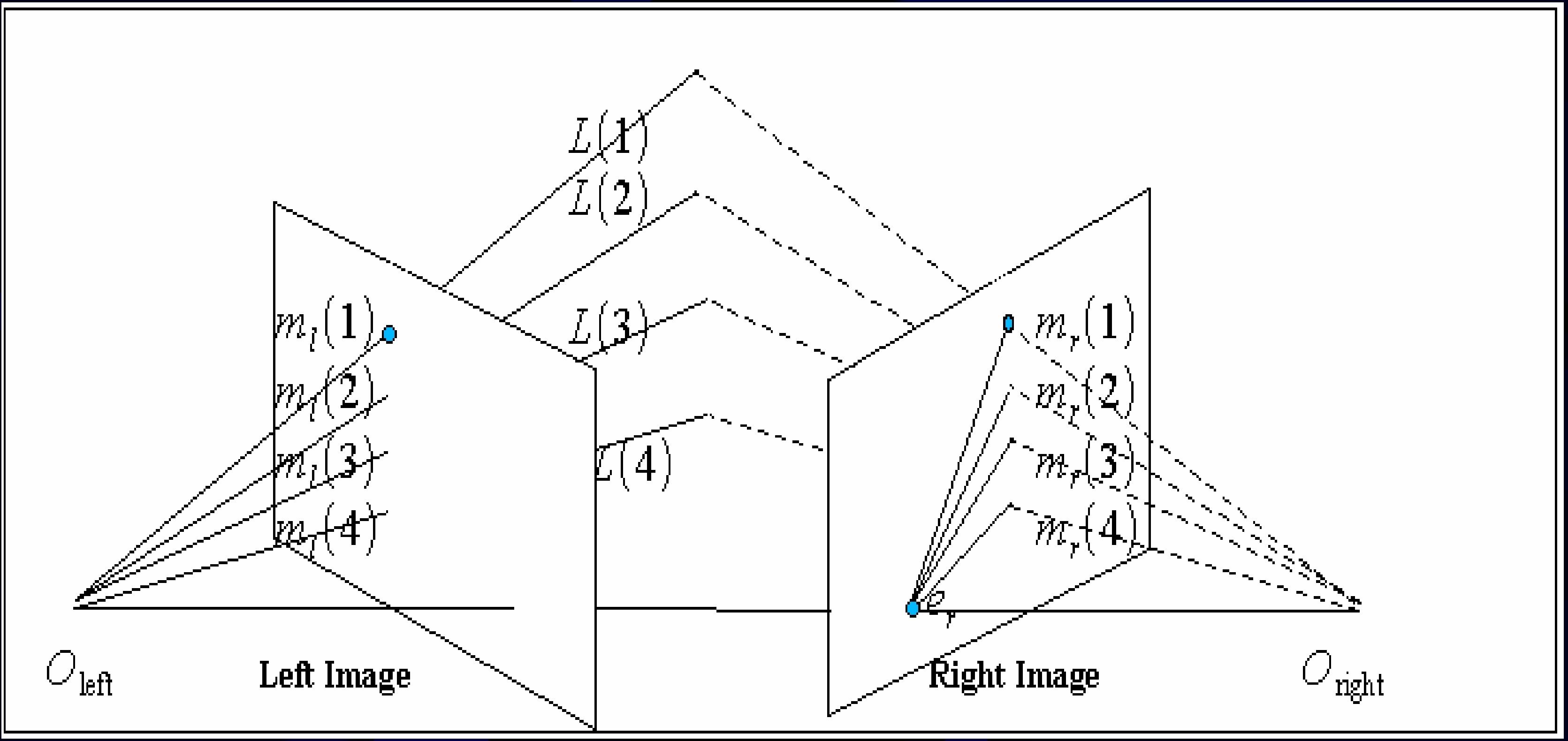
Stereo Reconstruction.

- Matching process.
- Triangulation.



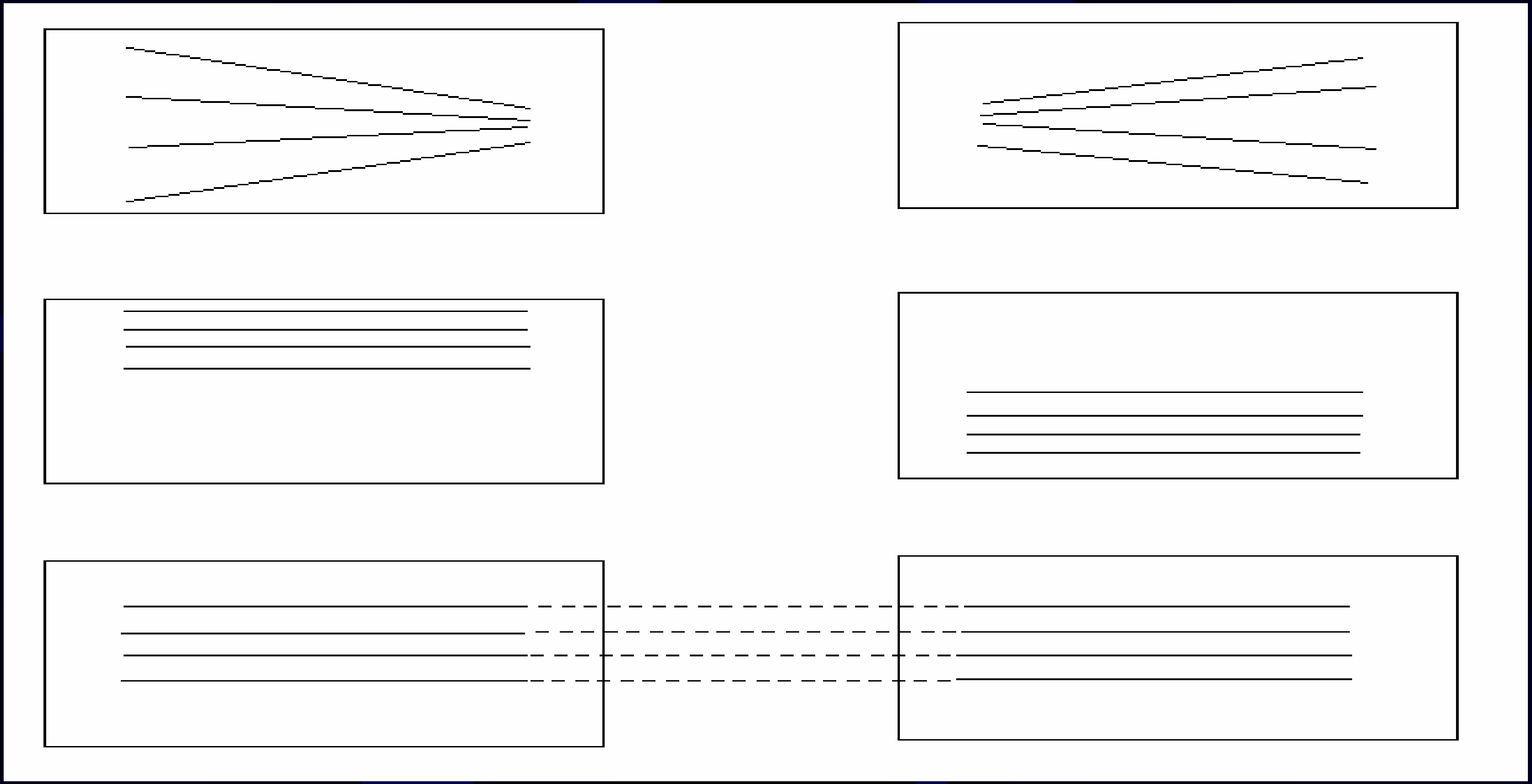
Epipolar lines.

- One dimensional search.

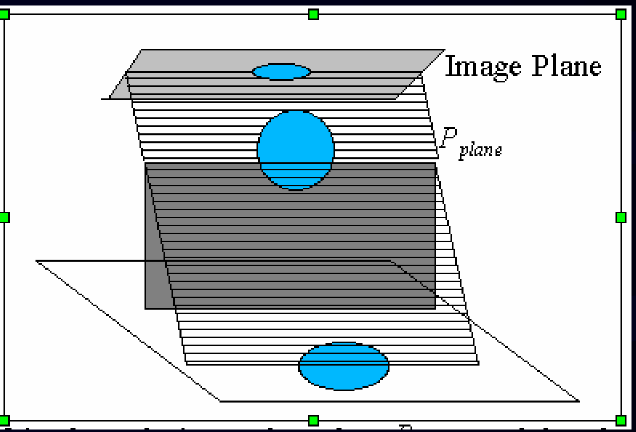


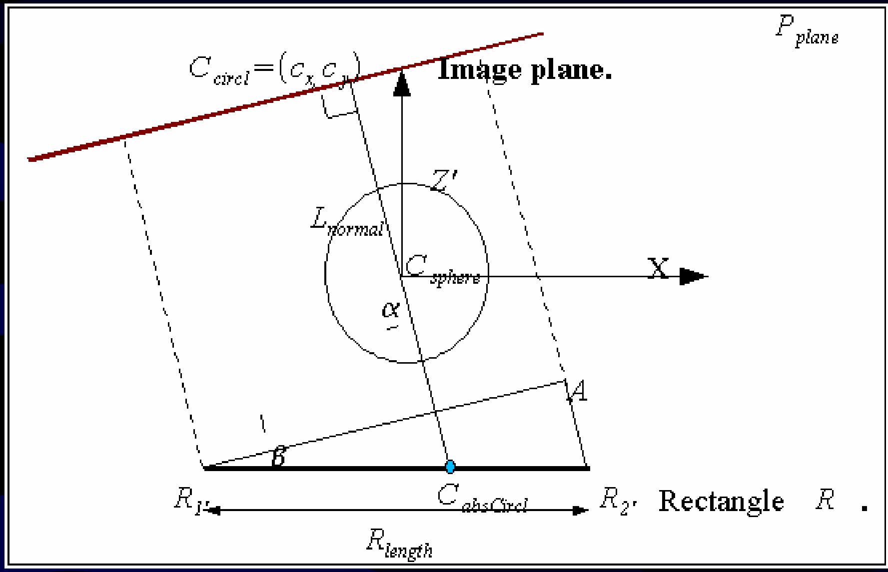
Rectification phase.

- Straightening, Blending and Shifting.



Camera Calibration Method.



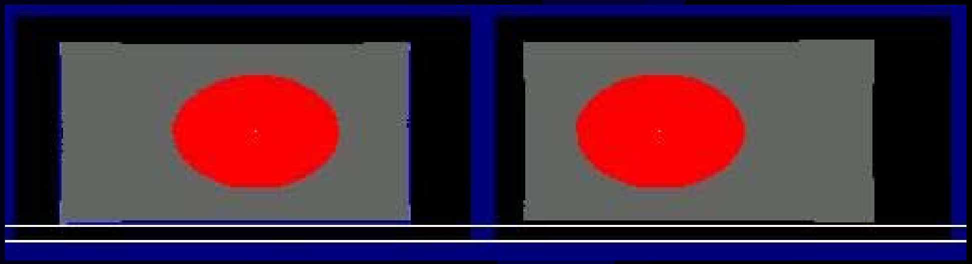


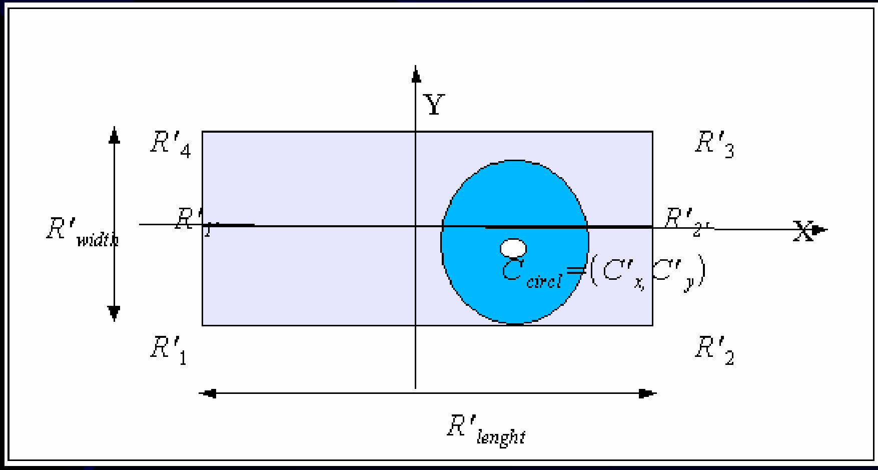
- Intrinsic parameters

$$p_w = \frac{d}{p_u} \quad , \quad p_h = \frac{d}{p_v}$$

- Extrinsic parameters

$$\frac{R'_{lenght}}{R_{lenght}} = \frac{R'_{width}}{R_{width}} = \frac{C'_x}{C_{ax}} = \frac{C'_y}{C_{ay}}$$





State of the the Art of Hand Gesture Recognition

- Hand gesture taxonomy and interaction model
- Hand gesture modeling
- Hand gesture Analysis
- Hand gesture recognition techniques

