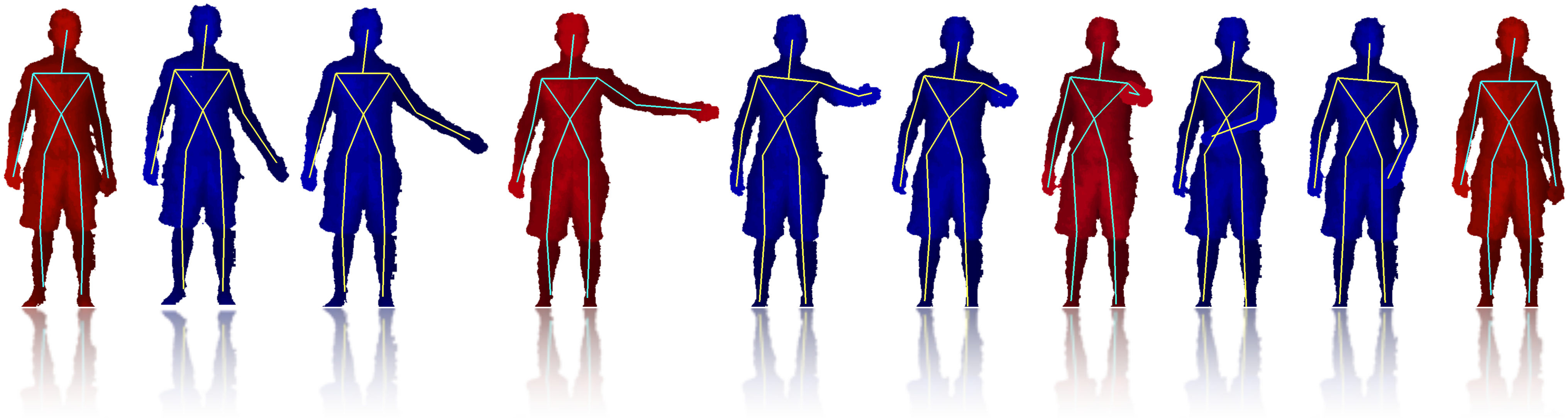
Real-time gesture recognition from depth data through key poses learning and decision forests

Leandro Miranda
Thales Vieira (presenter)
Dimas Martinez
Mathematics, UFAL

Thomas Lewiner
Mathematics, PUC-Rio

Antonio W. Vieira
Mario F. M. Campos
Computer Science, UFMG



Human Gesture Recognition

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Human Gesture Recognition

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Human Gesture Recognition

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Human Gesture Recognition

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Current Scenario

Popularization of real time depth sensors

Microsoft Kinect Sensor

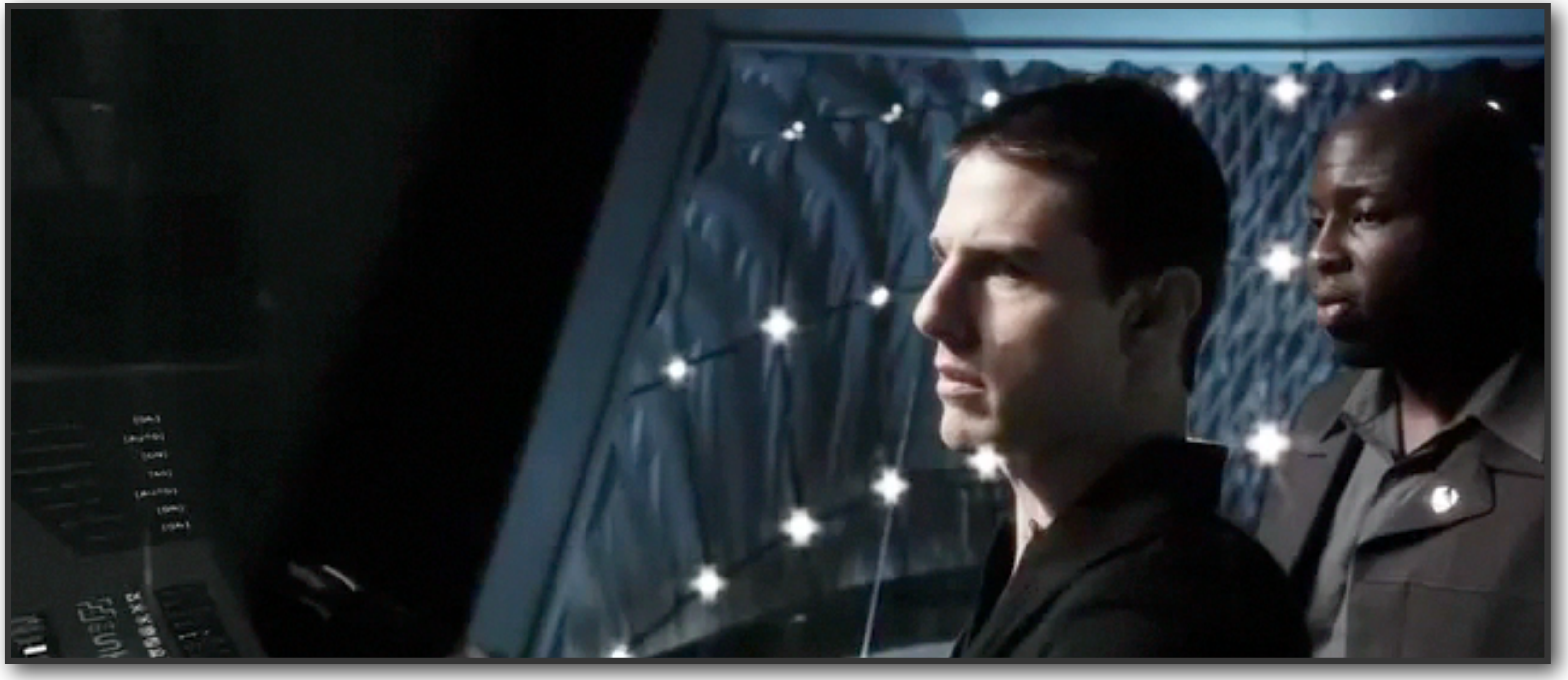
Development of high quality Natural User Interfaces (NUI)

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Current Scenario

Popularization of real time depth sensors

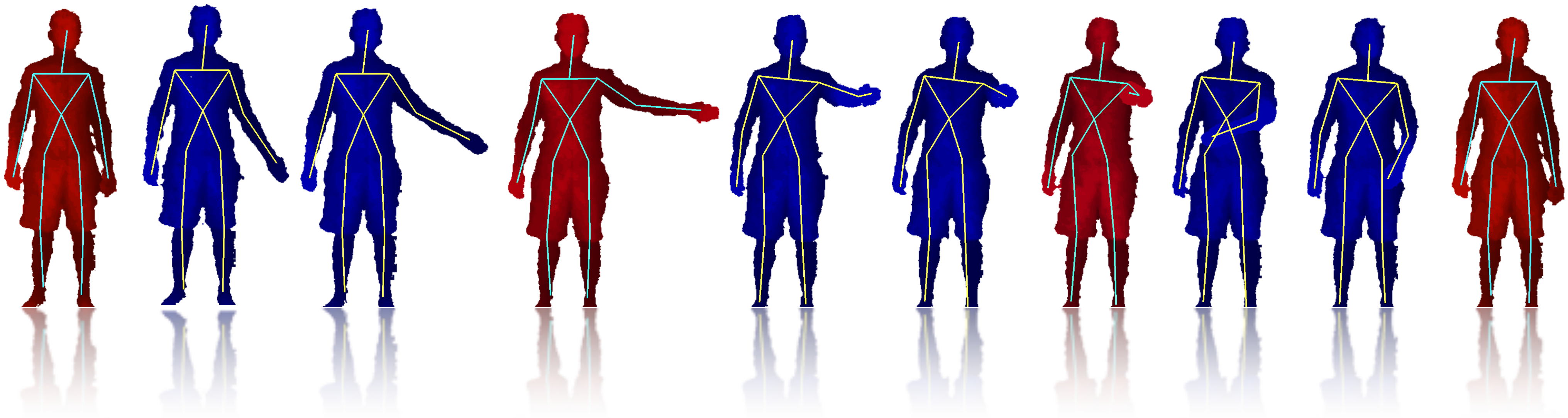
Microsoft Kinect Sensor



Challenging task! Gestures performed at different speeds and/or sequence of poses

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Our approach: key poses learning



Gestures can be characterized by a few extreme poses!

✓ Real-time gesture learning and recognition

✓ Ideal for the average inexperienced user

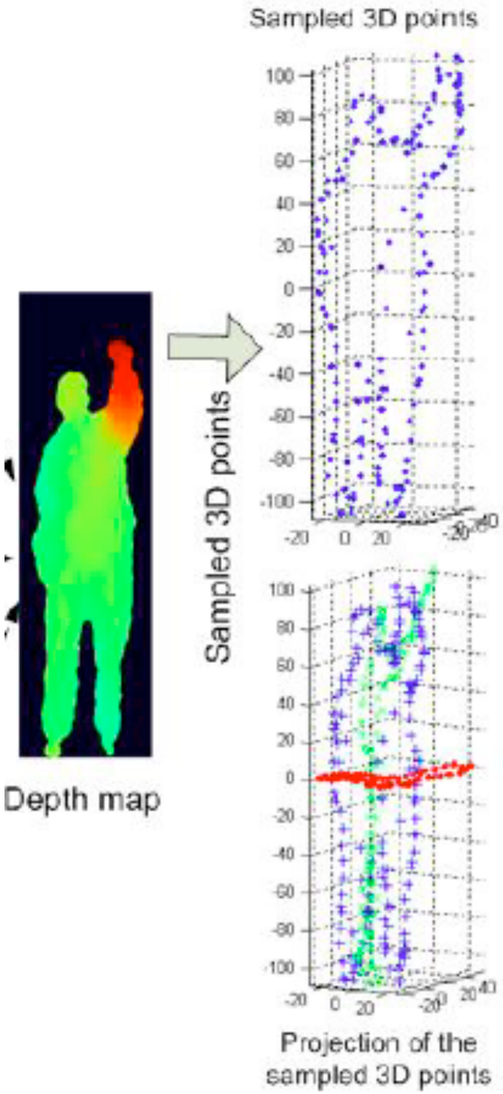
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Outline

1. Related Work
2. Overview
3. Joint-angles Representation
4. Key Pose Statistical Learning
5. Gesture Recognition Through Decision Forests
6. Results

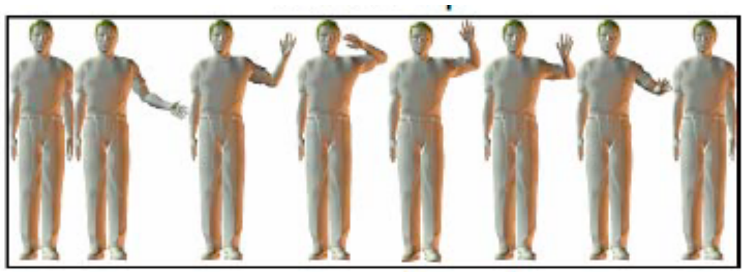
Related Work

Local methods



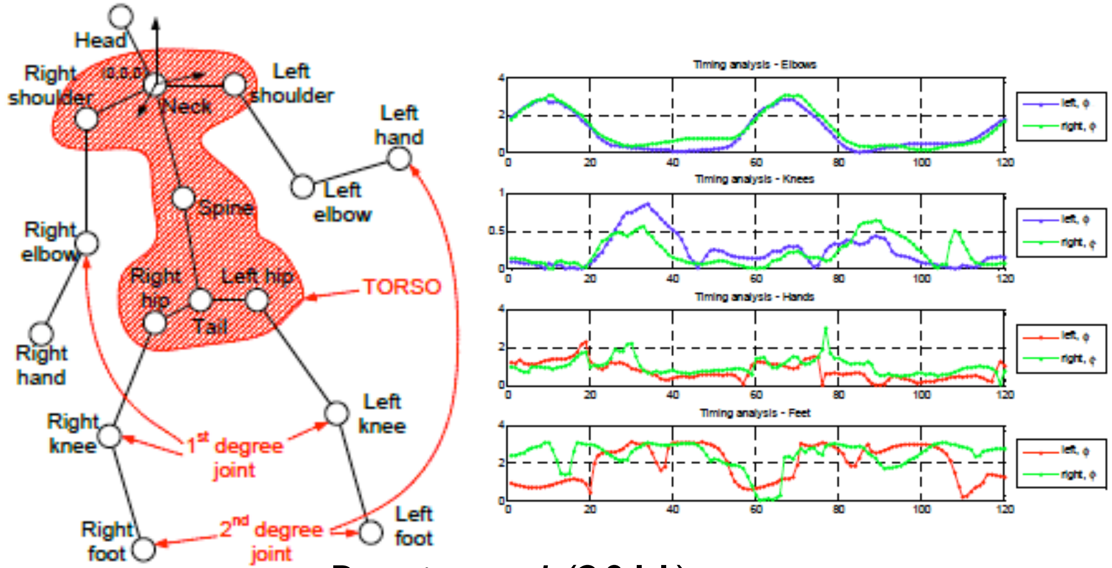
Li *et al* (2010)