Taxonomy of Gesture for Human-computer Interaction

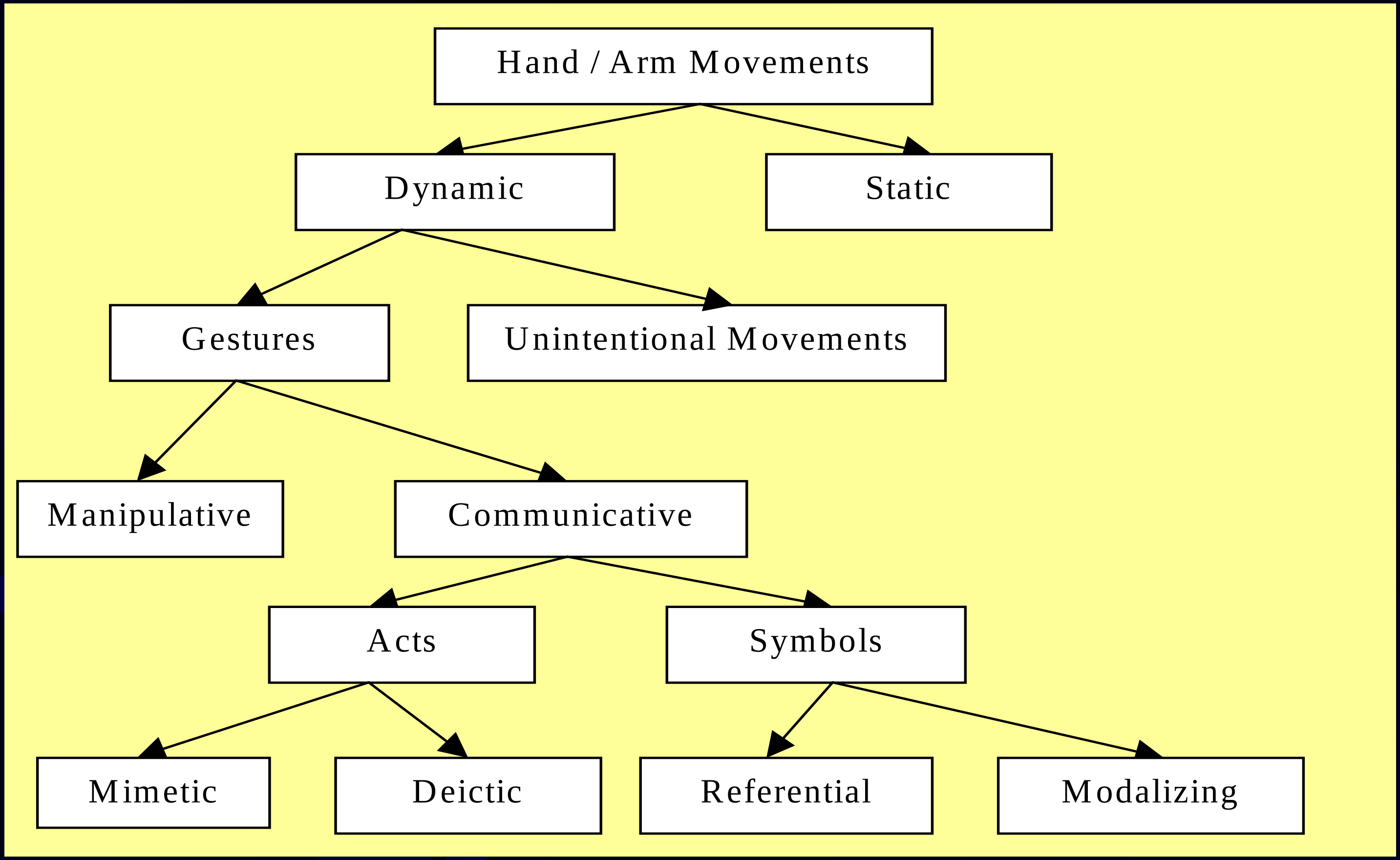


Fig.1: A Taxonomy of hand gestures for Human-computer Interaction. Meaningful gestures are differentiated from unintentional movements. Gestures used for manipulation of objects are separated from the gestures which possess inherent communicational character. Symbols are those gestures having a linguistic role. They symbolize some referential action or are used as modalizers, often of speech.

State of the the Art of Hand Gesture Recognition

- Hand gesture taxonomy and interaction model
- Hand gesture modeling
- Hand gesture Analysis
- Hand gesture recognition techniques

Taxonomy of Gesture for Human-computer Interaction

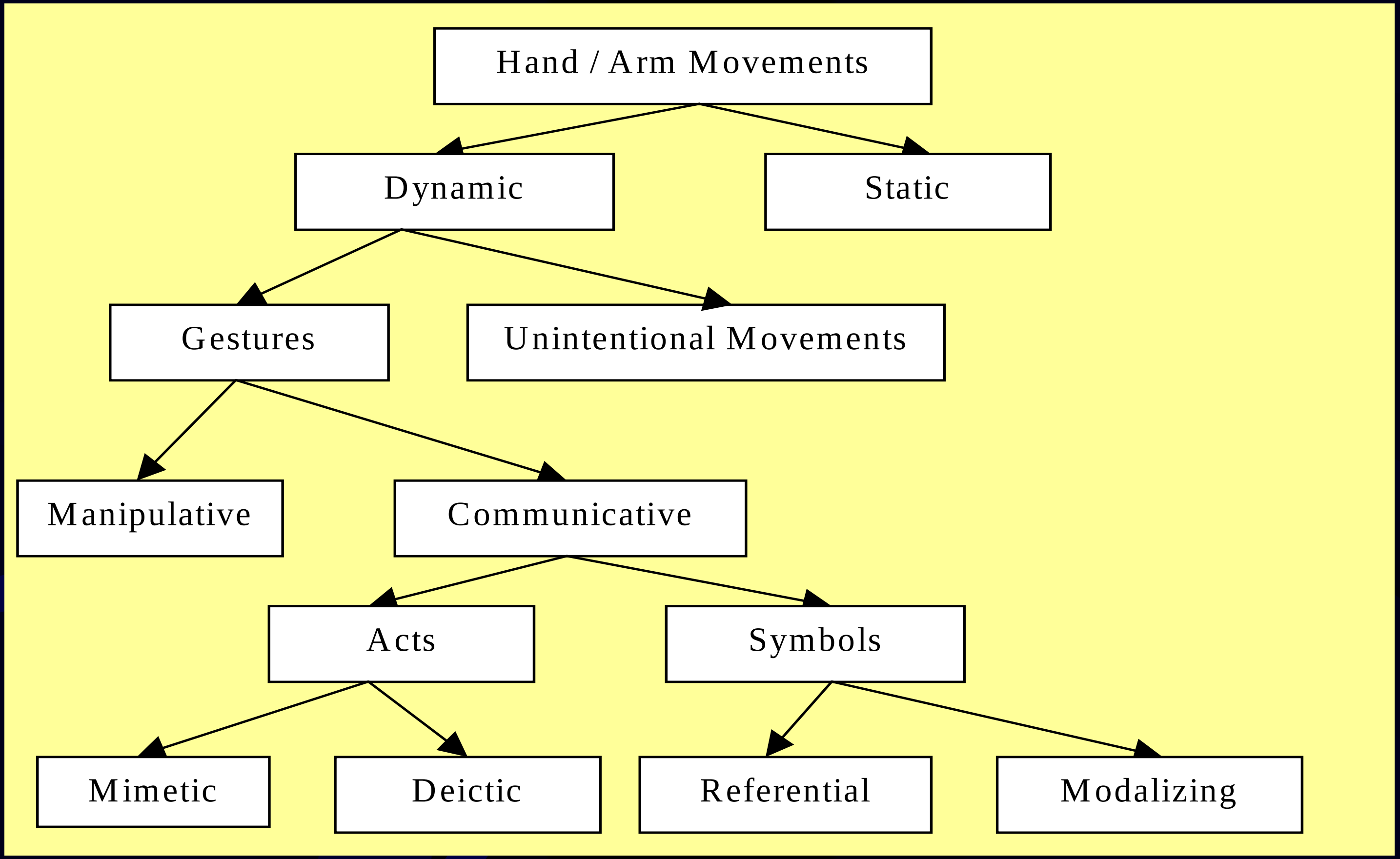
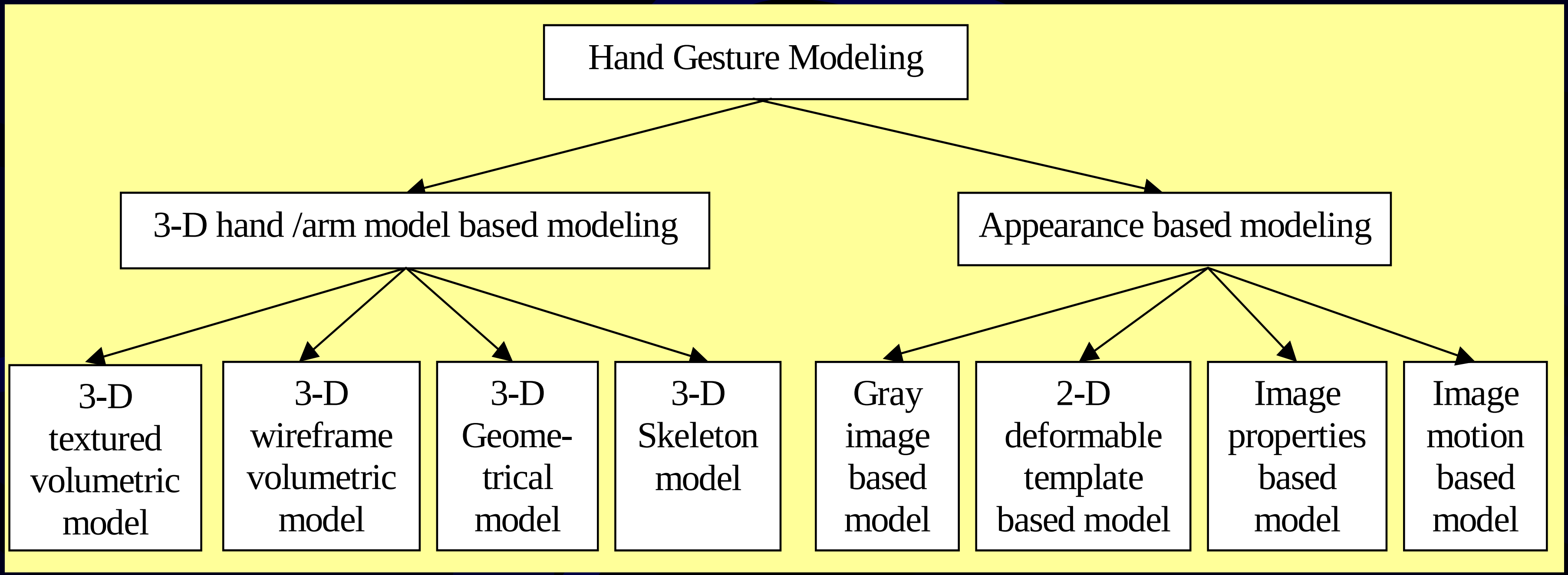


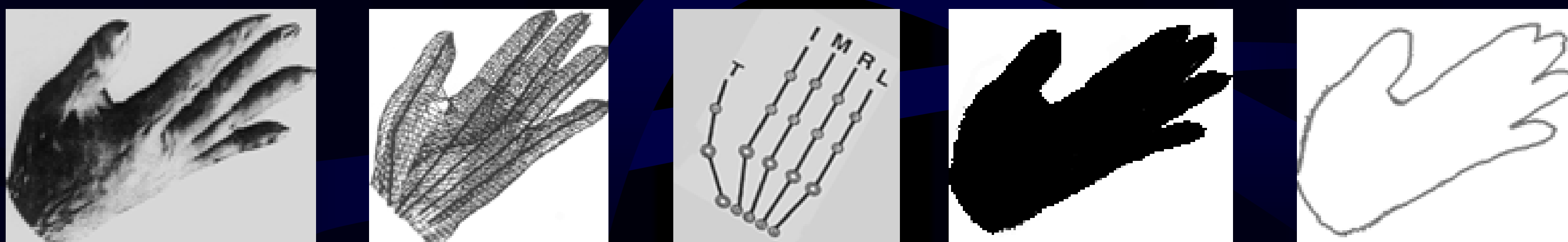
Fig.1: A Taxonomy of hand gestures for Human-computer Interaction. Meaningful gestures are differentiated from unintentional movements. Gestures used for manipulation of objects are separated from the gestures which possess inherent communicational character. Symbols are those gestures having a linguistic role. They symbolize some referential action or are used as modalizers, often of speech.

Hand Gesture Modeling



Classification of hand gesture models

Hand Gesture Modeling



(a) (b) (c) (d) (e)

Fig.3: Representing the same hand posture by different hand models. (a) 3-D textured volumetric model; (b) 3-D wireframe volumetric model; (c) 3-D skeletal model; (d) Binary silhouette; (e) Contour model.