Global methods



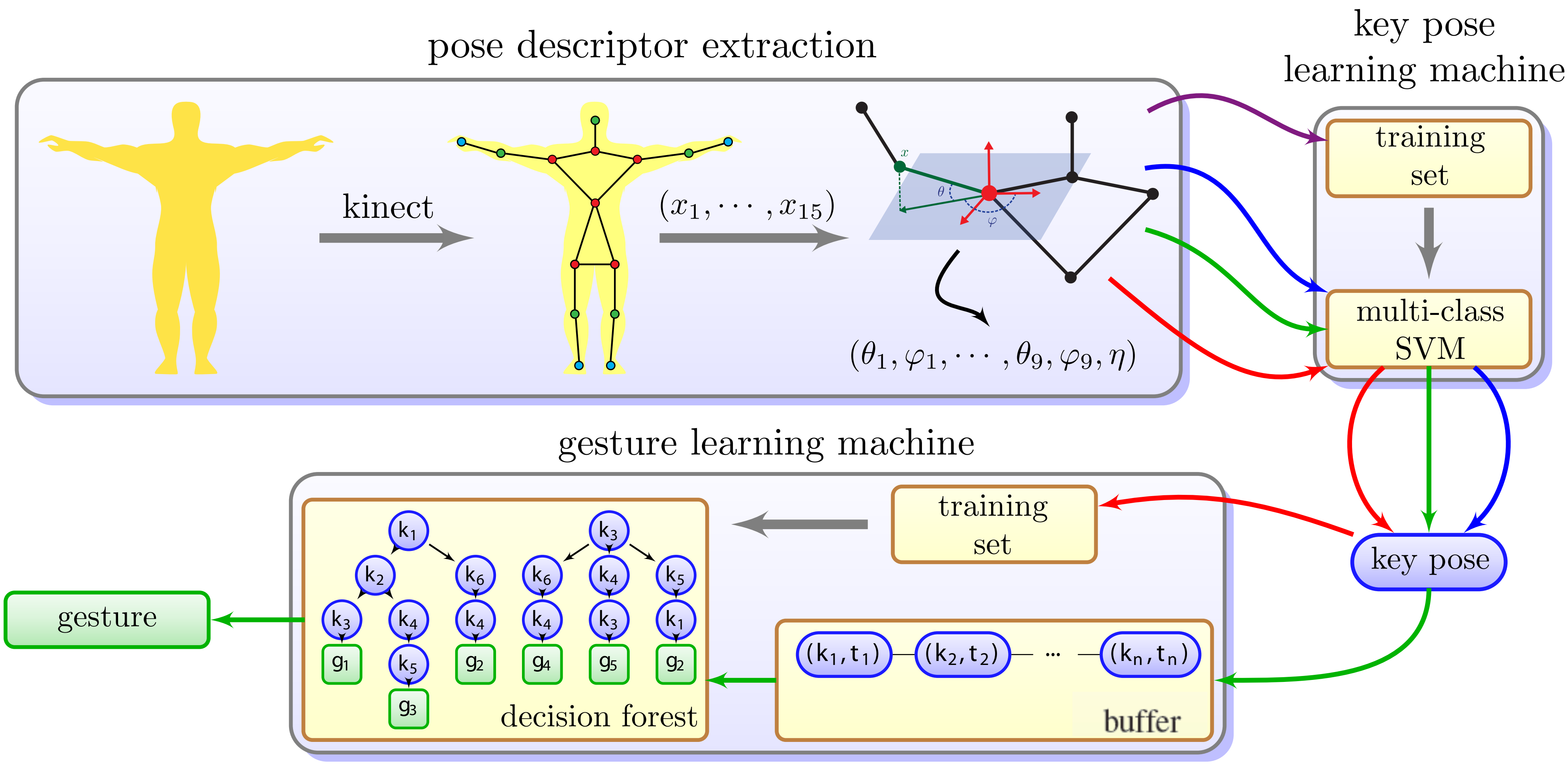
Lv and Nevatia (2007)

Parametric methods



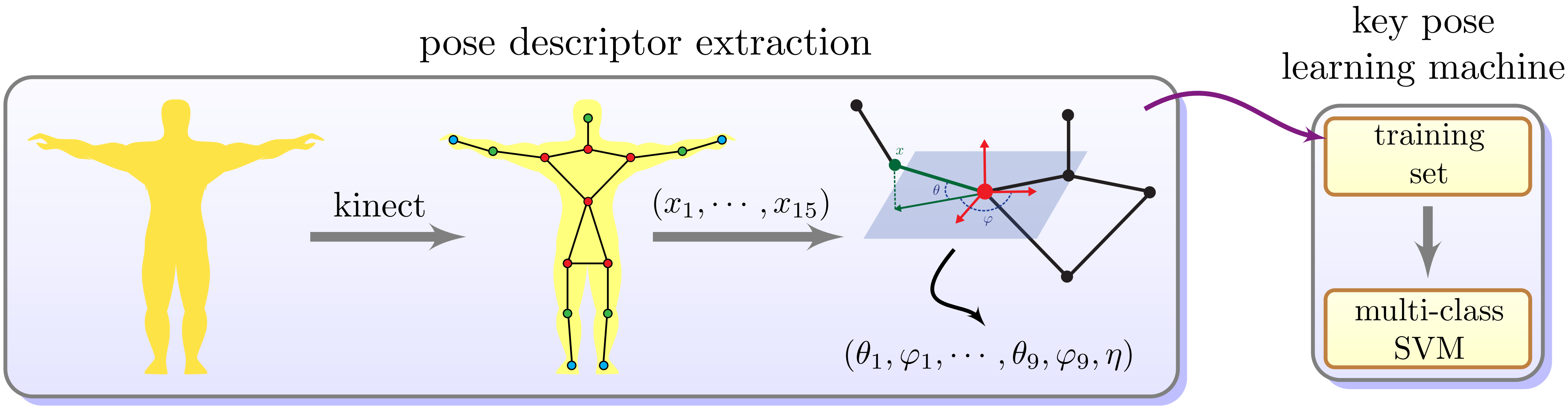
Raptis *et al* (2011)

Overview

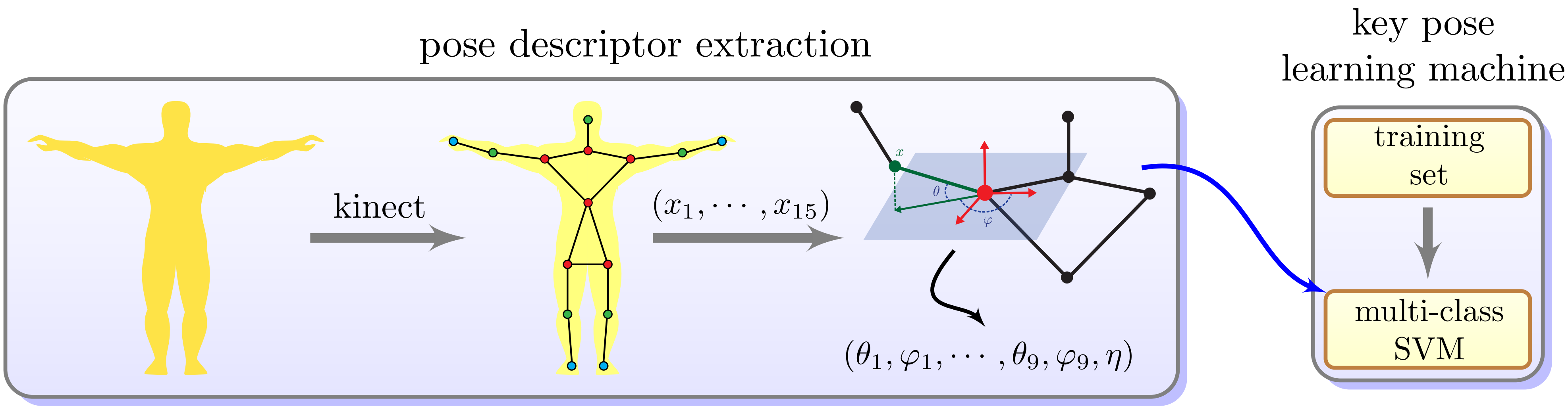


 Real-time gesture recognition from depth data through key poses learning and decision forests

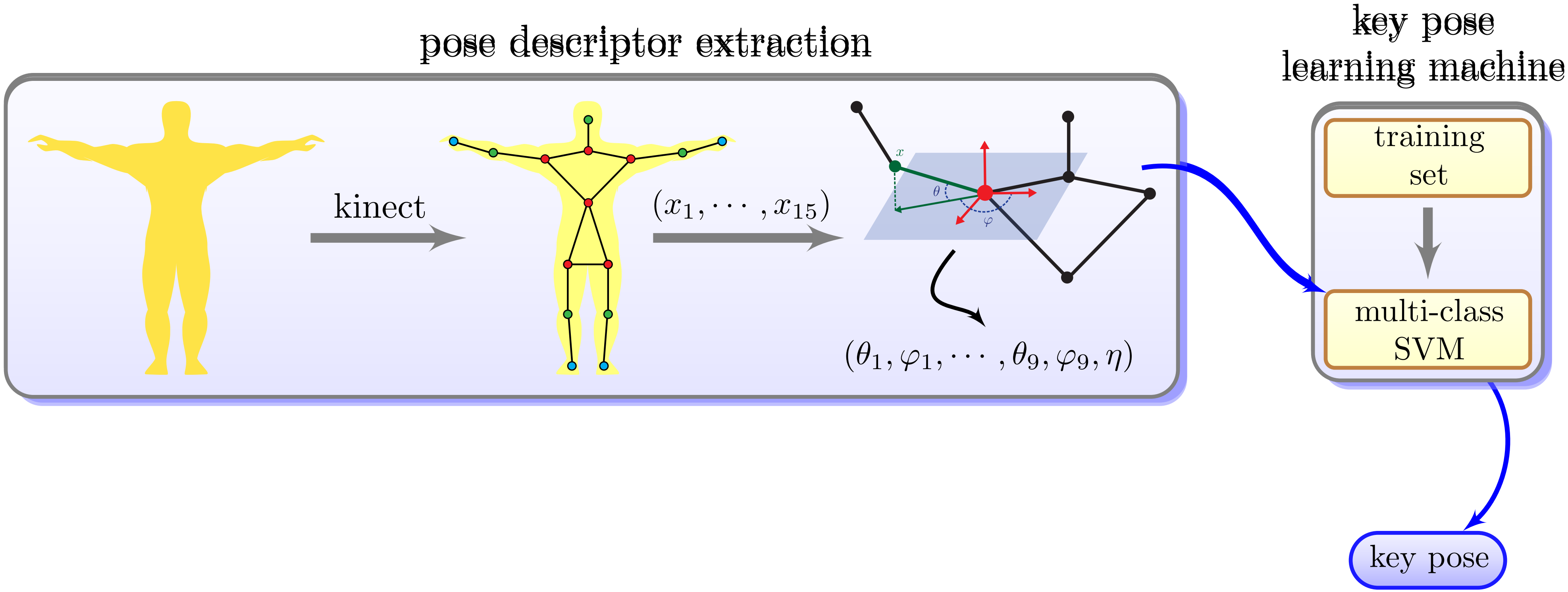
Overview: training key poses



Overview: recognizing key poses

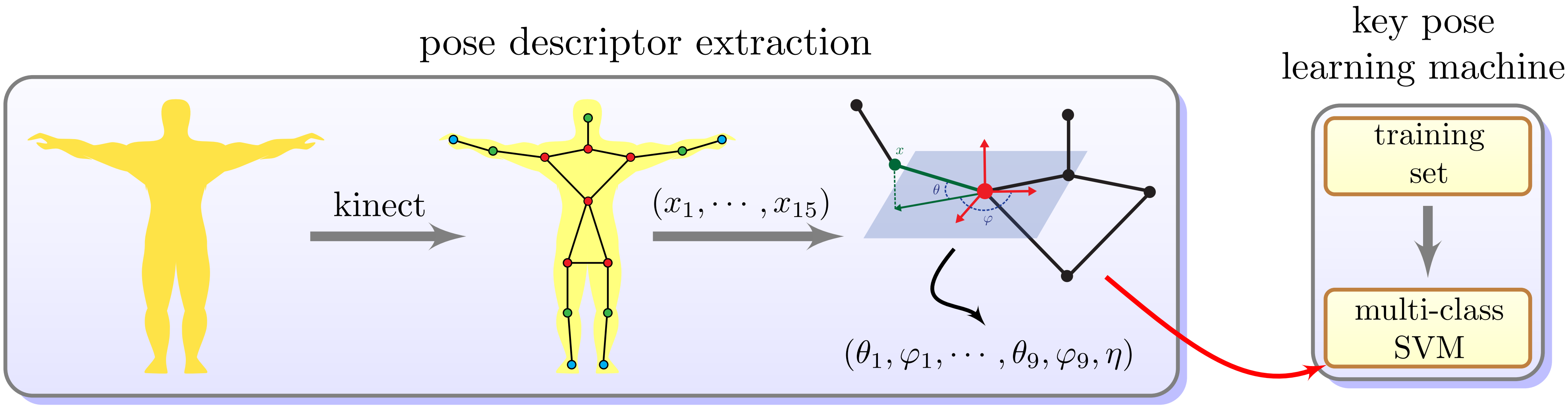


Overview: recognizing key poses

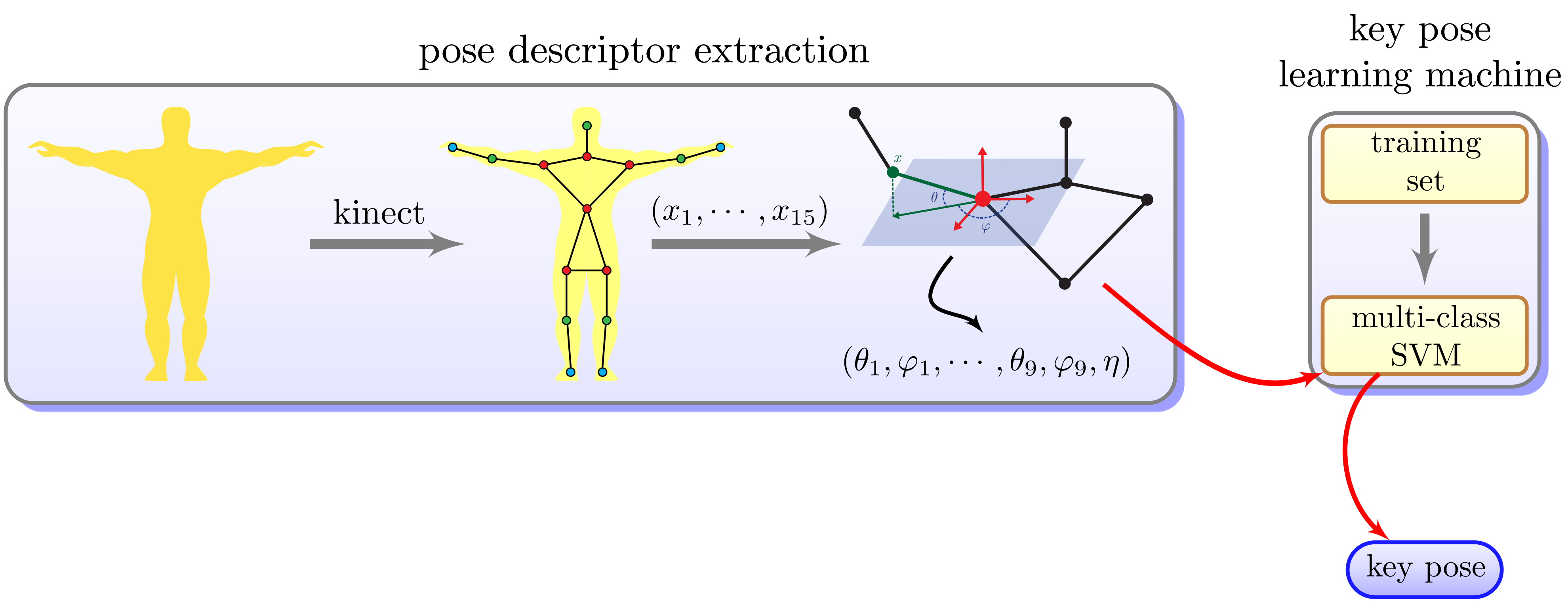


Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

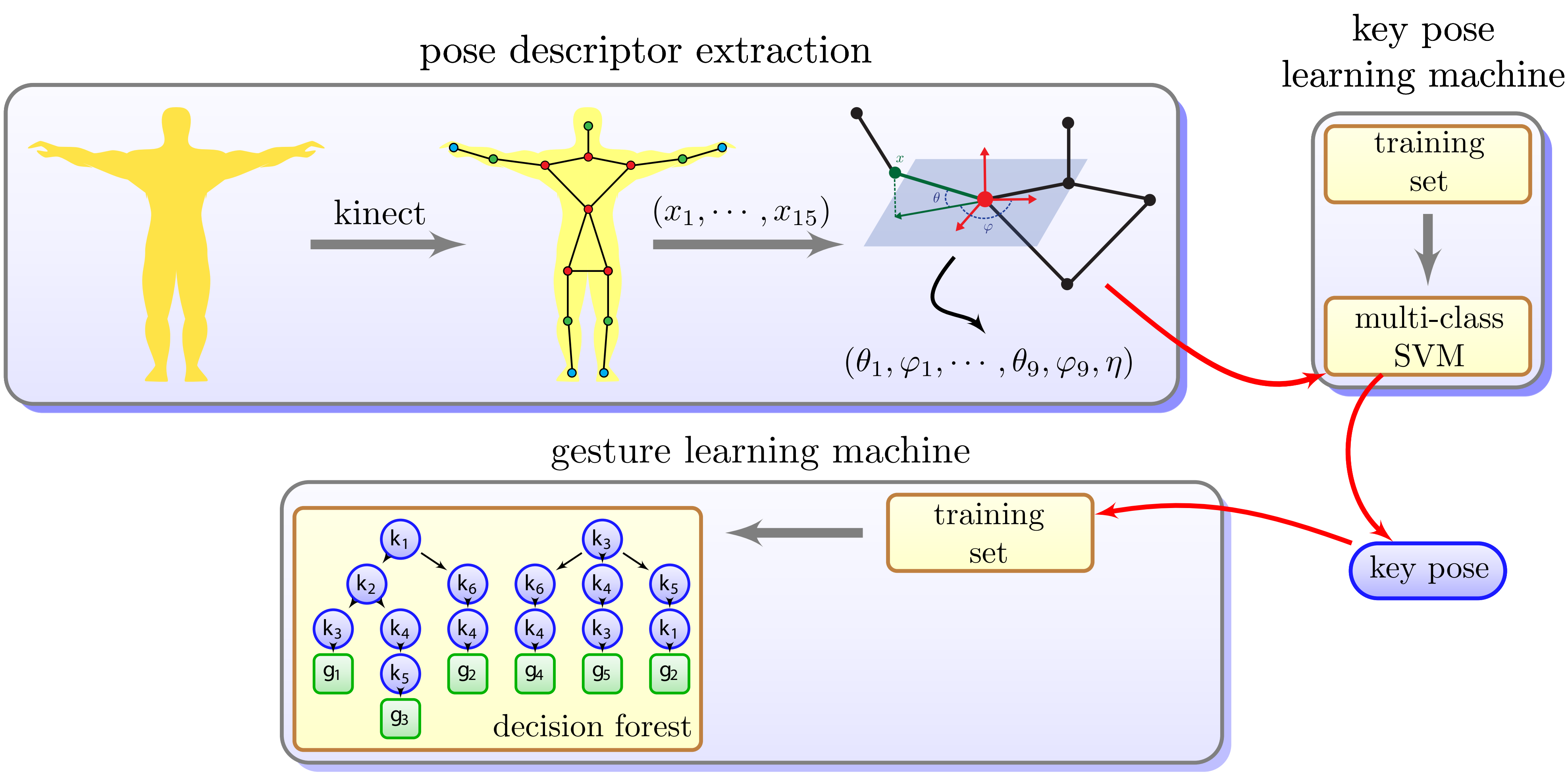
Overview: training gestures



Overview: training gestures

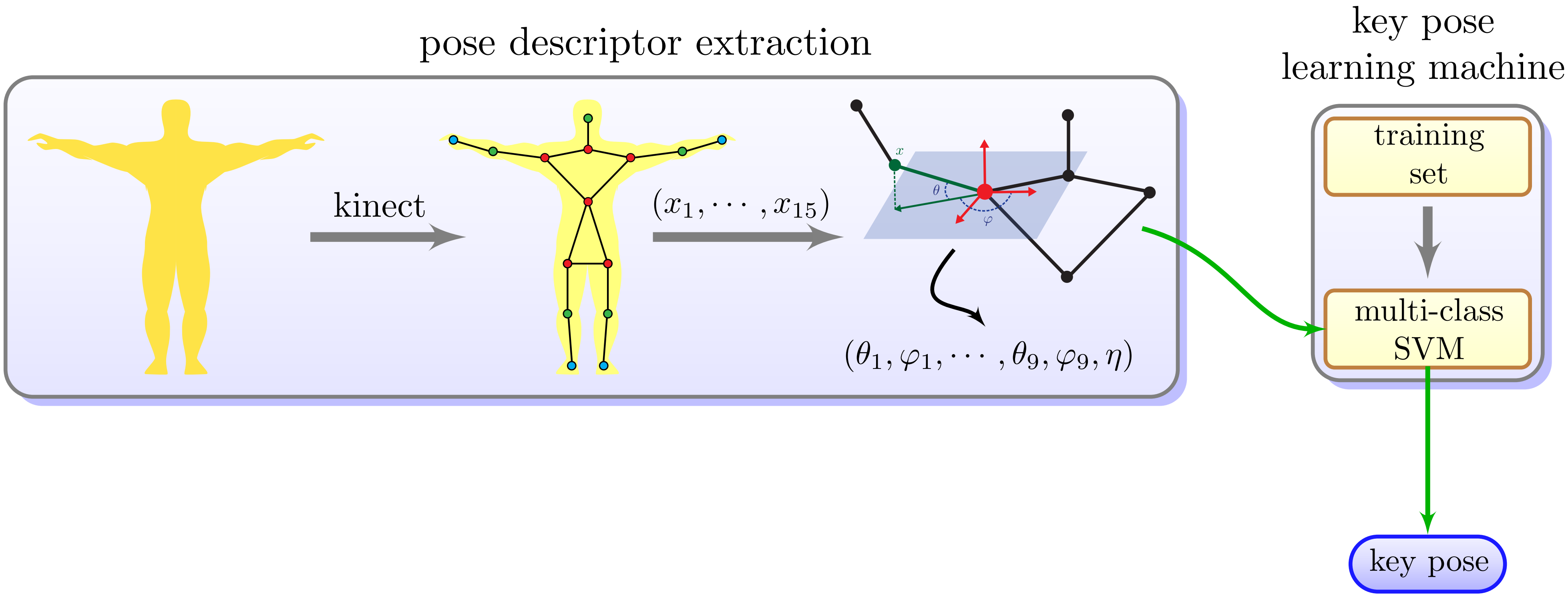


Overview: training gestures

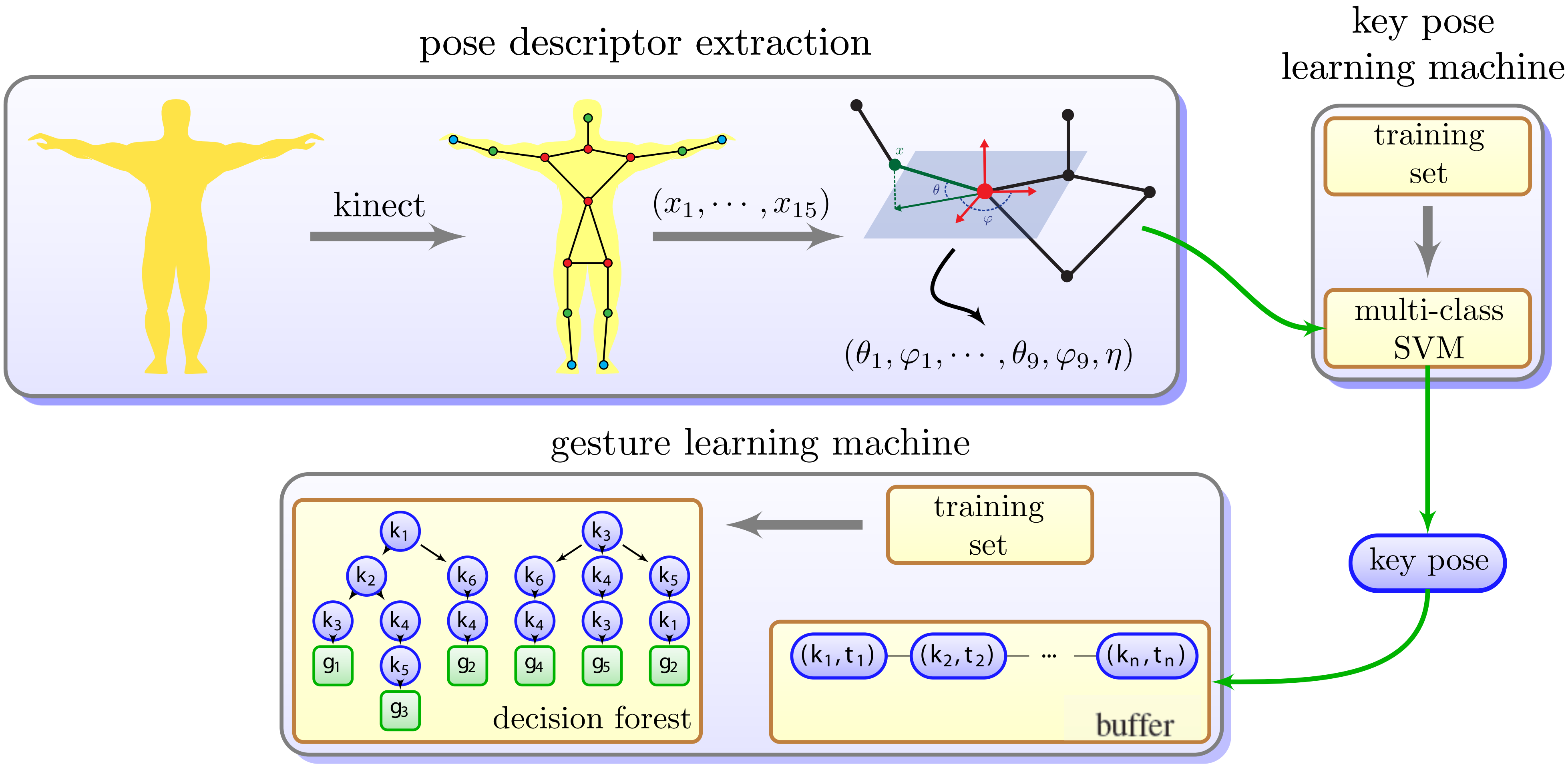


Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Overview: recognizing gestures