Gesture Analysis

1 Gesture detection and feature extraction

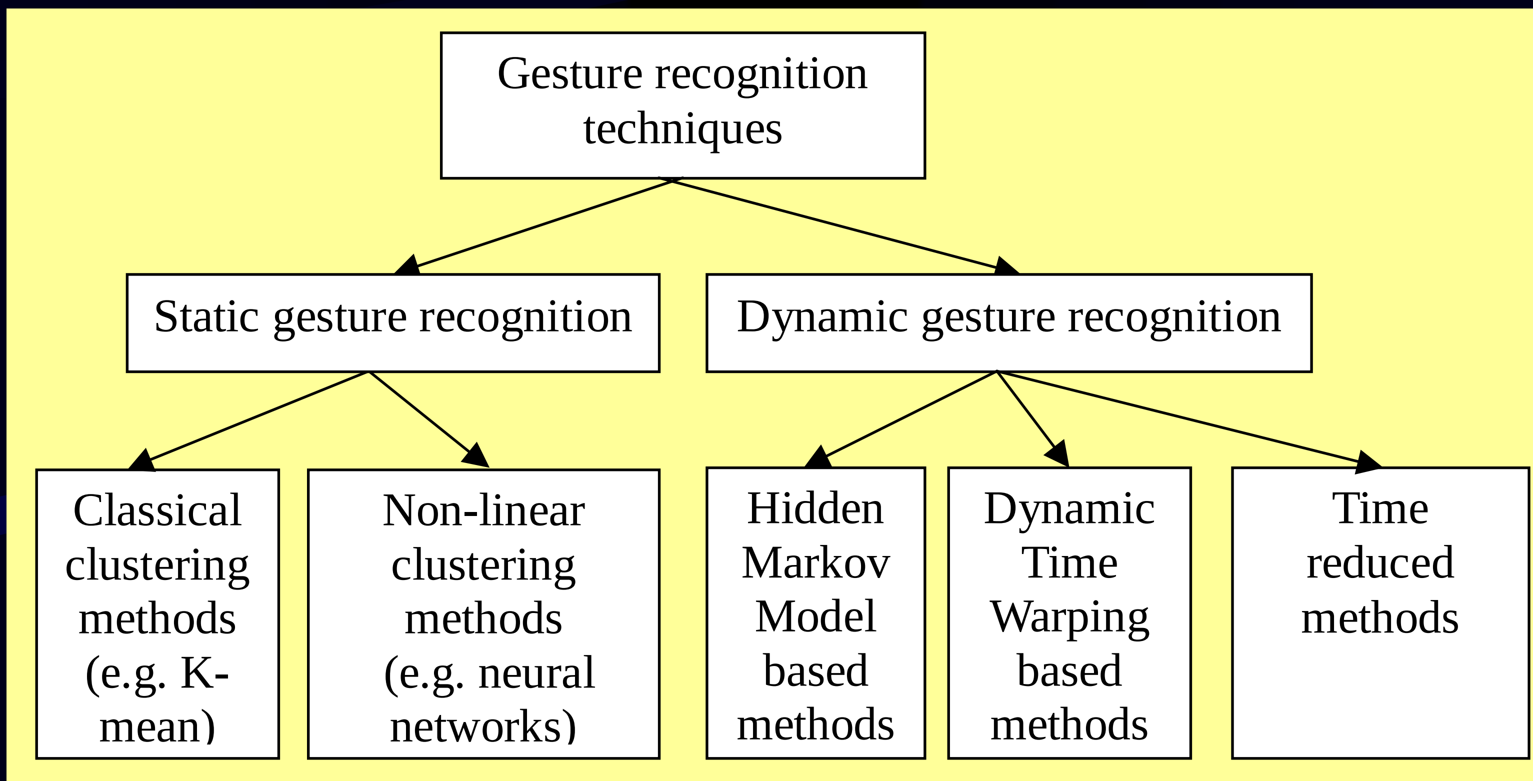
- skin color clues based approaches
- motion clues based approaches
- multiple clues based approaches
- features include gray image, binary silhouette, moving region, edge, contour, and so on.

Gesture Analysis

Recovering gesture model parameters

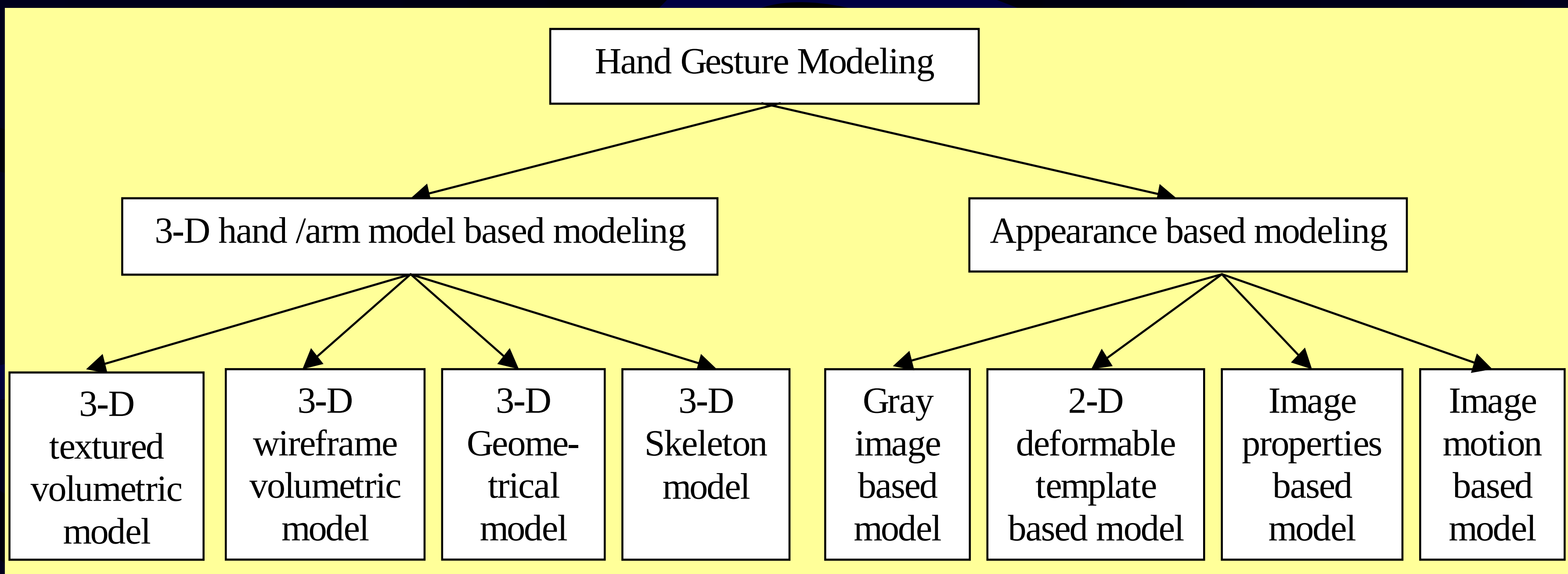
- Estimation of 3-D hand /arm model parameters
  - two sets of parameters: angular (joint angles) and linear (palm dimensions)
  - the initial parameter estimation
  - the parameter update as the hand gesture evolve in time.
- Estimation of appearance based model parameters
  - image motion estimation (e.g. optical flow)
  - shape analysis (e.g. computing moments)
  - histogram based feature parameters (e.g.)
  - active contour model.

Gesture Recognition Techniques



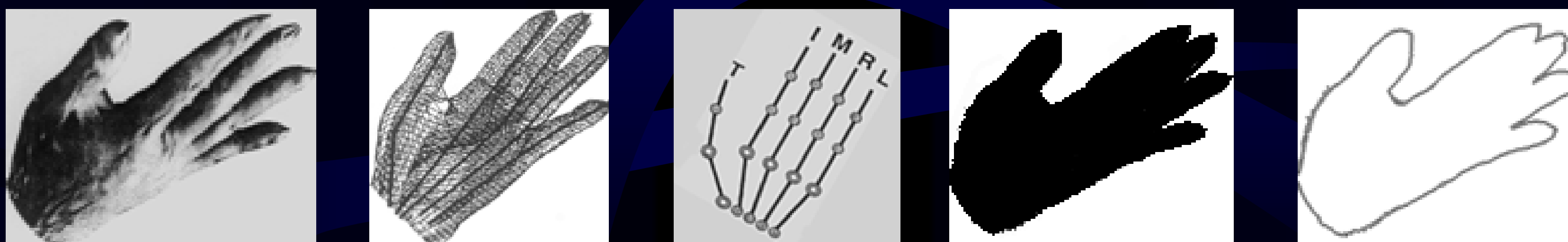
Classification of hand gesture recognition techniques

Hand Gesture Modeling



Classification of hand gesture models

Hand Gesture Modeling



(a) (b) (c) (d) (e)

Fig.3: Representing the same hand posture by different hand models. (a) 3-D textured volumetric model; (b) 3-D wireframe volumetric model; (c) 3-D skeletal model; (d) Binary silhouette; (e) Contour model.

Gesture Analysis

1 Gesture detection and feature extraction

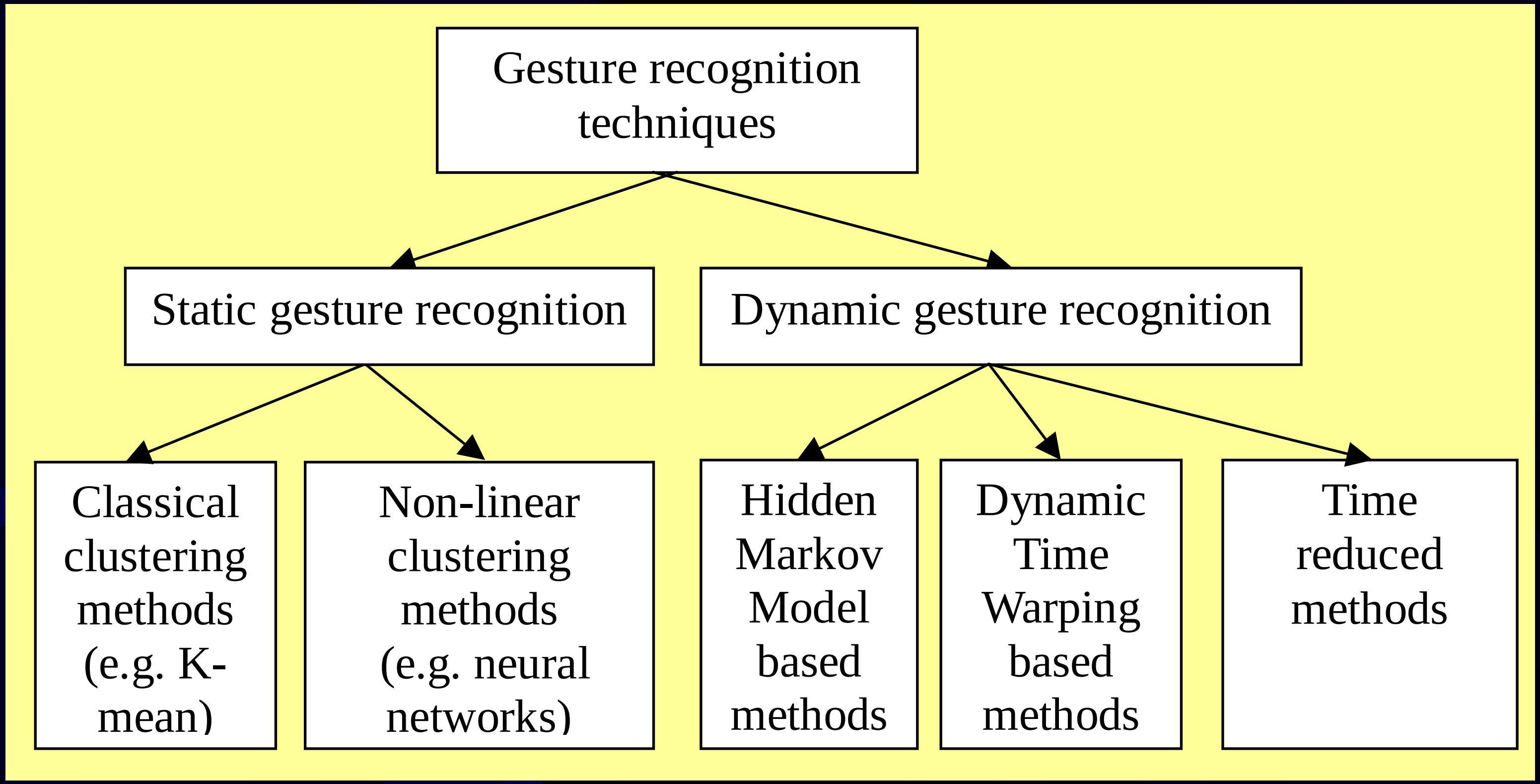
- skin color clues based approaches
- motion clues based approaches
- multiple clues based approaches
- features include gray image, binary silhouette, moving region, edge, contour, and so on.