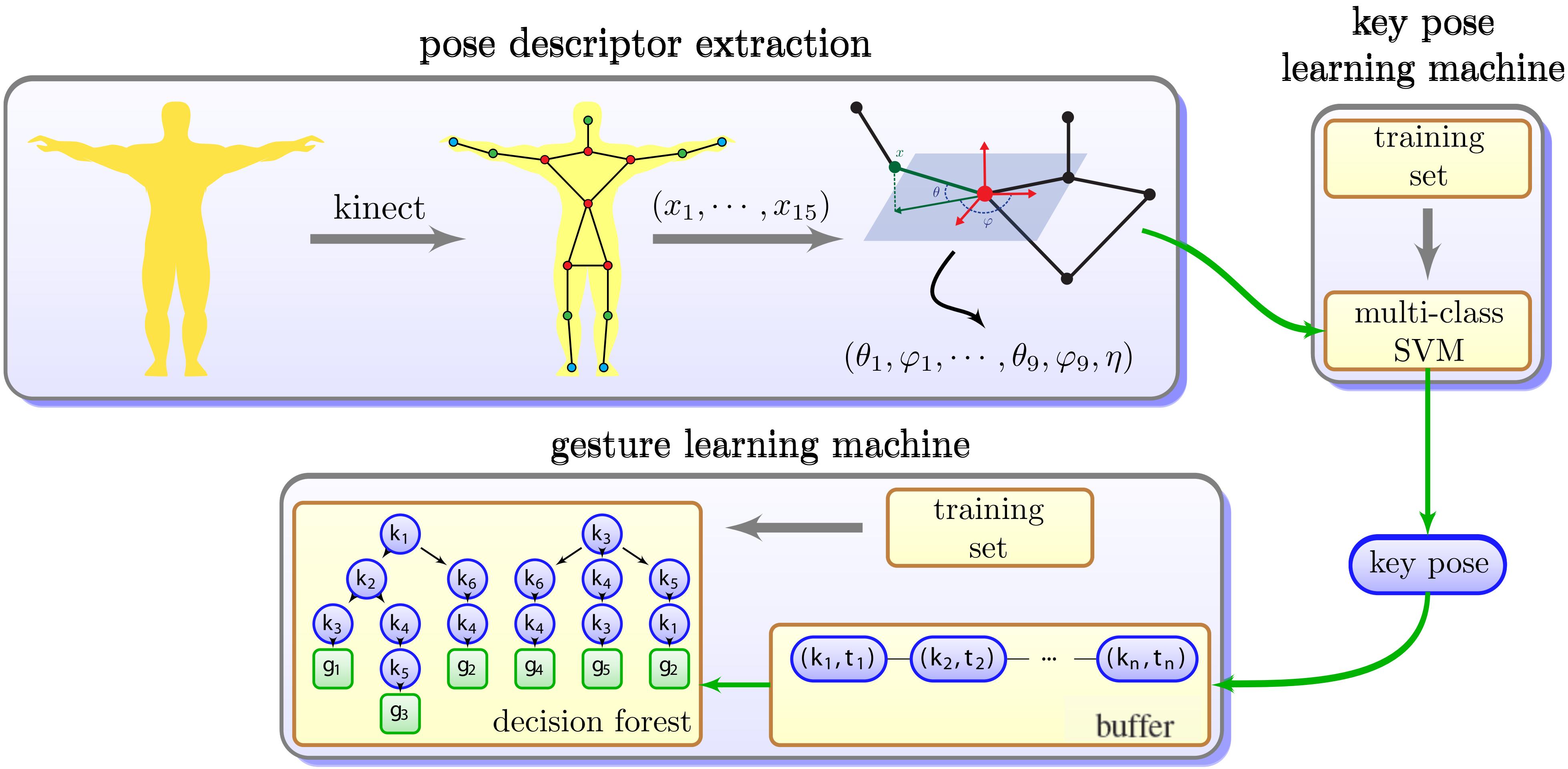


Overview: recognizing gestures



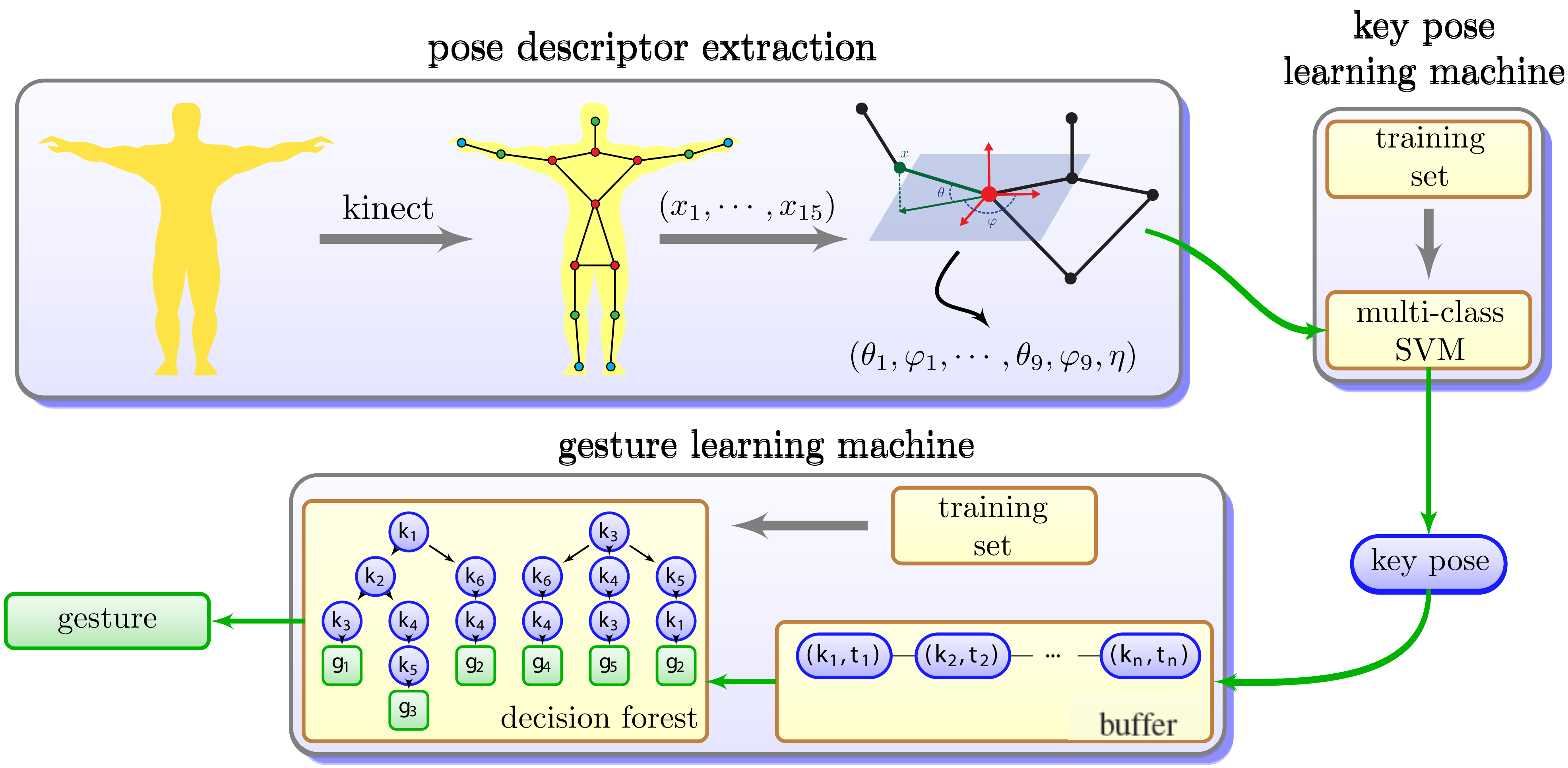
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Overview: recognizing gestures



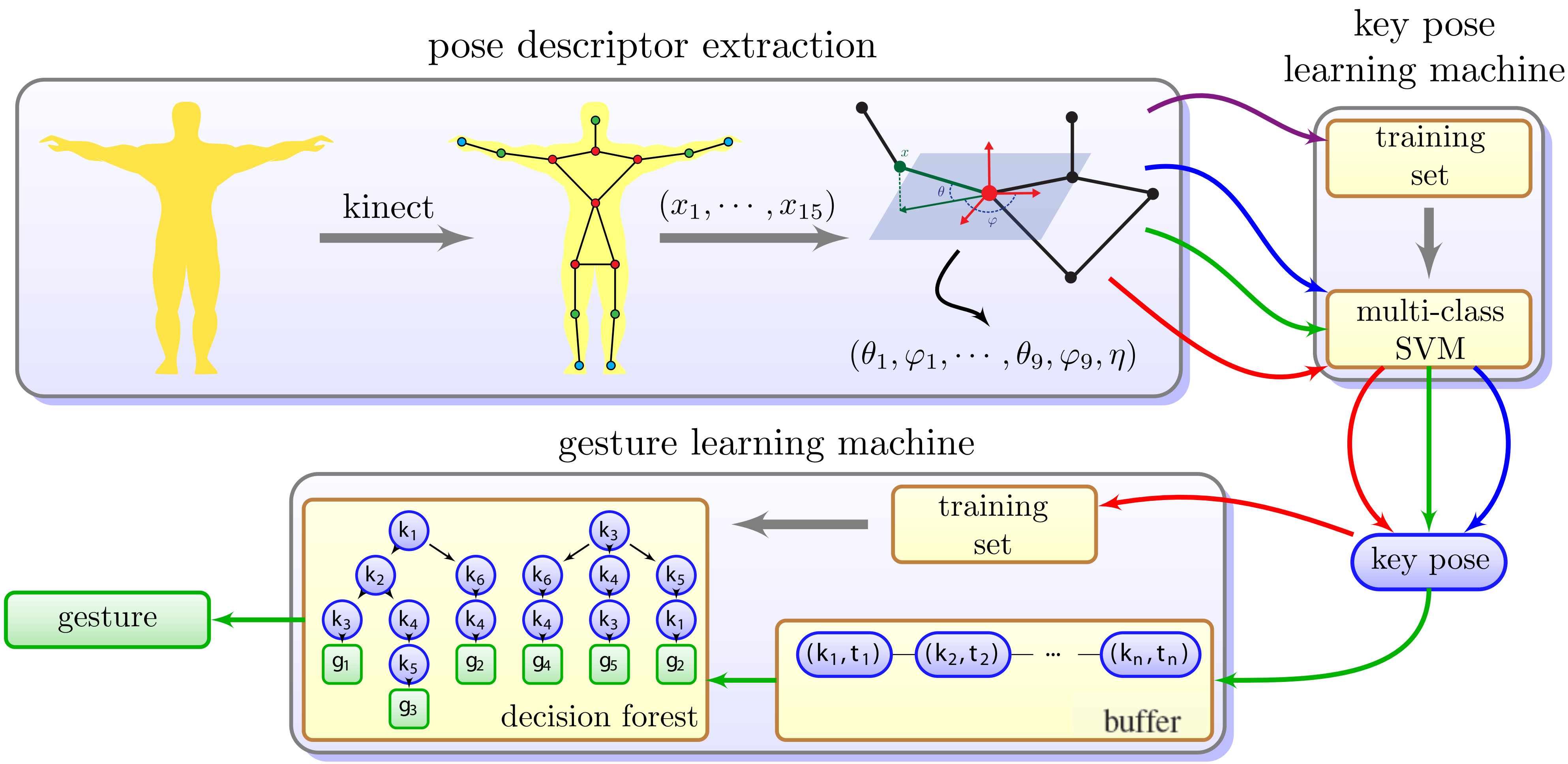
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Overview: recognizing gestures



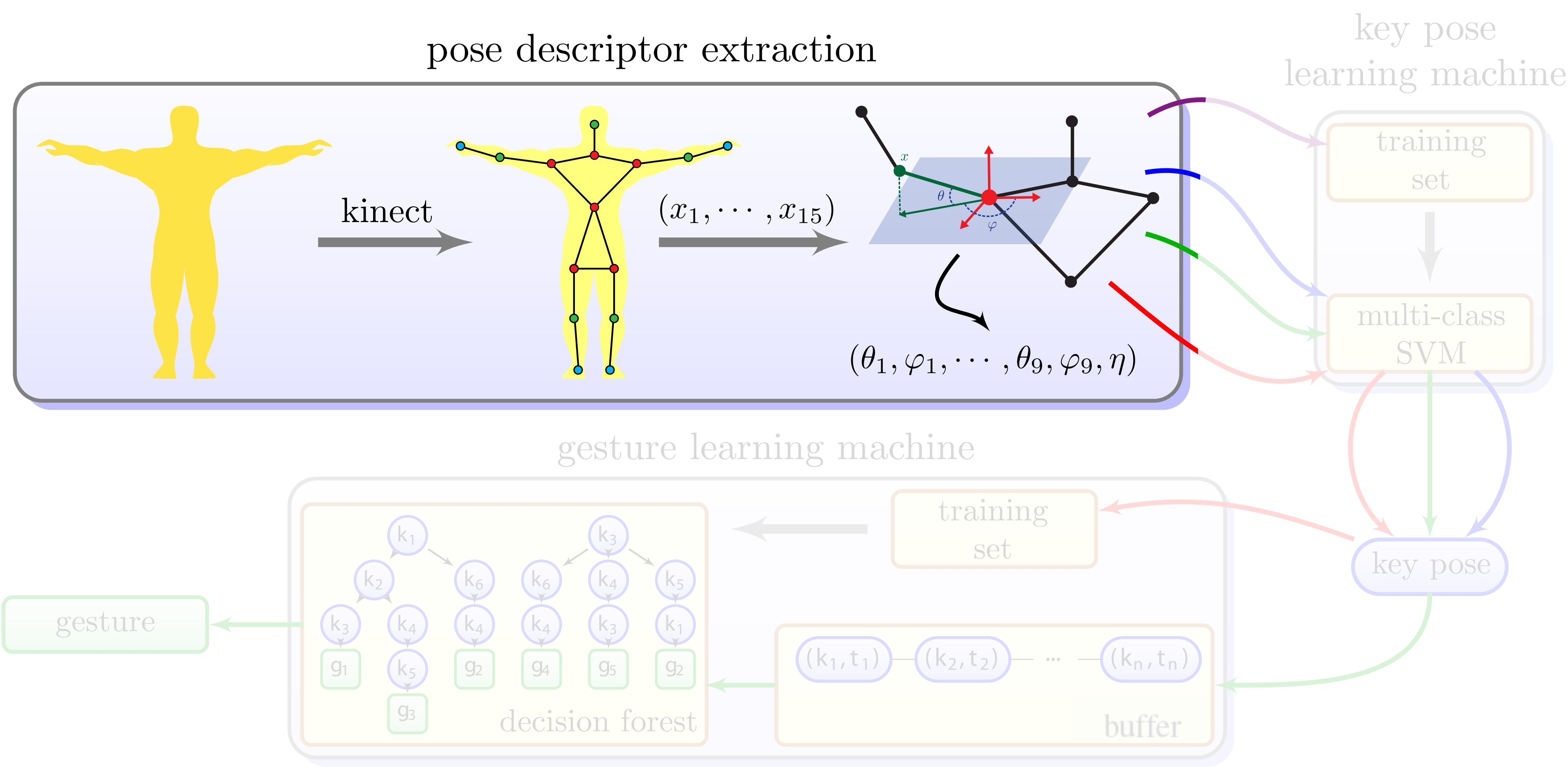
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Overview



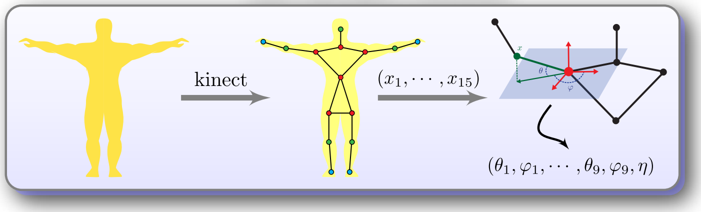
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

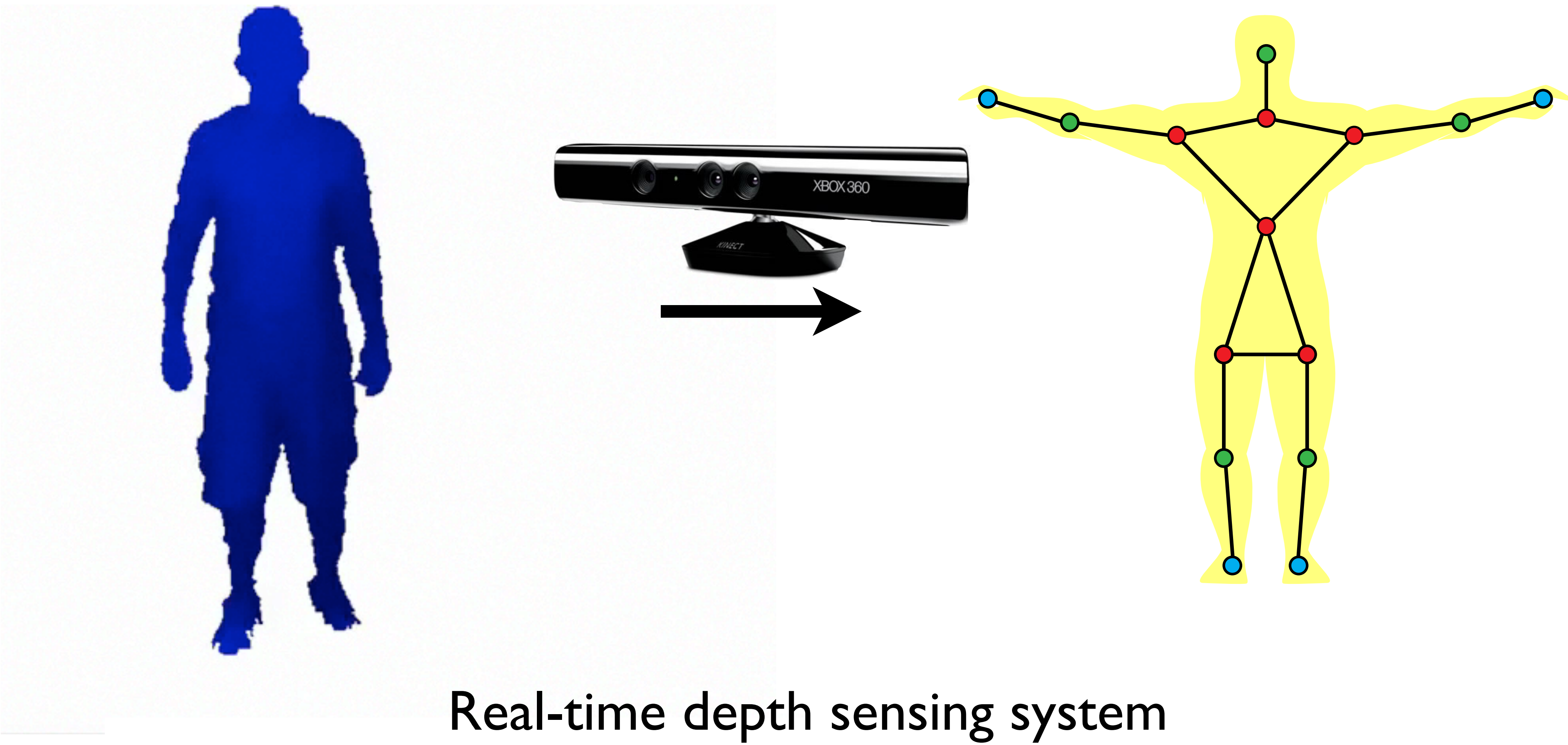
Overview



Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Skeletons from Kinect Sensor