Gesture Analysis

Recovering gesture model parameters

- Estimation of 3-D hand /arm model parameters
  - two sets of parameters: angular (joint angles) and linear (palm dimensions)
  - the initial parameter estimation
  - the parameter update as the hand gesture evolve in time.
- Estimation of appearance based model parameters
  - image motion estimation (e.g. optical flow)
  - shape analysis (e.g. computing moments)
  - histogram based feature parameters (e.g.)
  - active contour model.

Gesture Recognition Techniques



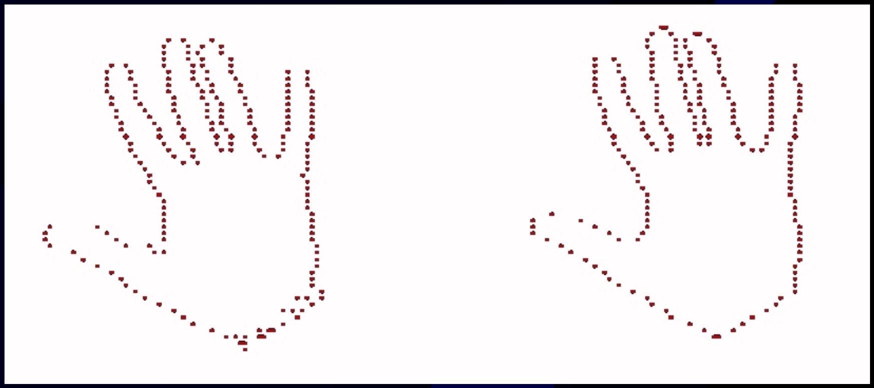
Classification of hand gesture recognition techniques

Stereo-Reconstruction.

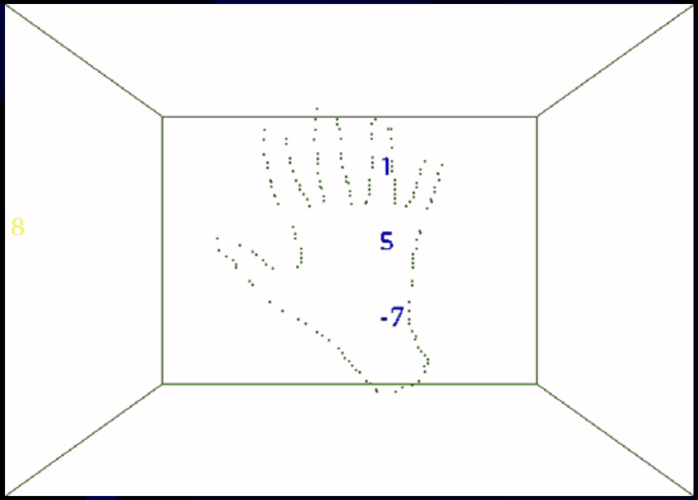
- Simple matching.

- Thresholding.

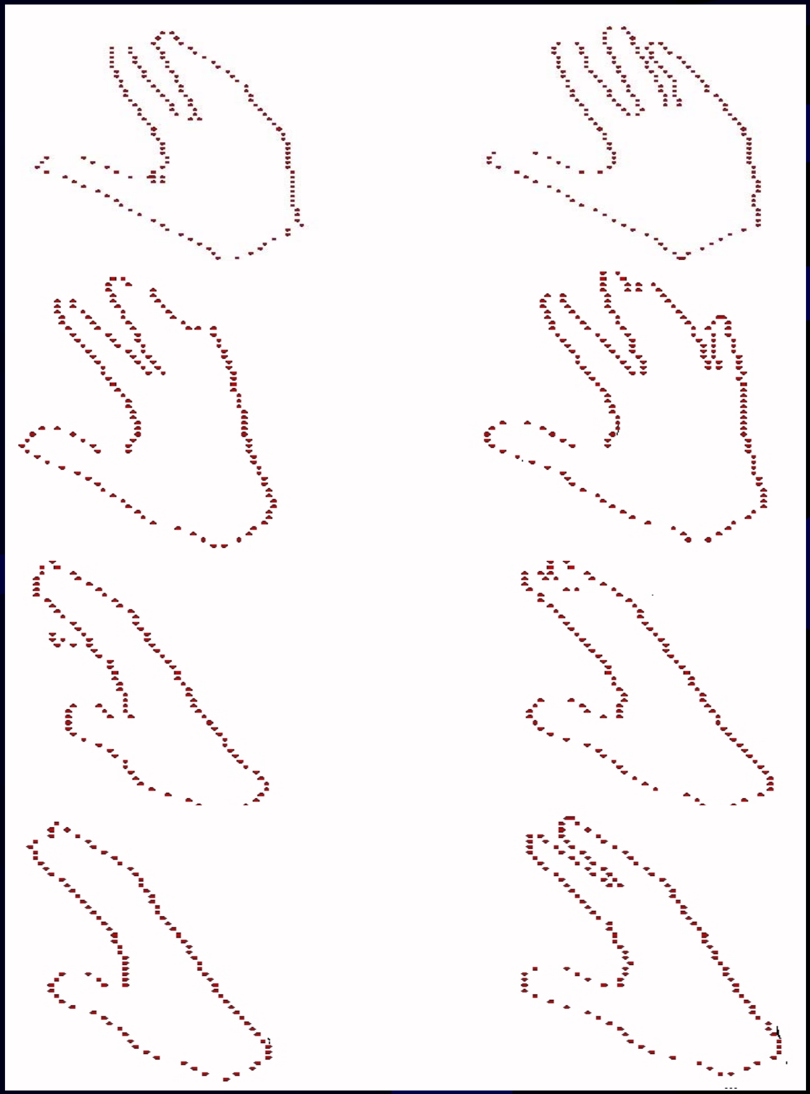
- Fast reconstruction.

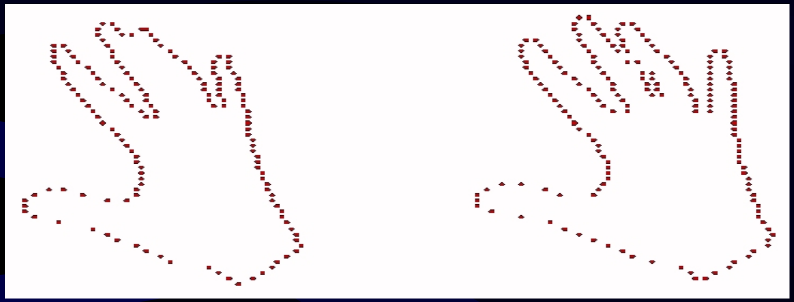


- 3D reconstruction.

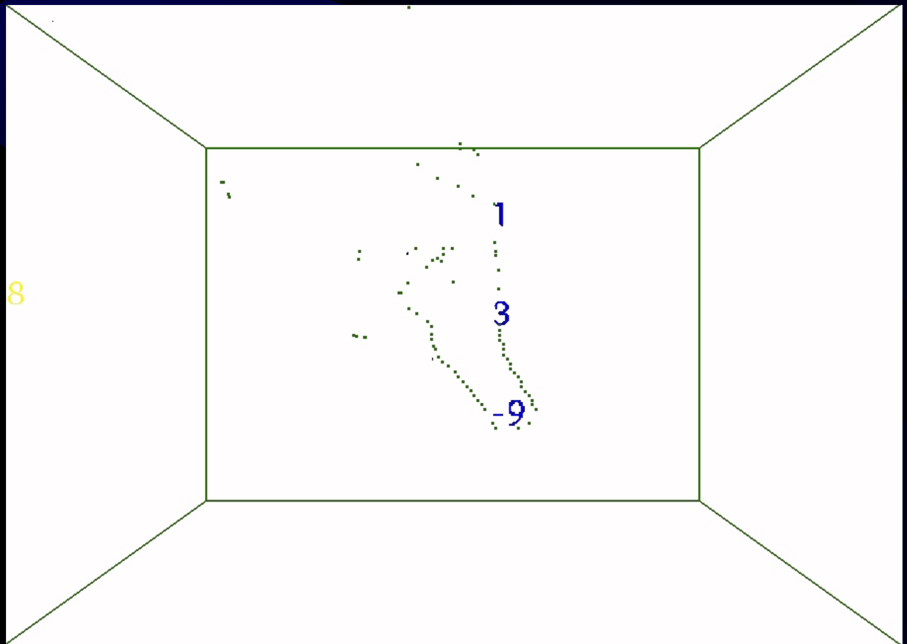


Problems.





- Order constraint, occlusion, merging.

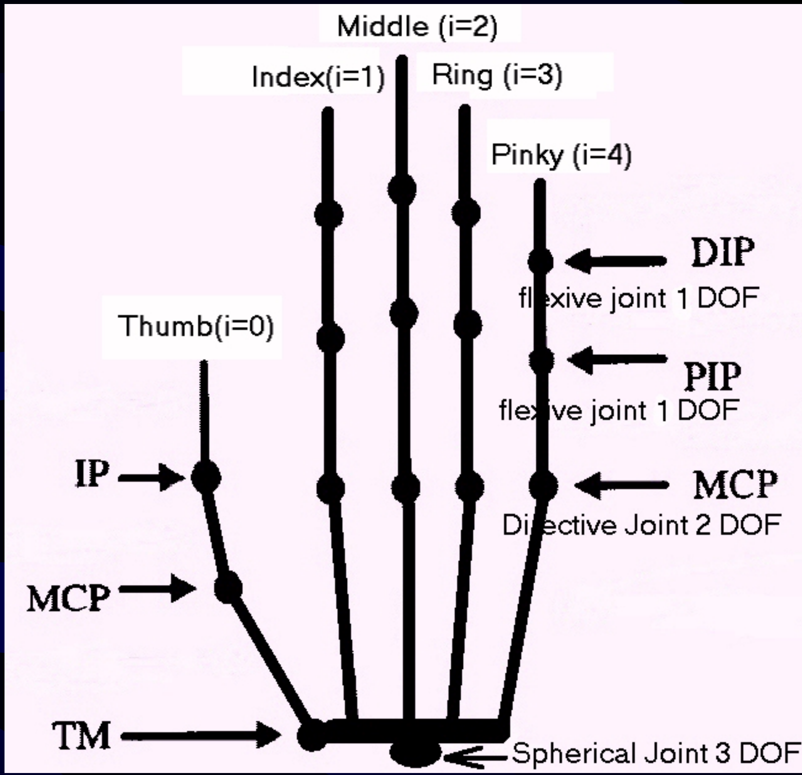


Hand Modeling.

- Dynamic Constraints for all four fingers.

$$\vartheta_{DIP.fe}(i)=\frac{2}{3}\vartheta_{PIP.fe}(i)$$

$$\vartheta_{MCP.aa}=\frac{\vartheta_{MPC.fe}}{90}\left(\vartheta_{MPC,converge}-\vartheta_{MCP.aa,s}\right)+\vartheta_{MCP.aa,s}$$



- Static Constraints for all four fingers.

$$0\le\vartheta_{DIP.fe}(i)\le s_{max}\left(\vartheta_{DIP.fe}(i)\right)\quad with\quad s_{max}\left(\vartheta_{DIP.fe}(i)\right)=90$$

$$0\le\vartheta_{PIP.fe}(i)\le s_{max}\left(\vartheta_{PIP.fe}(i)\right)\quad with\quad s_{max}\left(\vartheta_{PIP.fe}(i)\right)=110$$

$$0\le\vartheta_{MCP.fe}(i)\le s_{max}\left(\vartheta_{MCP.fe}(i)\right)\quad with\quad s_{max}\left(\vartheta_{MCP.fe}(i)\right)=90$$

$$-1\le\vartheta_{MCP.aa,o}(2)\le 1$$

$$-15\le\vartheta_{MCP.aa,o}(1,3,4)\le 15$$

- Kush, Wu.
- Agee 1982.

Dynamic Constraints.

- For separate fingers.

Index finger (i=1):

$$d_{max}(\vartheta_{MCP.fe}(1)) = min(\vartheta_{MCP.fe}(2)+25 \quad , \quad s_{max}(\vartheta_{MCP.fe}(1)))$$
$$d_{min}(\vartheta_{MCP.fe}(1)) = max(\vartheta_{MCP.fe}(2)-54 \quad , \quad 0)$$

Middle finger (i=2):

$$d_{max}(\vartheta_{MCP.fe}(2)) = min(\vartheta_{MCP.fe}(1)+54, \vartheta_{MCP.fe} \quad , \quad \vartheta_{MCP.fe}(3)+20 \quad , \quad s_{max}(\vartheta_{MCP.fe}(2)))$$
$$d_{min}(\vartheta_{MCP.fe}(2)) = max(\vartheta_{MCP.fe}(1)-25 \quad , \quad \vartheta_{MCP.fe}(3)-45 \quad , \quad 0)$$

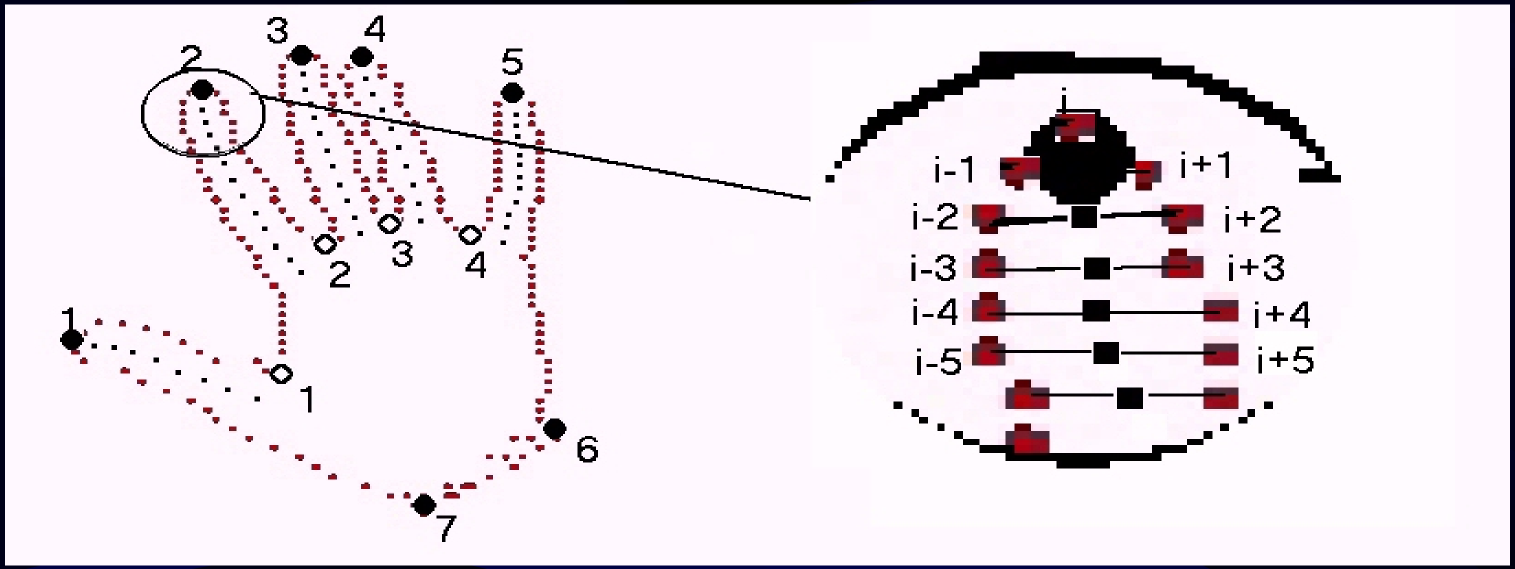
Ring finger (i=3):

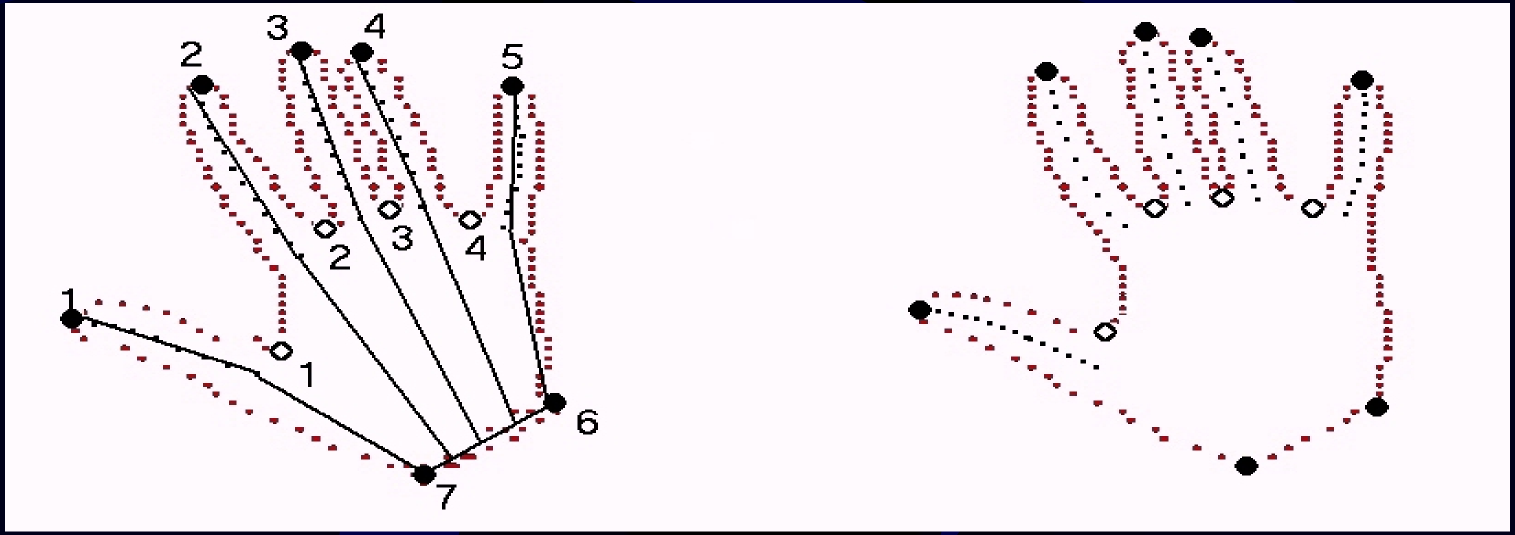
$$d_{max}(\vartheta_{MCP.fe}(3)) = min(\vartheta_{MCP.fe}(2)+45 \quad , \quad \vartheta_{MCP.fe}(4)+48, s_{max}(\vartheta_{MCP.fe}(3)))$$
$$d_{min}(\vartheta_{MCP.fe}(3)) = max(\vartheta_{MCP.fe}(2)-20 \quad , \quad \vartheta_{MCP.fe}(4)-44 \quad , \quad 0)$$

Pinky finger (i=4):

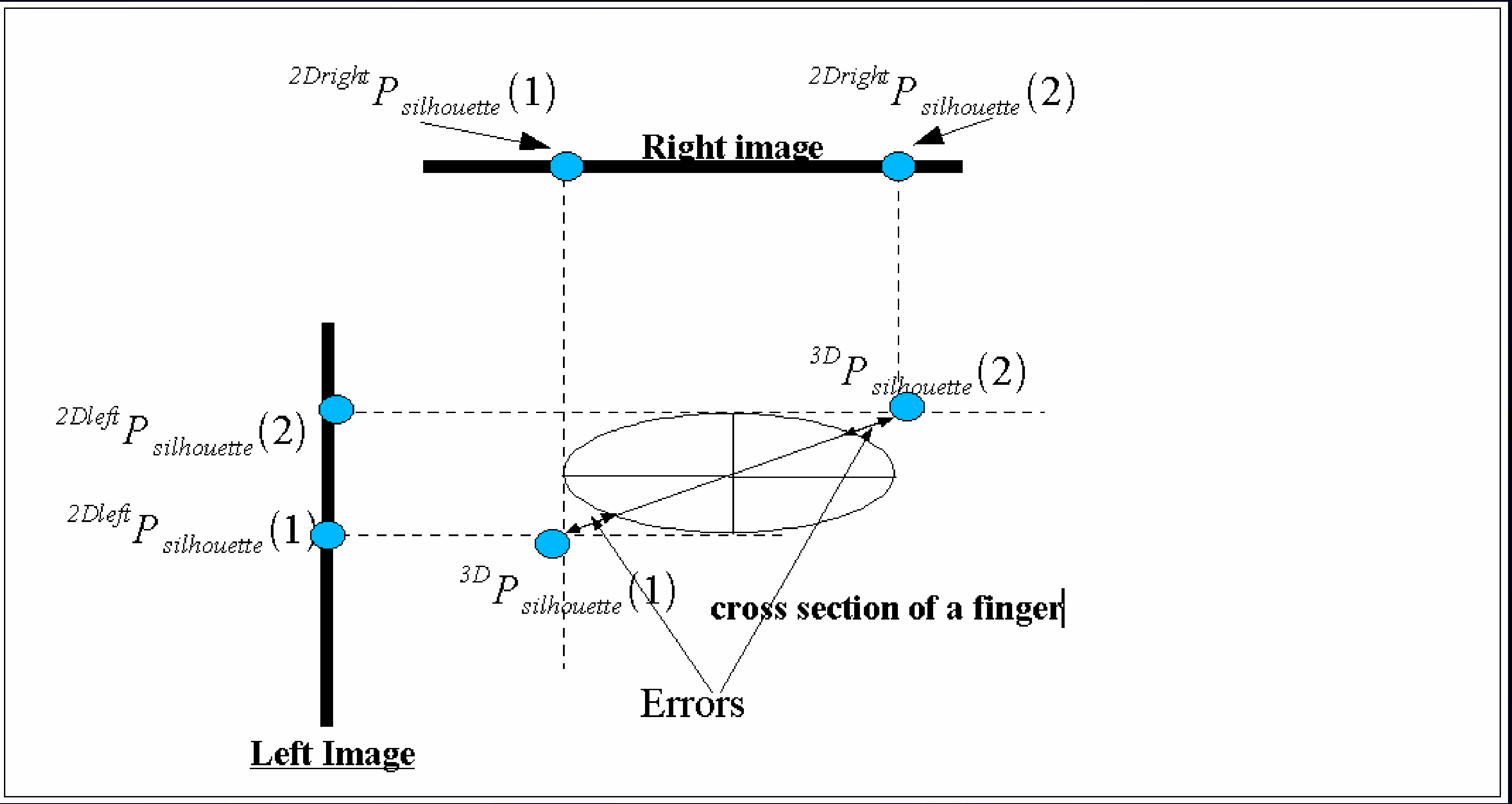
$$d_{max}(\vartheta_{MCP.fe}(4)) = min(\vartheta_{MCP.fe}(3)+44 \quad , \quad s_{max}(\vartheta_{MCP.fe}(4)))$$
$$d_{min}(\vartheta_{MCP.fe}(4)) = max(\vartheta_{MCP.fe}(3)-48 \quad , \quad 0)$$

Initial Pose of the Hand Model.



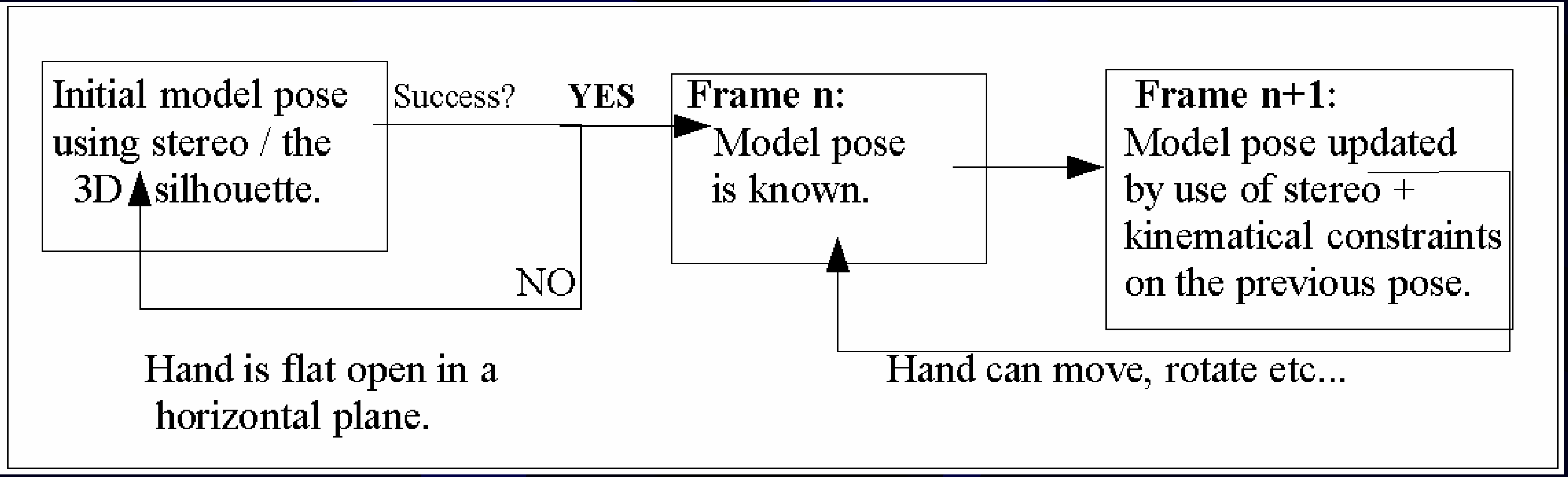


Precision of the Initial Pose.



Tracking the Hand.

- General Diagram:
  - Initial pose,
  - Real time tracking.