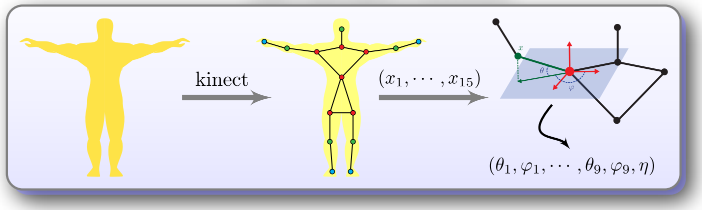


Real-time depth sensing system
streaming depth data and skeletons at 30fps

Miranda et al., 2012

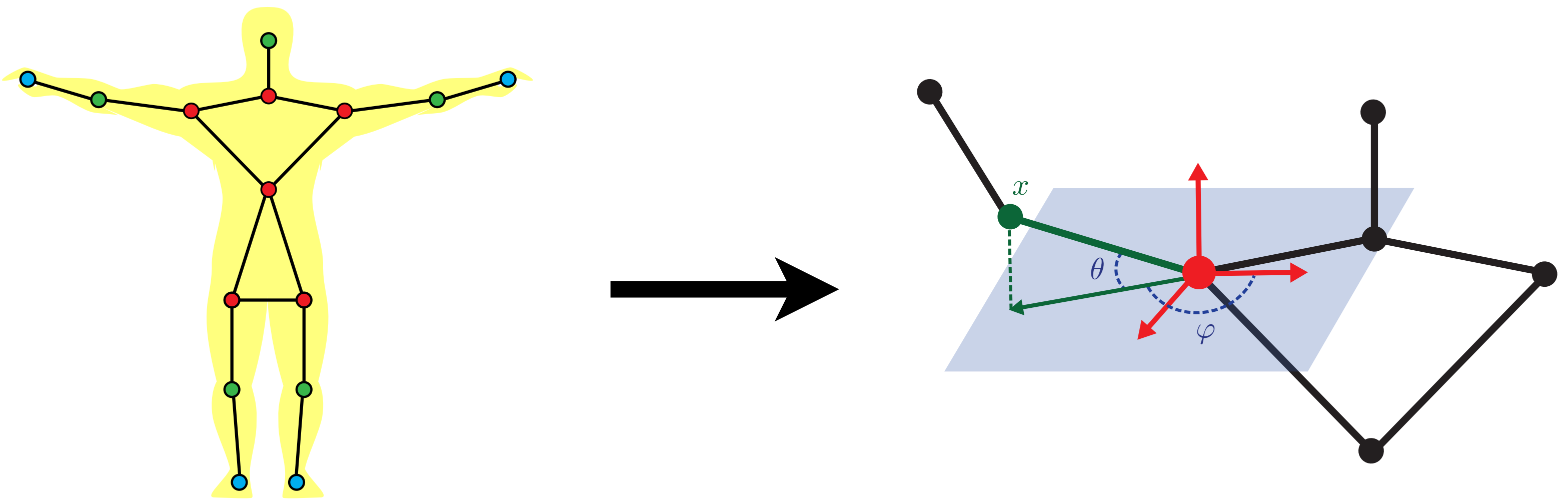
Real-time gesture recognition from depth data through key poses learning and decision forests

Joint-Angles Pose Descriptor



Objective: Concise and invariant representation of relevant pose information.

Improvement of Raptis *et al* (2011) local spherical coordinates.



$(x_1, x_2, \ldots, x_{15}) \in \mathbb{R}^{45}$

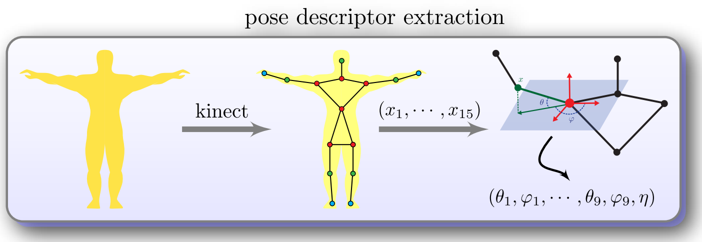
$(\theta_1, \varphi_1, \ldots, \theta_9, \varphi_9, \eta) \in \mathbb{R}^{19}$

1st degree joints: elbows, knees and head

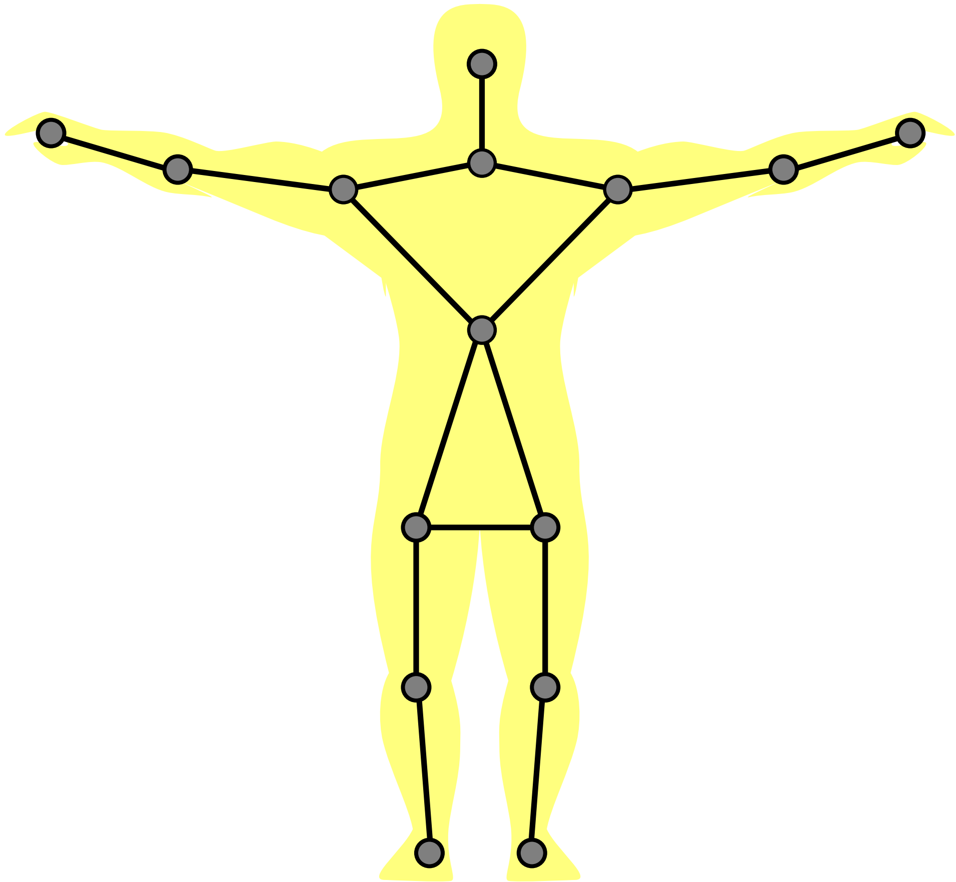
2nd degree joints: hands, feet.

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

How to compute the local bases?



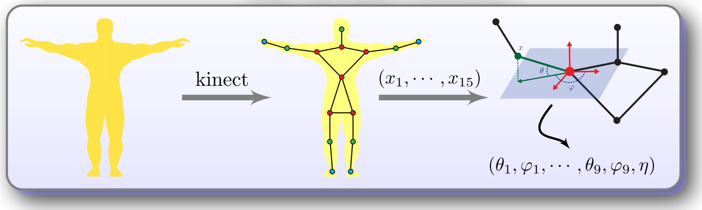
1st degree joints:



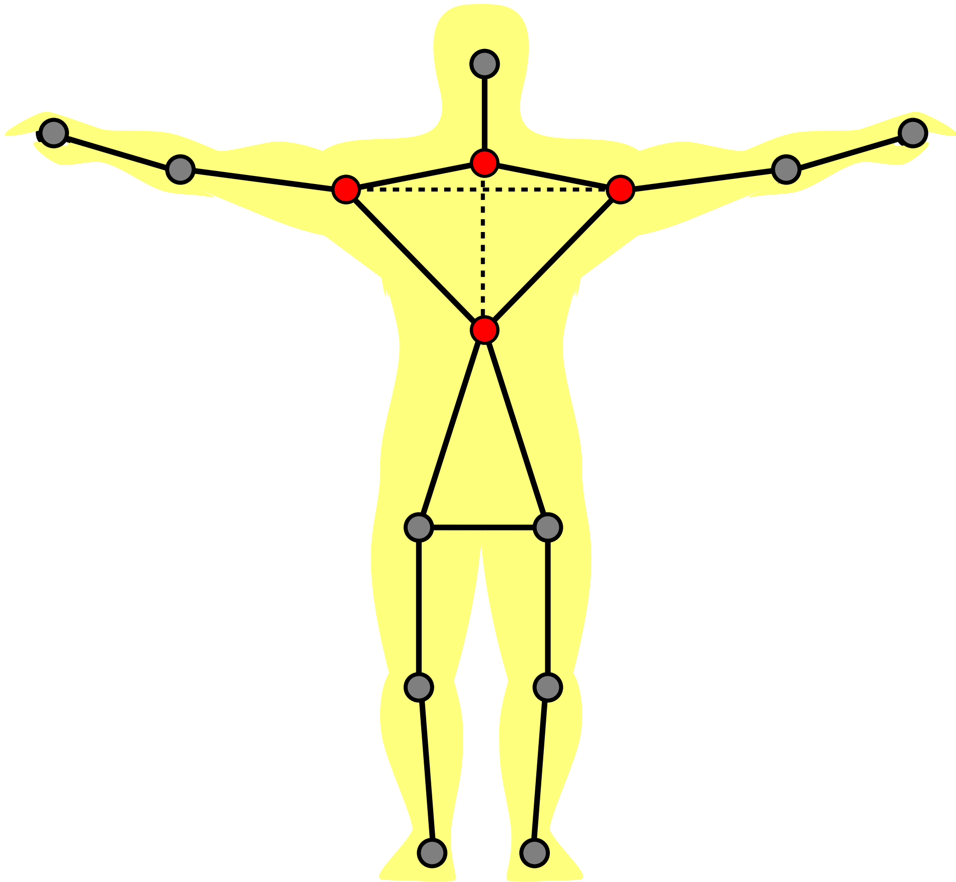
Miranda et al., 2012

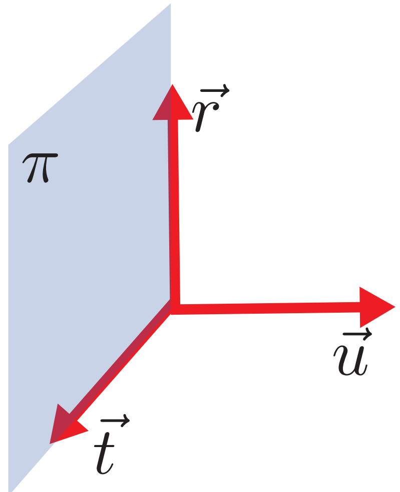
Real-time gesture recognition from depth data through key poses learning and decision forests

How to compute the local bases?