Linear Optimization.

- Frame N-1: Feature vector:

$$Hand_{pose}(N-1) = (\bar{\vartheta}_{DIP.fe}(i) , \bar{\vartheta}_{PIP.fe}(i) , \bar{\vartheta}_{MCP.fe}(i) , \bar{\vartheta}_{MCP.aa}(i) , \bar{\vartheta}_{IP.fe} , \bar{\vartheta}_{MCP.fe} \quad \bar{\vartheta}_{TM.fe} , \bar{\vartheta}_{TM.aa} , {}^{palm}\bar{B}.x , {}^{palm}\bar{B}.y , {}^{palm}\bar{B}.z , {}^{palm}\bar{B}.\vartheta , {}^{palm}\bar{B}.\alpha , {}^{palm}\bar{B}.\gamma)$$

- Frame N: Feature vector:

$$Hand_{pose}(N) = (\vartheta_{DIP.fe}(i) , \vartheta_{PIP.fe}(i) , \vartheta_{MCP.fe}(i) , \vartheta_{MCP.aa}(i) , \vartheta_{IP.fe} , \vartheta_{MCP.fe} \quad \vartheta_{TM.fe} , \vartheta_{TM.aa} , {}^{palm}B.x , {}^{palm}B.y , {}^{palm}B.z , {}^{palm}B.\vartheta , {}^{palm}B.\alpha , {}^{palm}B.\gamma).$$

- Minimization of:

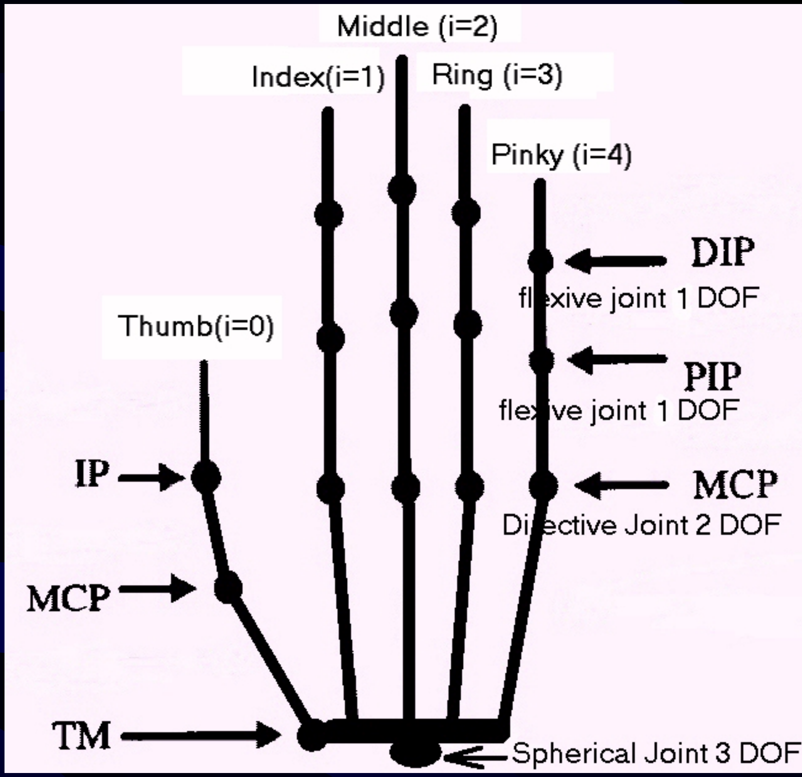
$$z = \Sigma_i [\bar{\vartheta}_{DIP.fe}(i) - \vartheta_{DIP.fe}(i)] + \Sigma [\bar{\vartheta}_{PIP.fe}(i) - \vartheta_{PIP.fe}(i)] + \Sigma [\bar{\vartheta}_{MCP.fe}(i) - \vartheta_{MCP.fe}(i)]$$
$$+ \Sigma [\bar{\vartheta}_{MCP.aa}(i) - \vartheta_{MCP.aa}(i)] + \bar{\vartheta}_{IP.fe} - \vartheta_{IP.fe} + \bar{\vartheta}_{MCP.fe} - \vartheta_{MCP.fe} + \bar{\vartheta}_{TM.fe} - \vartheta_{TM.fe} + \bar{\vartheta}_{TM.aa} - \vartheta_{TM.aa}$$
$$+ {}^{palm}\bar{B}.x - {}^{palm}B.x + {}^{palm}\bar{B}.y - {}^{palm}B.y + {}^{palm}\bar{B}.z - {}^{palm}B.z + {}^{palm}\bar{B}.\vartheta - {}^{palm}B.\vartheta$$
$$+ {}^{palm}\bar{B}.\alpha - {}^{palm}B.\alpha + {}^{palm}\bar{B}.\gamma - {}^{palm}B.\gamma$$

Hand Modeling.

- Dynamic Constraints.

$$\vartheta_{DIP.fe}(i)=\frac{2}{3}\vartheta_{PIP.fe}(i)$$

$$\vartheta_{MCP.aa}=\frac{\vartheta_{MPC.fe}}{90}\left(\vartheta_{MPC,converge}-\vartheta_{MCP.aa,s}\right)+\vartheta_{MCP.aa,s}$$



- Static Constraints.

$$0\le\vartheta_{DIP.fe}(i)\le s_{max}\left(\vartheta_{DIP.fe}(i)\right)\quad with\quad s_{max}\left(\vartheta_{DIP.fe}(i)\right)=90$$

$$0\le\vartheta_{PIP.fe}(i)\le s_{max}\left(\vartheta_{PIP.fe}(i)\right)\quad with\quad s_{max}\left(\vartheta_{PIP.fe}(i)\right)=110$$

$$0\le\vartheta_{MCP.fe}(i)\le s_{max}\left(\vartheta_{MCP.fe}(i)\right)\quad with\quad s_{max}\left(\vartheta_{MCP.fe}(i)\right)=90$$

$$-1\le\vartheta_{MCP.aa,o}(2)\le 1$$

$$-15\le\vartheta_{MCP.aa,o}(1,3,4)\le 15$$

Dynamic Constraints.

Index finger (i=1):

$$d_{max}(\vartheta_{MCP.fe}(1)) = min(\vartheta_{MCP.fe}(2)+25 \quad , \quad s_{max}(\vartheta_{MCP.fe}(1)))$$

$$d_{min}(\vartheta_{MCP.fe}(1)) = max(\vartheta_{MCP.fe}(2)-54 \quad , \quad 0)$$

Middle finger (i=2):

$$d_{max}(\vartheta_{MCP.fe}(2)) = min(\vartheta_{MCP.fe}(1)+54, \vartheta_{MCP.fe} \quad , \quad \vartheta_{MCP.fe}(3)+20 \quad , \quad s_{max}(\vartheta_{MCP.fe}(2)))$$

$$d_{min}(\vartheta_{MCP.fe}(2)) = max(\vartheta_{MCP.fe}(1)-25 \quad , \quad \vartheta_{MCP.fe}(3)-45 \quad , \quad 0)$$

Ring finger (i=3):

$$d_{max}(\vartheta_{MCP.fe}(3)) = min(\vartheta_{MCP.fe}(2)+45 \quad , \quad \vartheta_{MCP.fe}(4)+48, s_{max}(\vartheta_{MCP.fe}(3)))$$

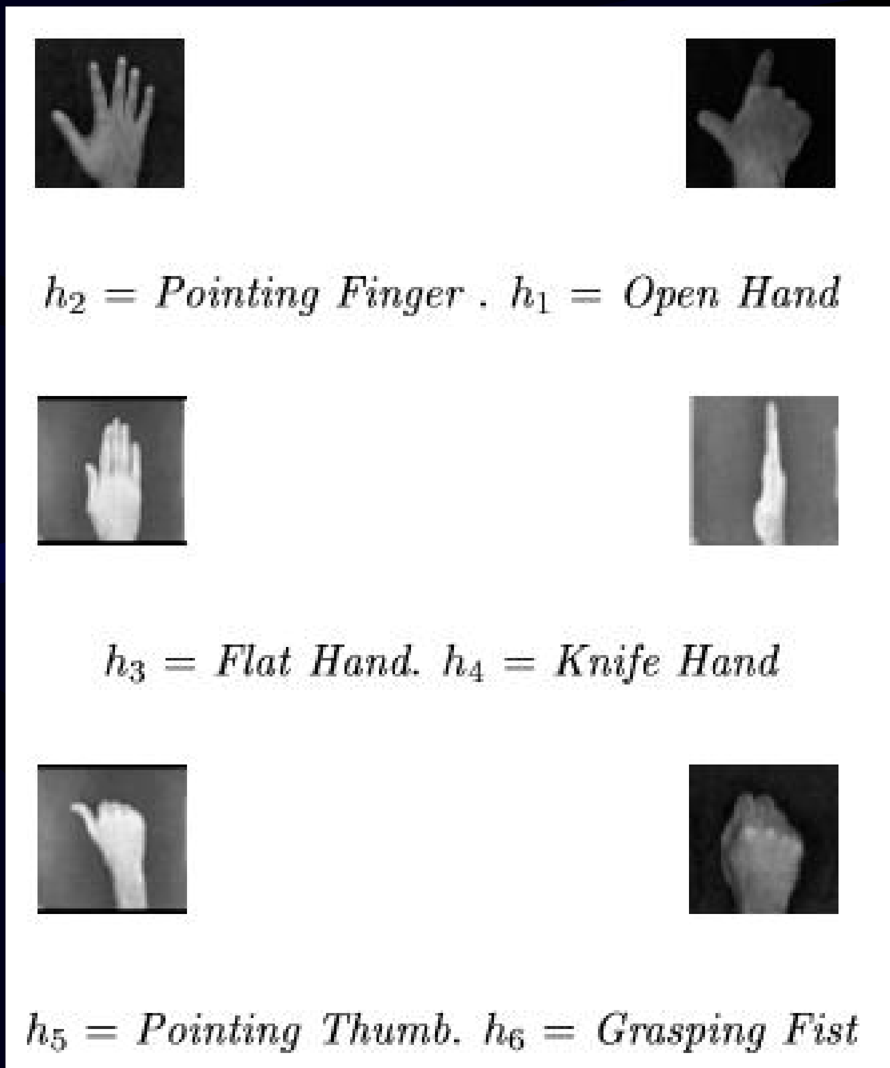
$$d_{min}(\vartheta_{MCP.fe}(3)) = max(\vartheta_{MCP.fe}(2)-20 \quad , \quad \vartheta_{MCP.fe}(4)-44 \quad , \quad 0)$$

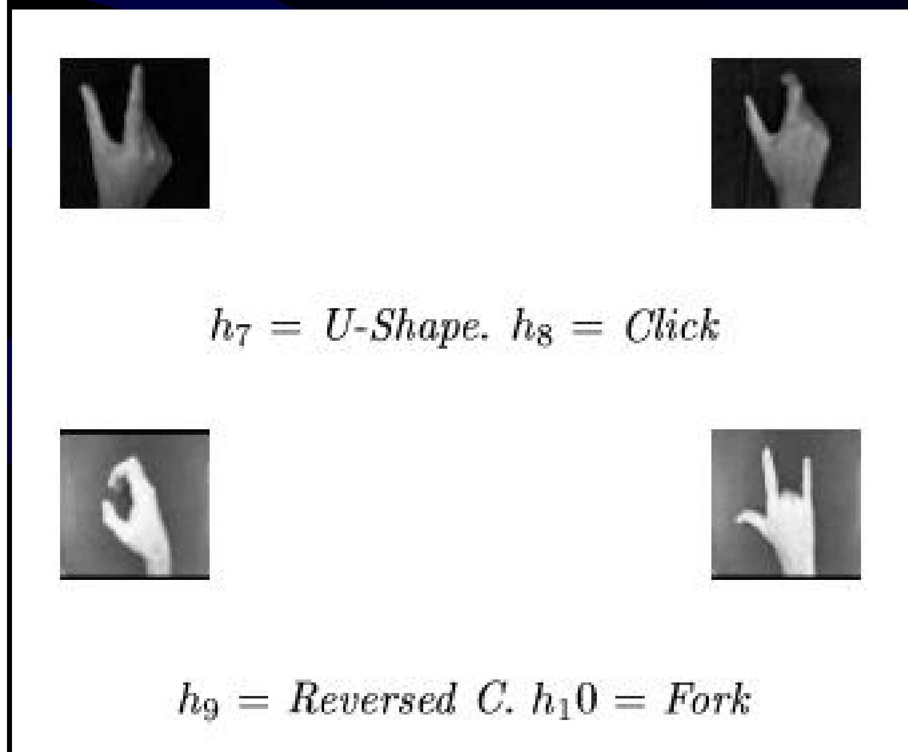
Pinky finger (i=4):

$$d_{max}(\vartheta_{MCP.fe}(4)) = min(\vartheta_{MCP.fe}(3)+44 \quad , \quad s_{max}(\vartheta_{MCP.fe}(4)))$$

$$d_{min}(\vartheta_{MCP.fe}(4)) = max(\vartheta_{MCP.fe}(3)-48 \quad , \quad 0)$$

SVM gesture recognizer.

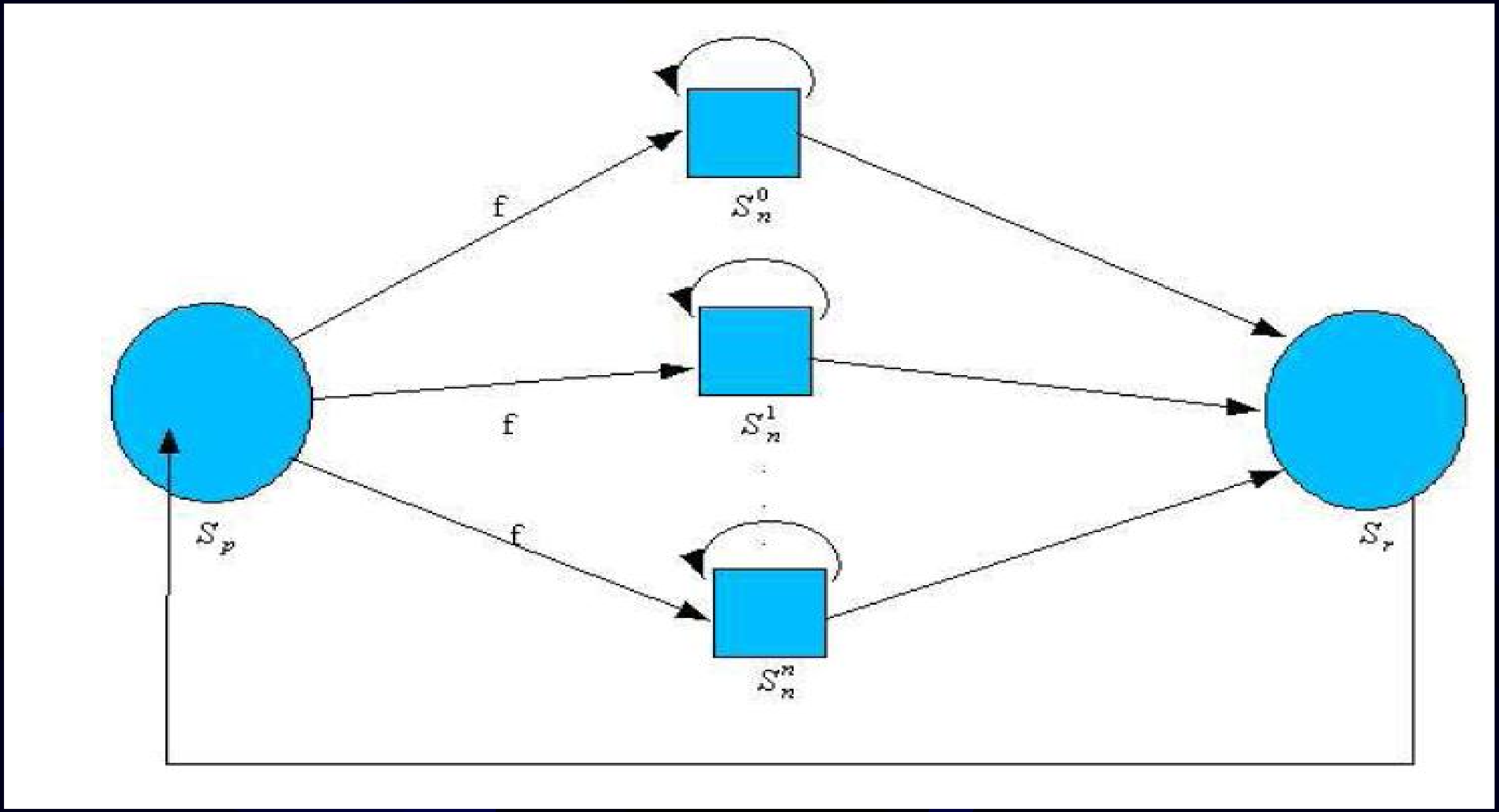




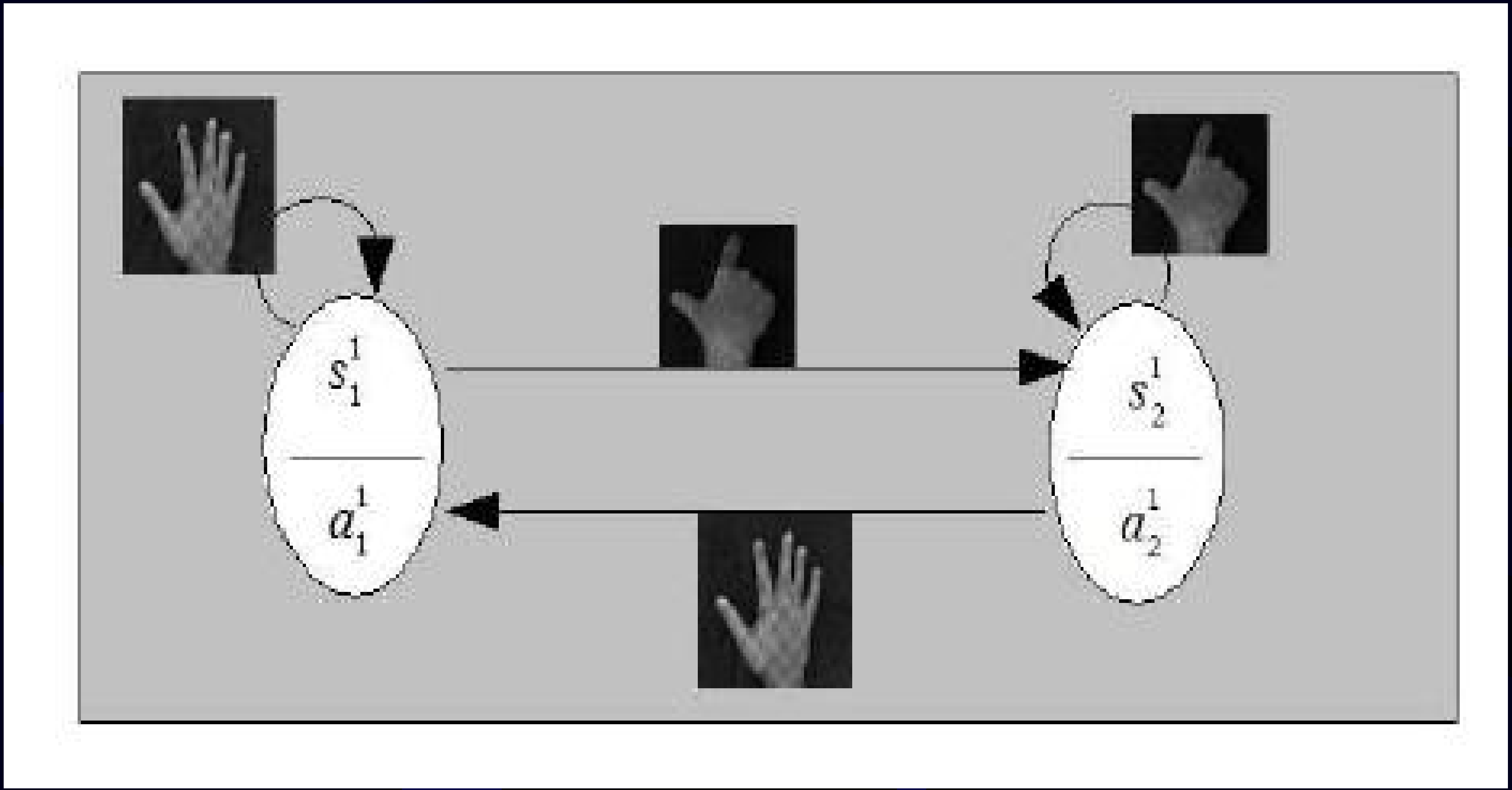
Gestural phases: Kendon.

- 1- Preparation phase: *prepares the hand from its idle state, by moving into a recognizable form.*
- 2- The Nucleus phase: *which has a definite form and is the peak or stroke of the gesture*
- 3- The retraction phase: *which usually returns the hand to the resting position.*

Super-State Machine.



Mini-State Machine S1.

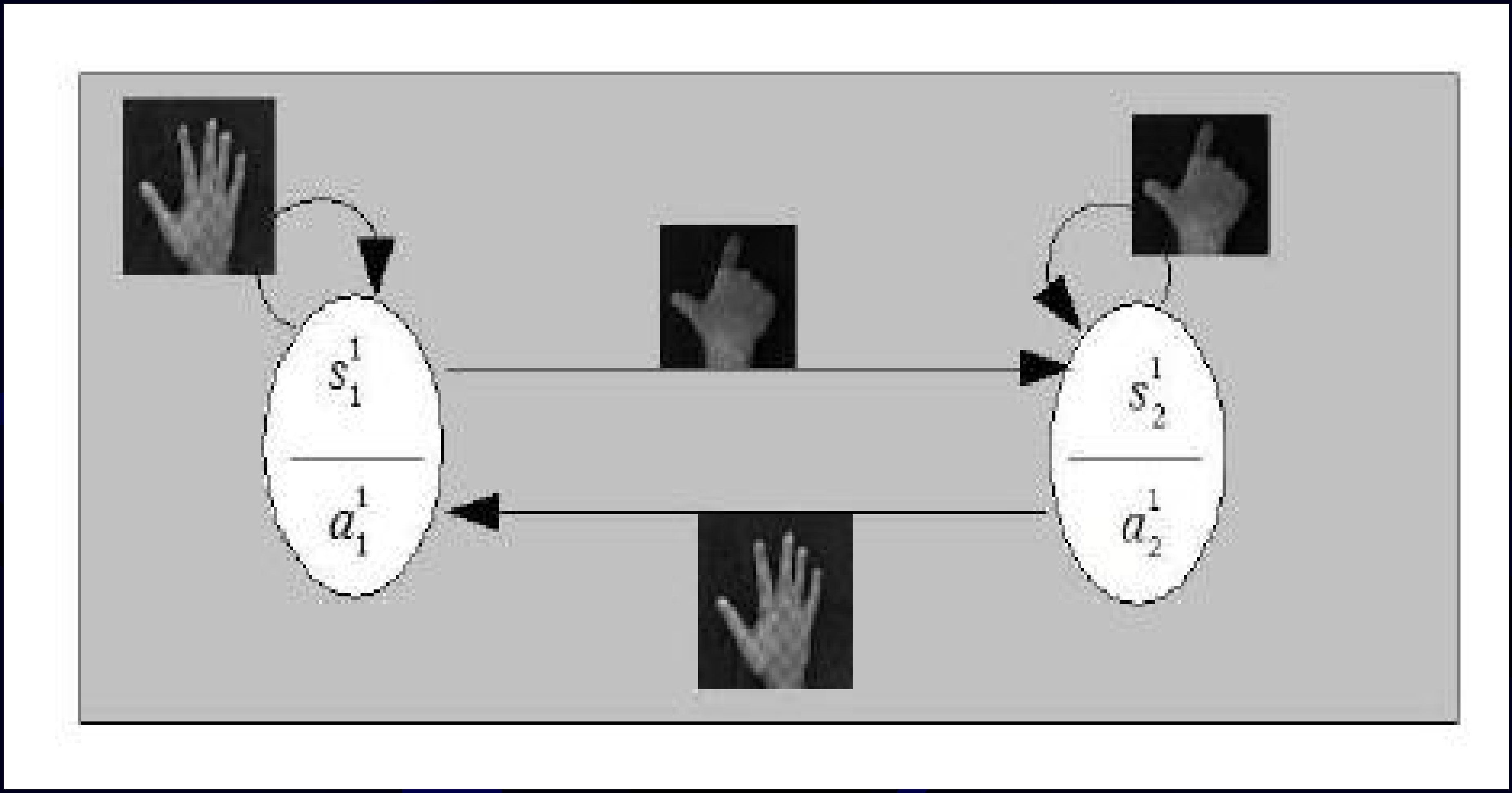


States and Input events.

- $S^1 = \{s_1^1, s_2^1\}$ with s_1^1 is the moving state, and s_2^1 is the rotation or looking around state.

- I^1 is the input event set $I^1 = \{h_1, h_2\} \subset H$

Mini-State Machine S1.