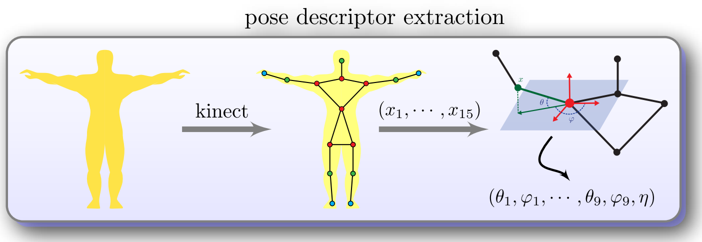
1st degree joints:



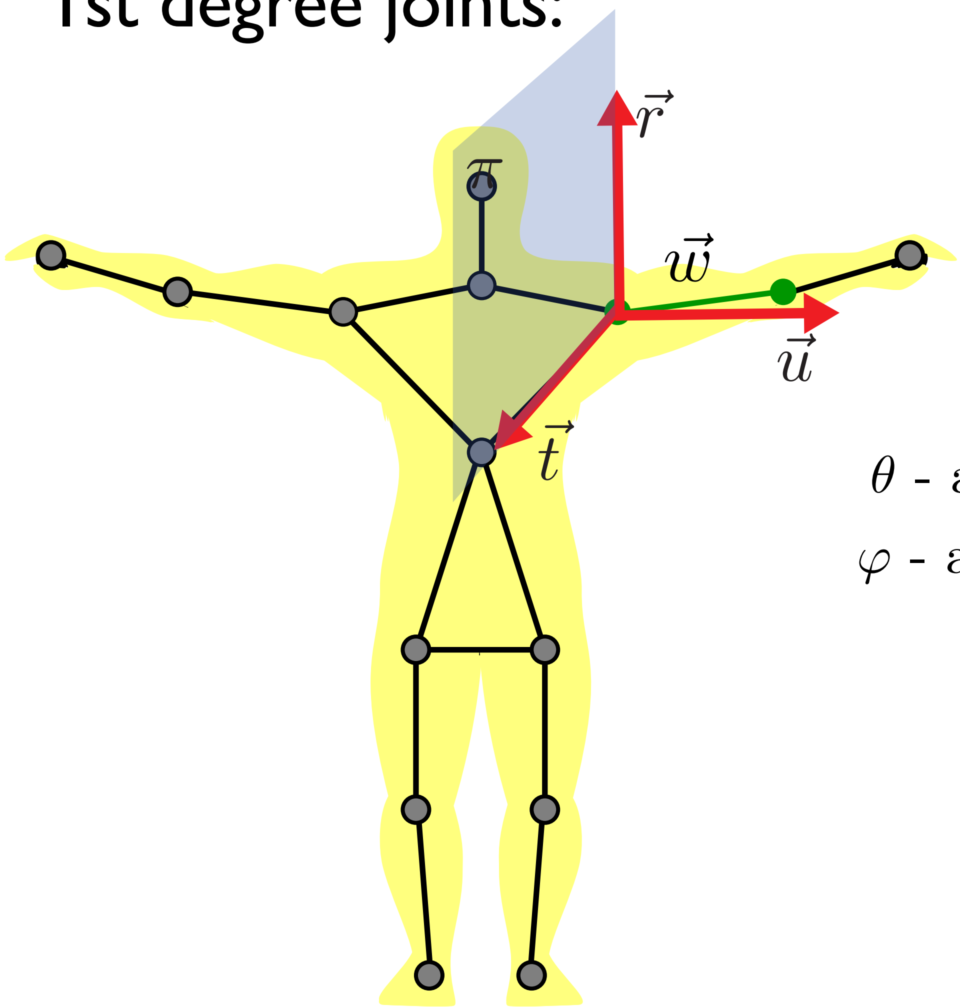


Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

How to compute the local bases?



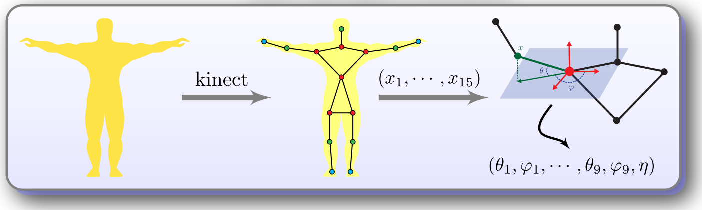
1st degree joints:



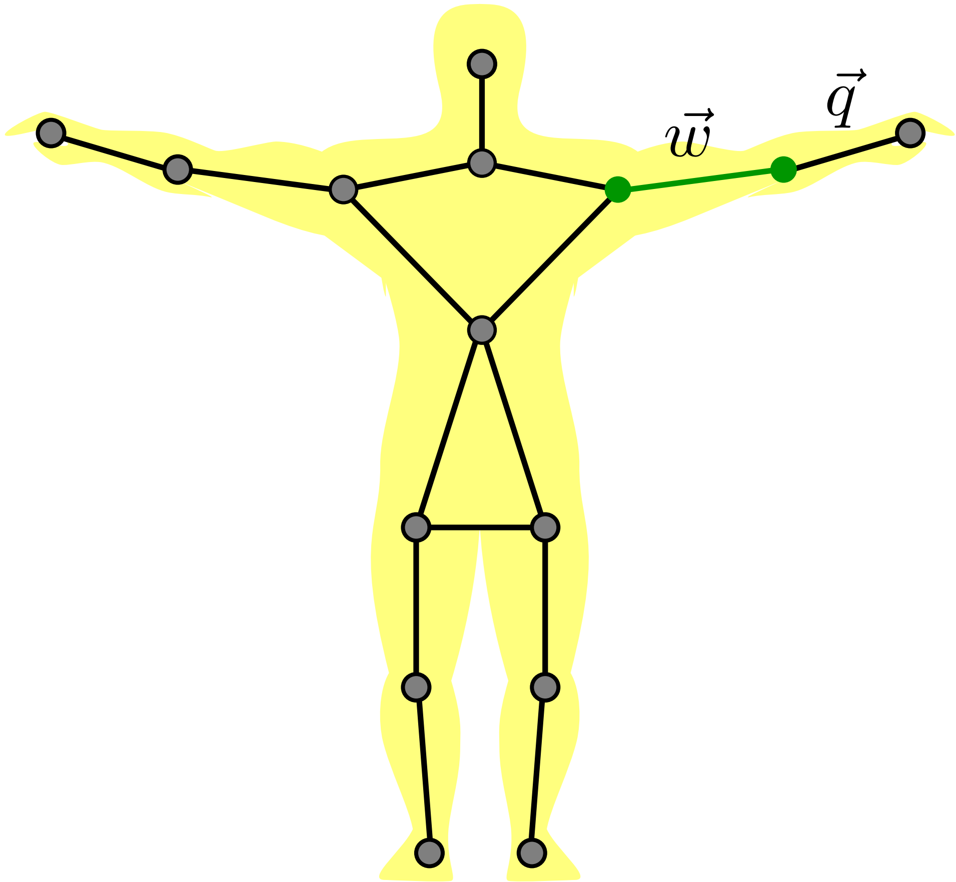
θ - angle between $\vec{u}$ and $\vec{w}$

φ - angle between $\vec{t}$ and the projection of $\vec{w}$ in π

How to compute the local bases?



2nd degree joints:

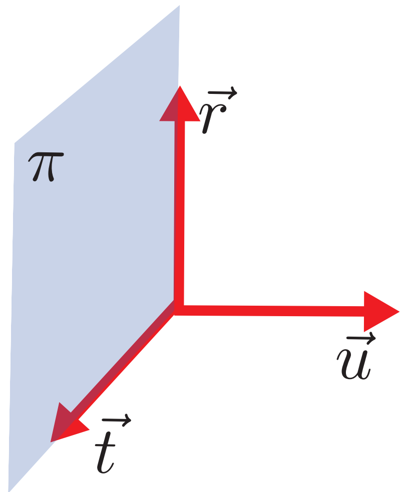


Rotate $\{\vec{u}, \vec{r}, \vec{t}\}$ by

$$\beta = \arccos(\vec{w}, \vec{u})$$

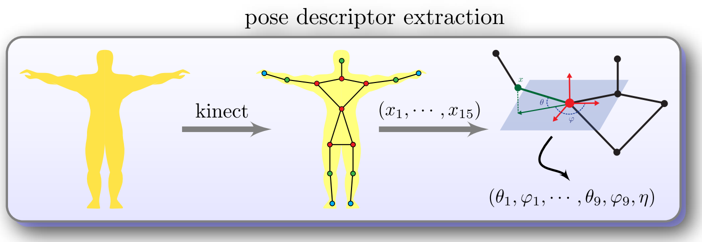
around

$$b = \vec{w} \times \vec{u}$$

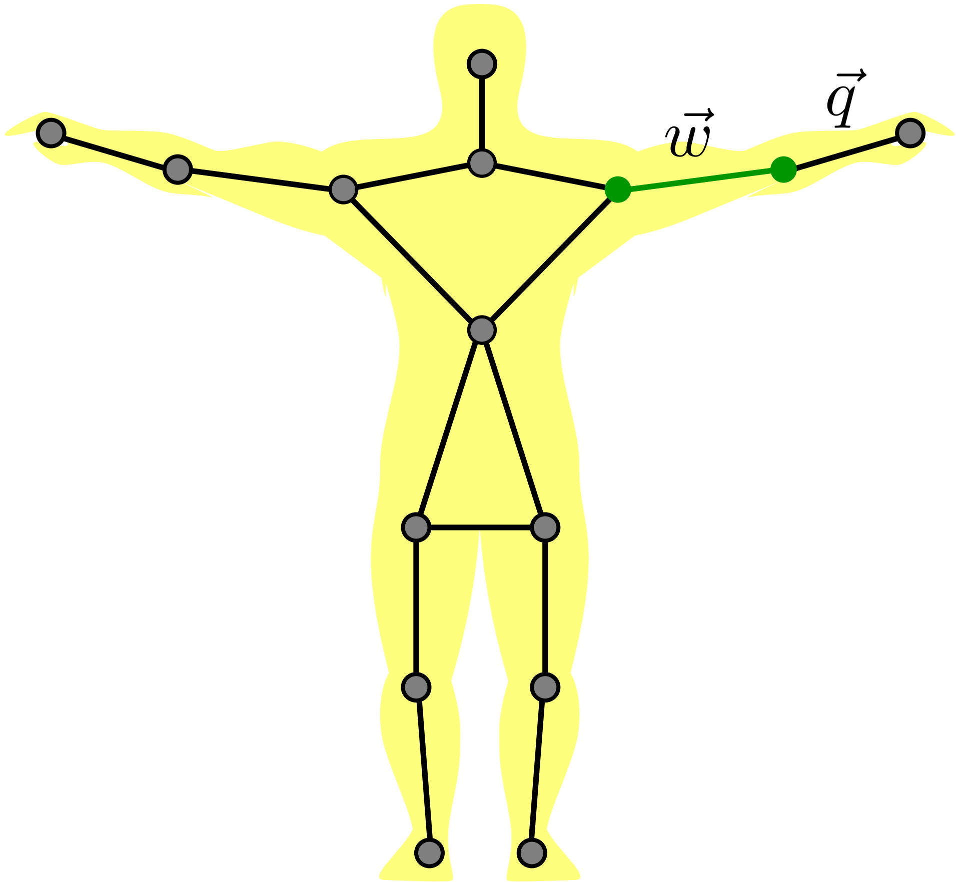


Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

How to compute the local bases?



2nd degree joints:

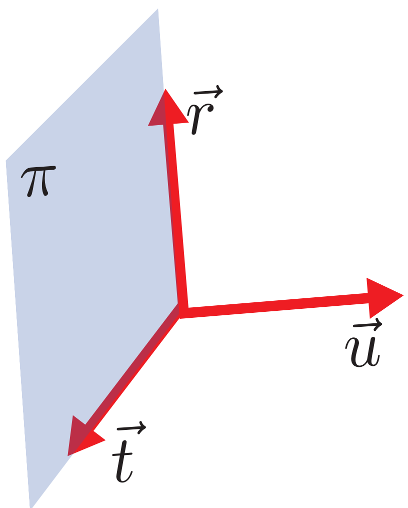


Rotate $\{\vec{u}, \vec{r}, \vec{t}\}$ by

$$\beta = \arccos(\vec{w}, \vec{u})$$

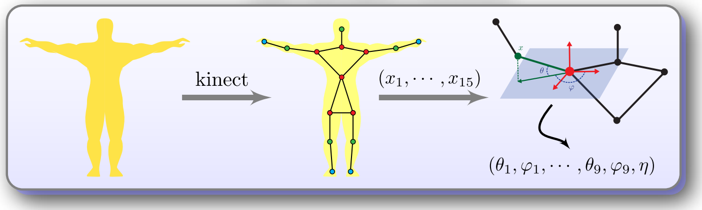
around

$$b = \vec{w} \times \vec{u}$$

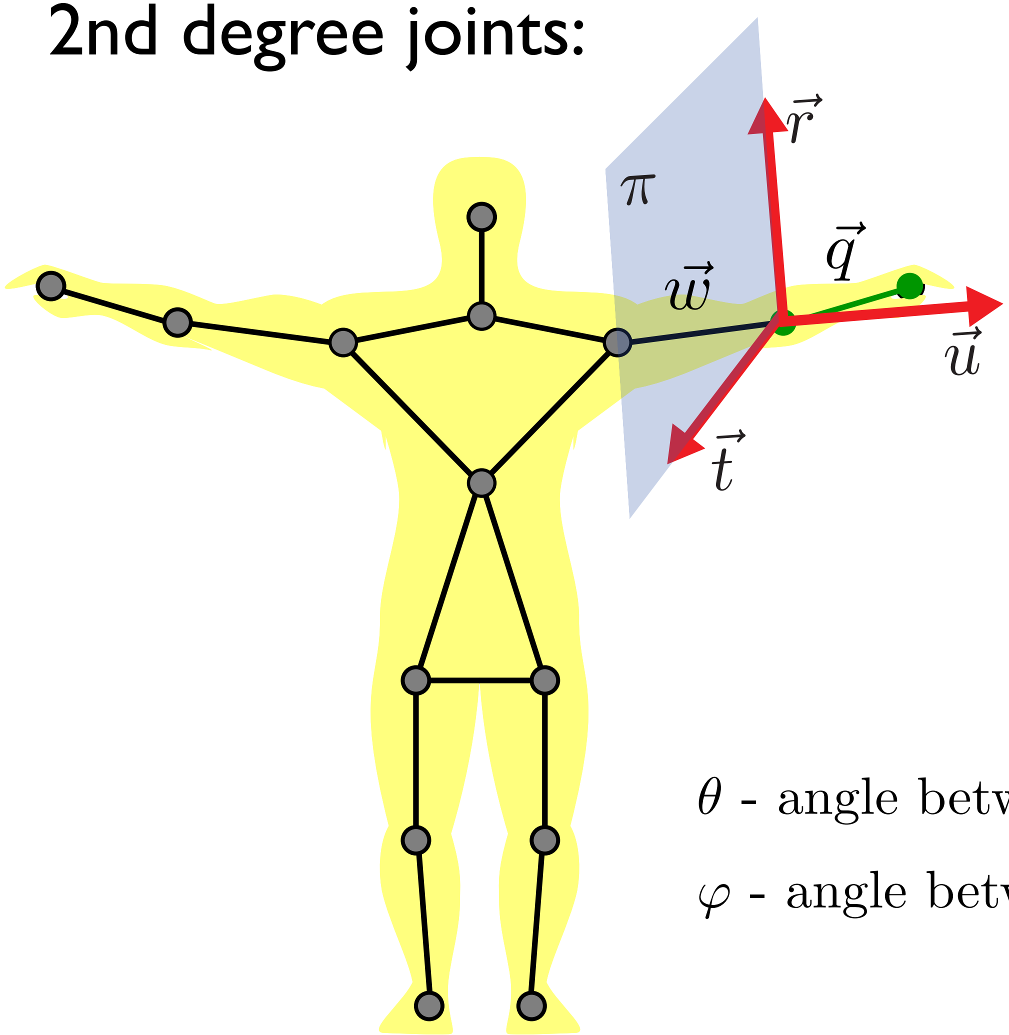


Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

How to compute the local bases?