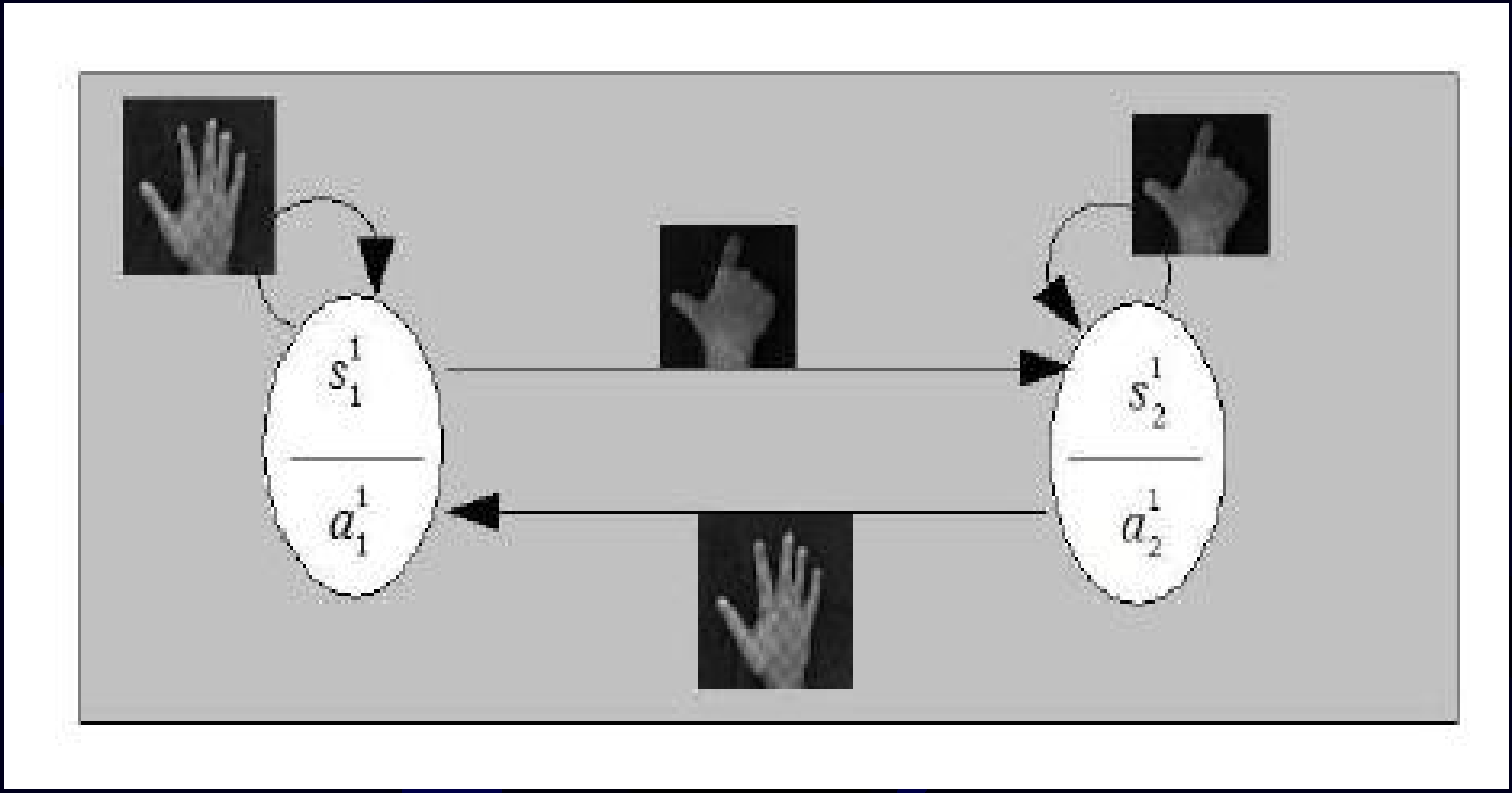


Actions

- $a_1^1 = a^1_{openhand} = g_1^1(p_1^1)$ is an action performed by the *open hand*, the action being translating the view point of the camera along the x, y, and z coordinates,

- $a_2^1 = a^1_{pointingfinger} = g_2^1(p_2^1)$ is an action performed by the pointing finger, the action being rotating the view point of the camera in pitch, yaw and roll.

Mini-State Machine S1.

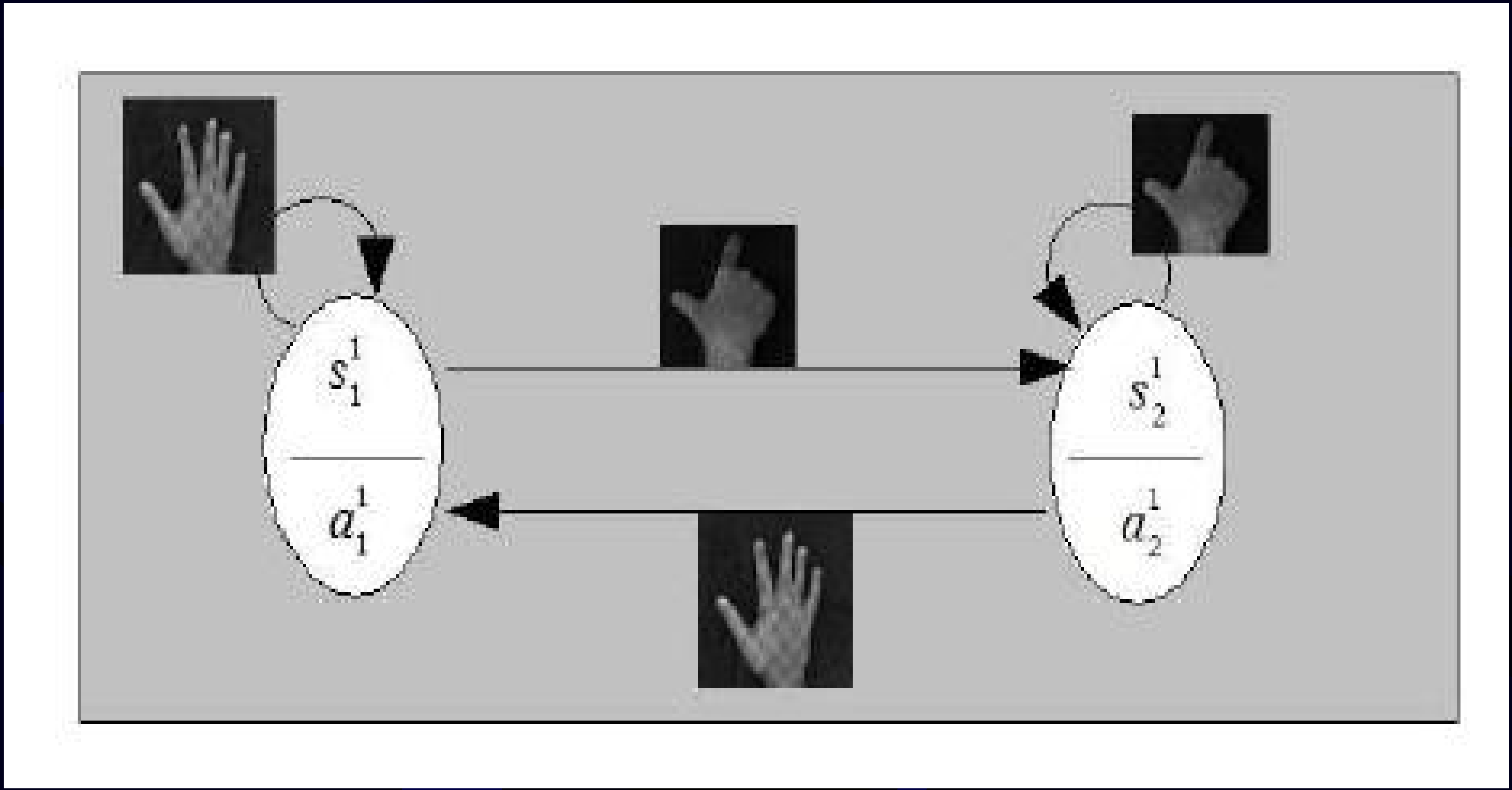


Parameters.

- $p_1^1 \in P$ with $p_1^1 = p_{openhand}^1 = (x, y, z)$ is the center of gravity of the static hand sign h_1 (*i.e. open hand*), in the absolute coordinate system and

- $p_2^1 = p_{pointingfinger}^1 = (\alpha, \beta, \gamma)$ be the direction of the pointing finger in static hand sign h_2 (*i.e. pointing finger*)

Mini-State Machine S1.

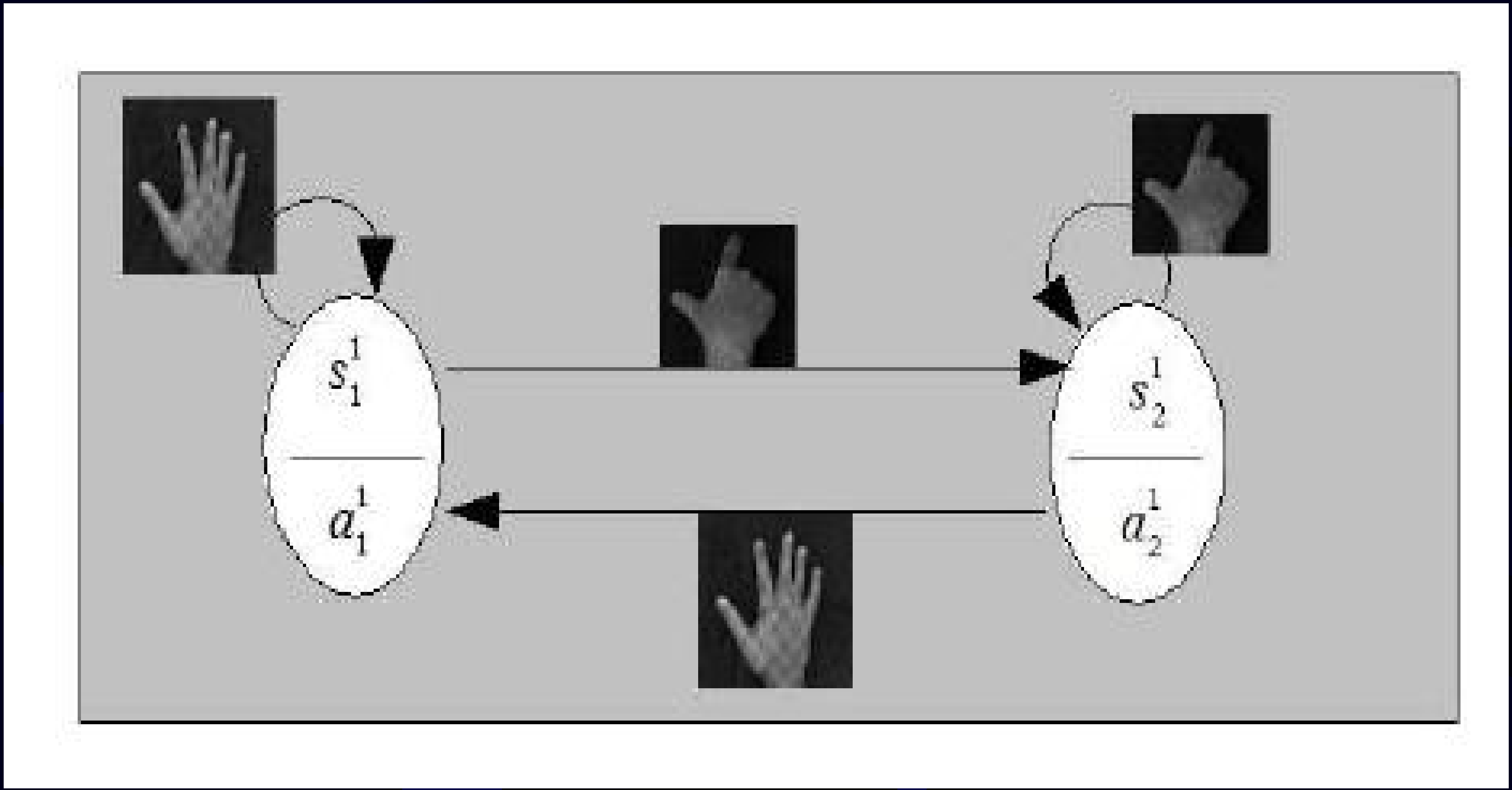


Functions.

- $g_1^1 \in G$ with $g_1^1 = g^1_{openhand} : p^1_{openhand} \rightarrow a^1_{openhand}$ in other words we can write: $g_1^1(p_1^1) = g_1^1(x, y, z) = \sqrt{(x^2 + y^2 + z^2)}$ = the velocity of motion of the virtual camera in the (x,y,z) direction = $a^1_{openhand}$.

- $g_2^1 = g^1_{pointingfinger} : p^1_{pointingfinger} \rightarrow a^1_{pointingfinger}$ in other words we can write: $g_2^1(p_2^1) = g_2^1(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos\alpha & sin\alpha \\ 0 & -sin\alpha & cos\alpha \end{pmatrix} \times \begin{pmatrix} cos\beta & 0 & -sin\beta \\ 0 & 1 & 0 \\ sin\beta & 0 & cos\beta \end{pmatrix} \times$ $\begin{pmatrix} cos\gamma & sin\gamma & 0 \\ -sin\gamma & cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = a_2^1$ which is the action of rotation of the camera in yaw, roll and pitch.

Mini-State Machine S1.



Compensatory. Pursuit.