2nd degree joints:



Rotate $\{\vec{u}, \vec{r}, \vec{t}\}$ by

$$\beta = \arccos(\vec{w}, \vec{u})$$

around

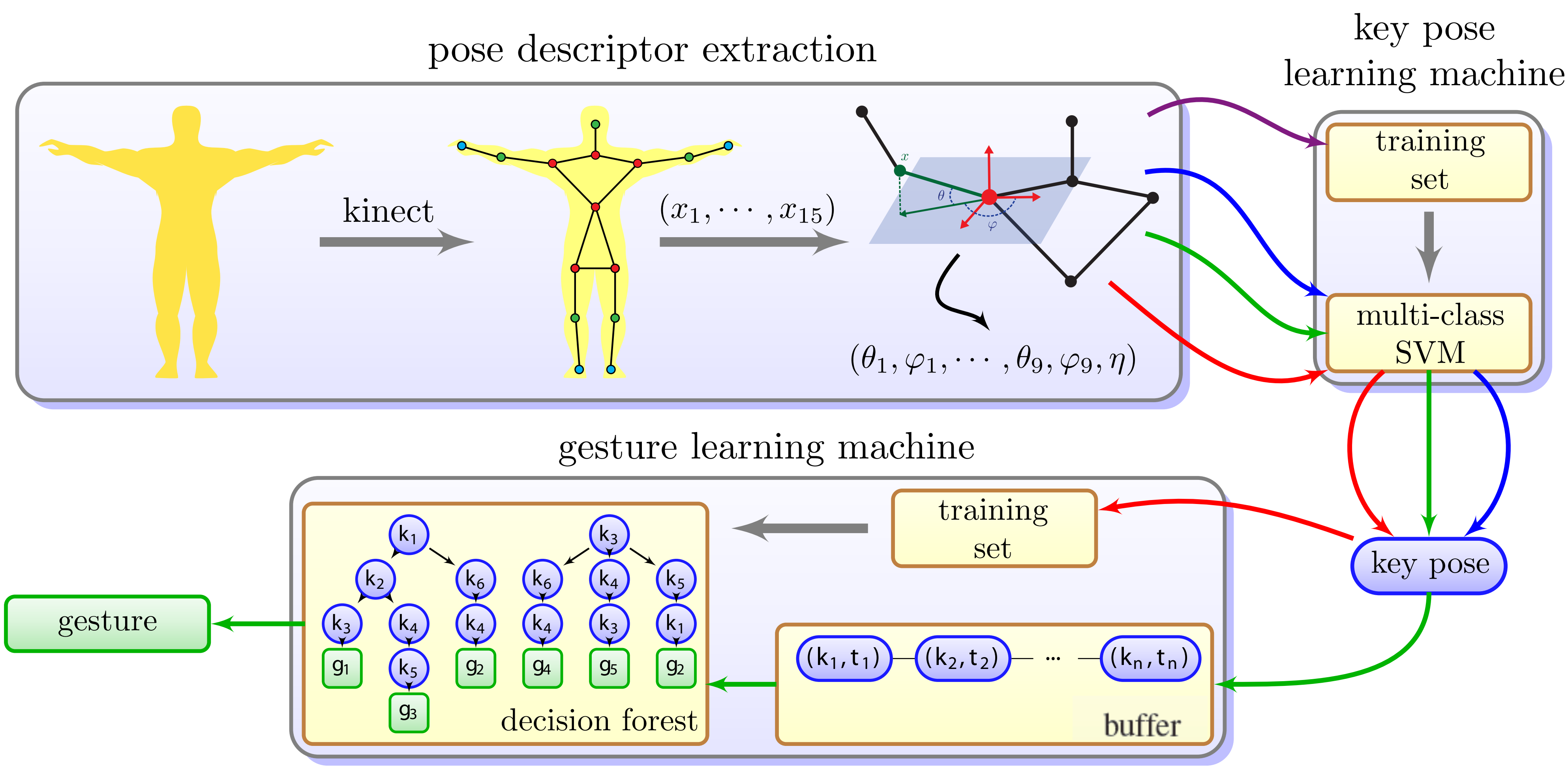
$$b = \vec{w} \times \vec{u}$$

θ - angle between rotated $\vec{u}$ and $\vec{q}$

φ - angle between rotated $\vec{t}$ and the projection of $\vec{q}$ in π

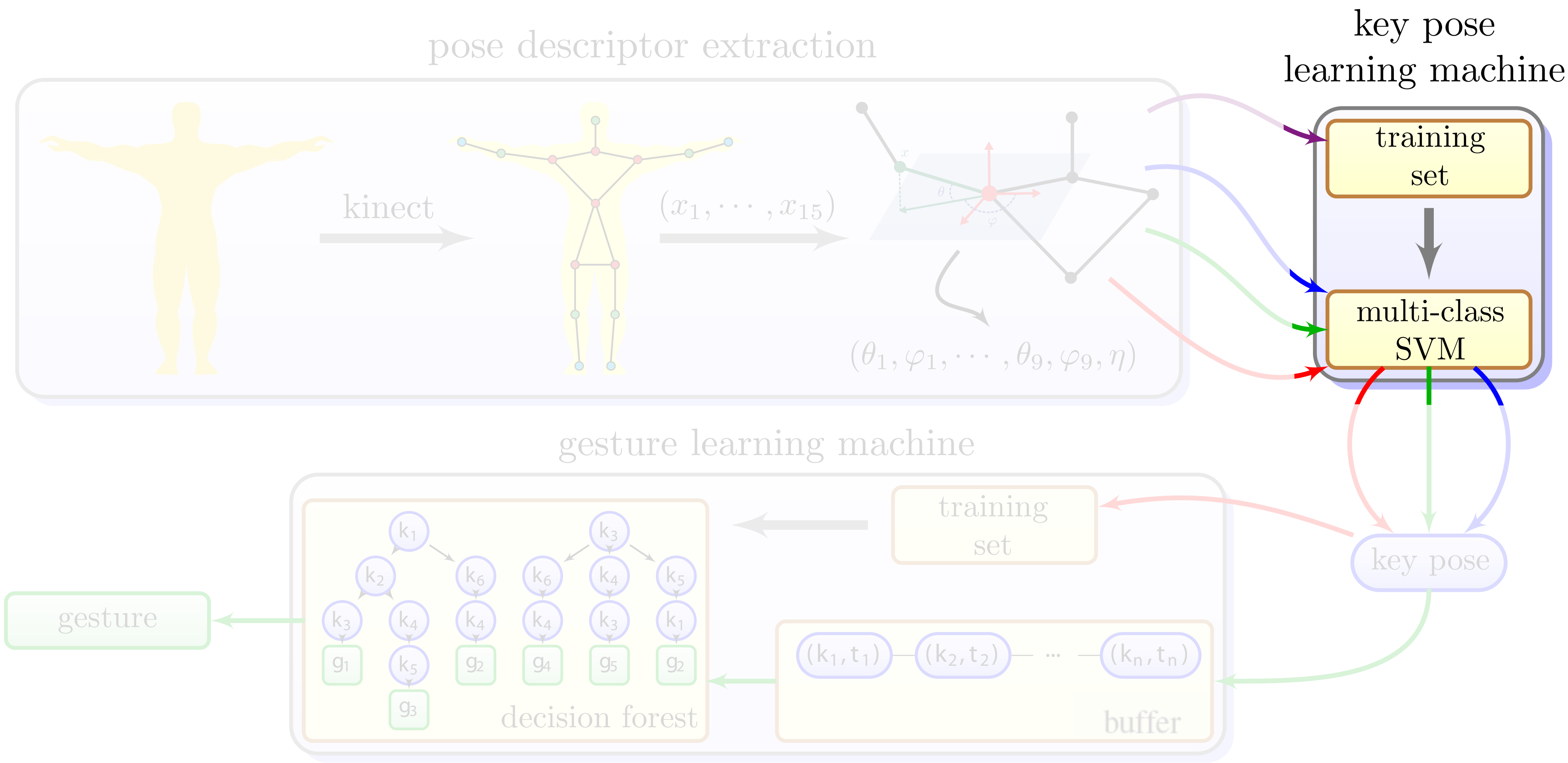
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

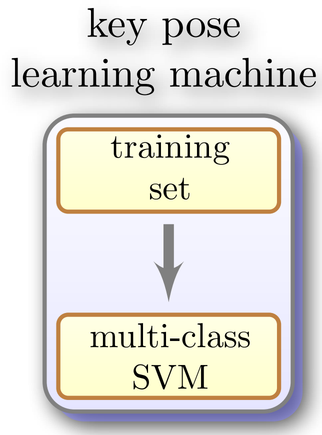
Overview



pose descriptor extraction
kinect
$(x_1, \cdots, x_{15})$
$(\theta_1, \varphi_1, \cdots, \theta_9, \varphi_9, \eta)$
key pose learning machine
training set
multi-class SVM
key pose
gesture learning machine
training set
decision forest
buffer
(k_1, t_1) — (k_2, t_2) — ... — (k_n, t_n)
gesture

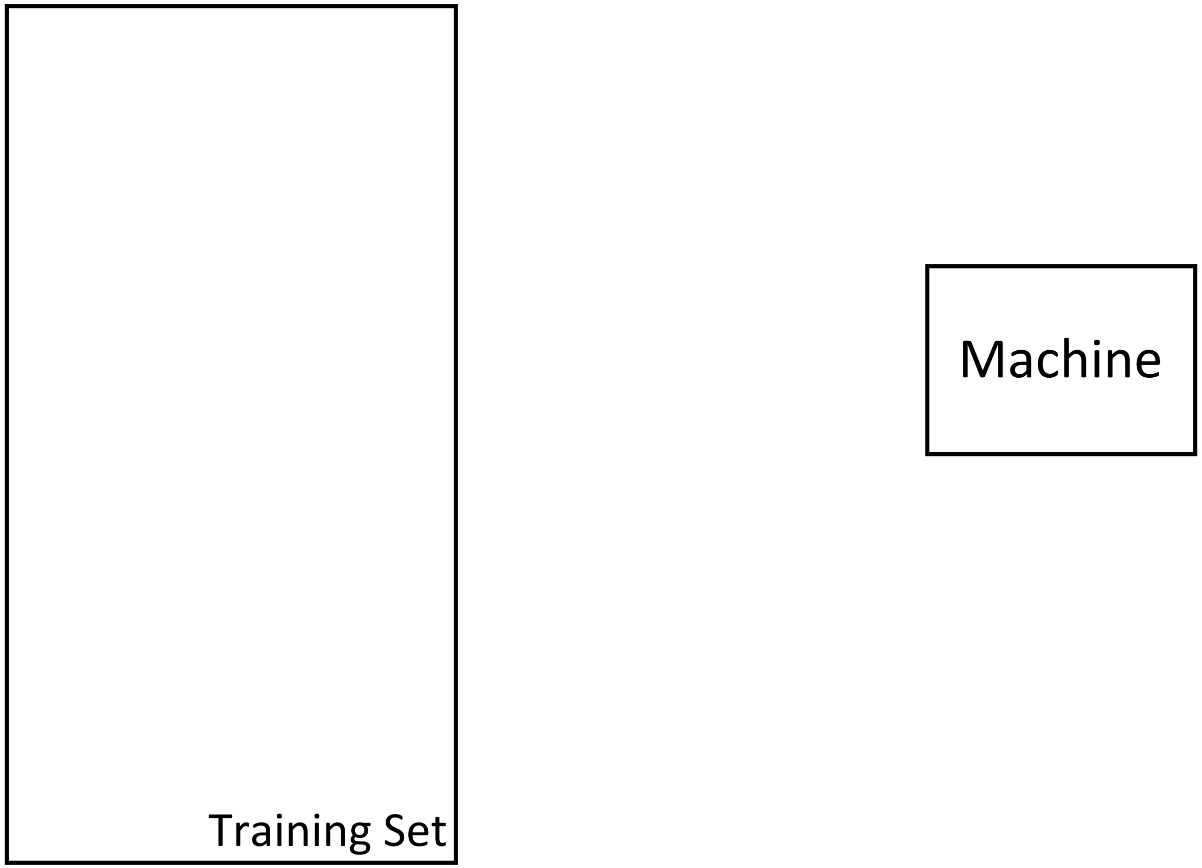
Overview


key pose
learning machine
pose descriptor extraction
kinect
$(x_1, \cdots, x_{15})$
$(\theta_1, \varphi_1, \cdots, \theta_9, \varphi_9, \eta)$
training
set
multi-class
SVM
gesture learning machine
training
set
key pose
gesture
decision forest
buffer
(k_1, t_1)
(k_2, t_2)
(k_n, t_n)

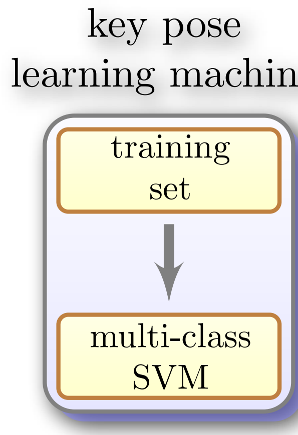


Supervised Learning Machine

Predefined key pose classes: $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$



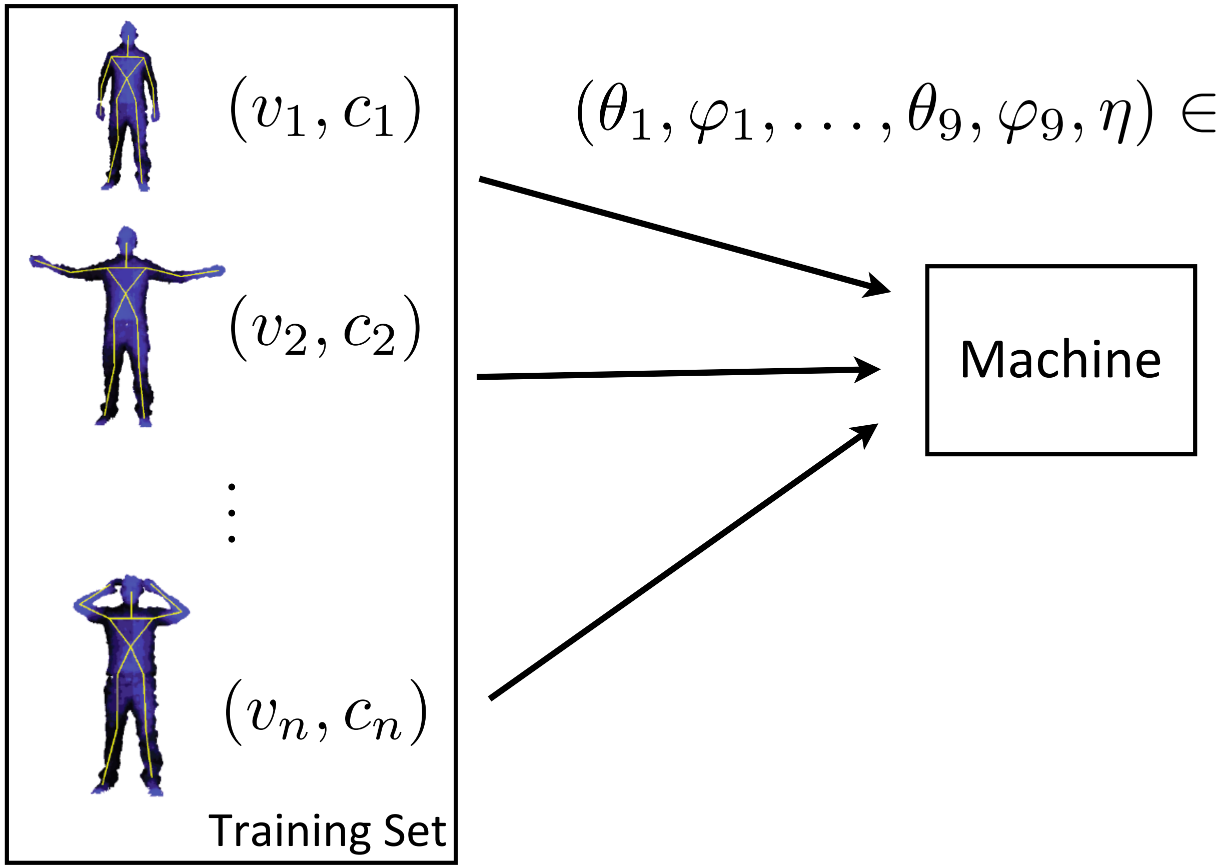
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

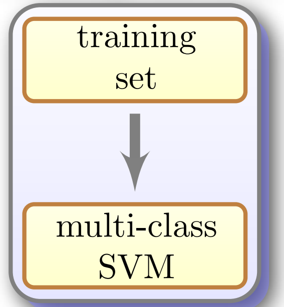


Supervised Learning Machine

Predefined key pose classes: $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$

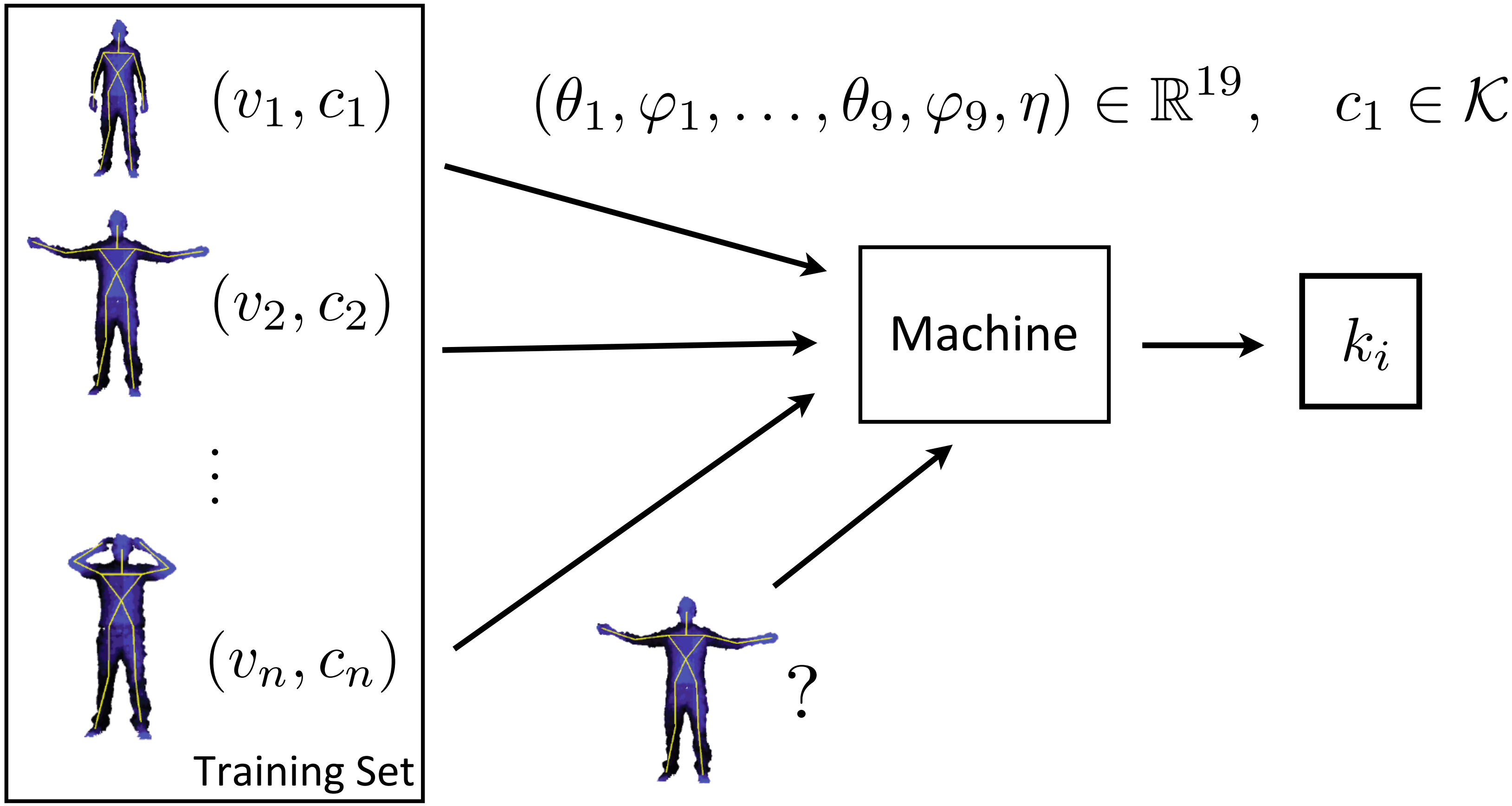
$$(\theta_1, \varphi_1, \ldots, \theta_9, \varphi_9, \eta) \in \mathbb{R}^{19}, \quad c_1 \in \mathcal{K}$$



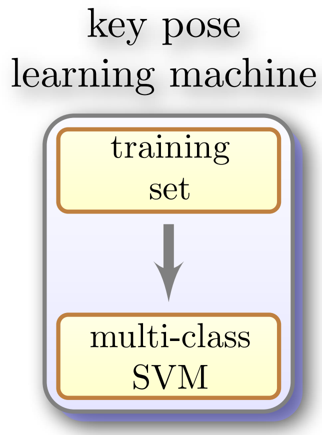


Supervised Learning Machine

Predefined key pose classes: $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$



Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*



Support Vector Machines (SVM)

Binary classifier

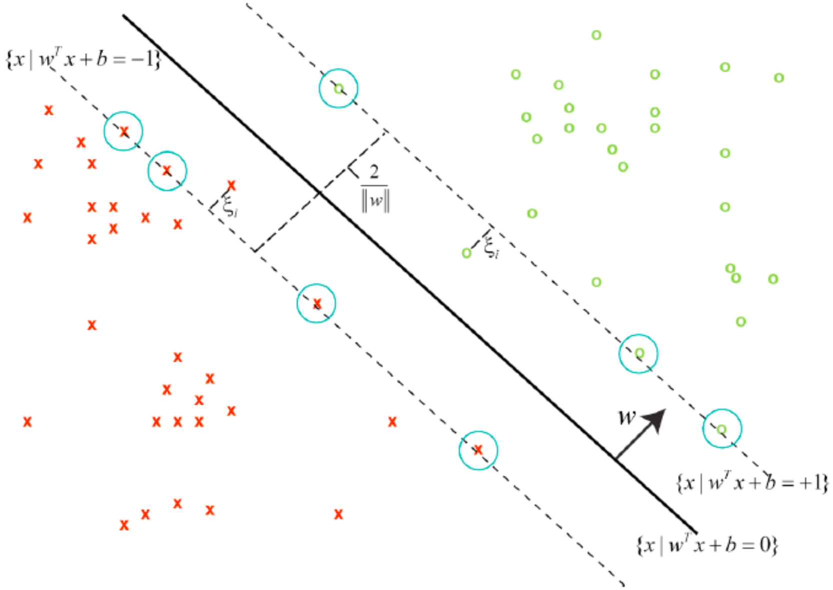
$$\hat{g} : \mathbb{R}^k \rightarrow \{-1, 1\}$$

$$v \rightarrow sign\left(\hat{f}(v)\right) = \{-1, 1\}$$

$$\hat{f}(v) = \sum_j \alpha_j \; s_j \langle \varphi(v_j), \varphi(v) \rangle + b$$

$$\underset{w,\gamma}{\text{MAX}} \quad \gamma - C \sum_{i=1}^{l} \varepsilon_i$$

subject to $\quad y_i \langle w, \Phi(x_i) \rangle \geq \gamma - \varepsilon_i \;,\; \varepsilon_i \geq 0 \;, \quad \|w\|^2 =$ $\quad {}_{i=1,\ldots,l}$

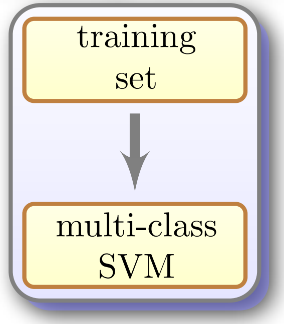


✓ Non-linear classification

✓ Efficiently computed for small training sets

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

key pose
learning machine



Multi-class SVM formulation

One-versus-all approach

One binary classifier for each key pose $\mathbf{p} \in \mathcal{K}$:

$$\hat{f}_{\mathbf{p}}(\mathbf{v}) = \sum_{j \in SV} \alpha_j \psi_{\mathbf{p}}(\mathbf{c}_j)\ \phi(\mathbf{v}_j, \mathbf{v}) + b,$$

where $$\psi_p(\mathbf{c}) = \begin{cases} 1 & \text{if } \mathbf{c} = \mathbf{p}, \\ -1 & \text{otherwise,} \end{cases}$$

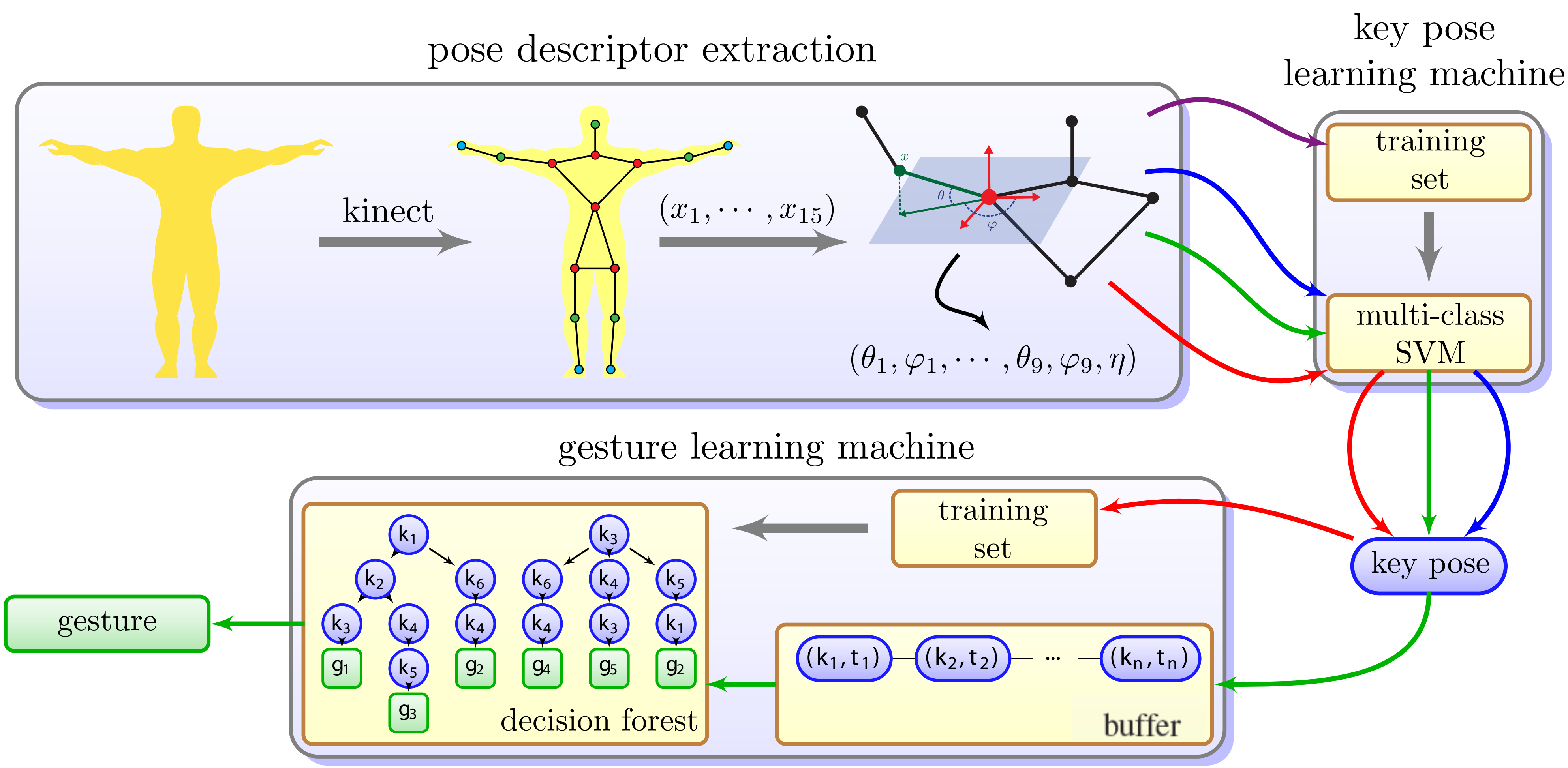
$$\phi(\mathbf{v}_1, \mathbf{v}_2) = \exp\left(-\frac{\|\mathbf{v_2} - \mathbf{v_1}\|^2}{2\sigma^2}\right)$$

Voting process:

$$\hat{f}(\mathbf{v}) = \begin{cases} \mathbf{q} = \arg\max_{\mathbf{p}} \hat{f}_{\mathbf{p}}(\mathbf{v}) & \text{if } \hat{f}_{\mathbf{q}}(\mathbf{v}) > 0, \\ \text{-1} & \text{otherwise.} \end{cases}$$

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

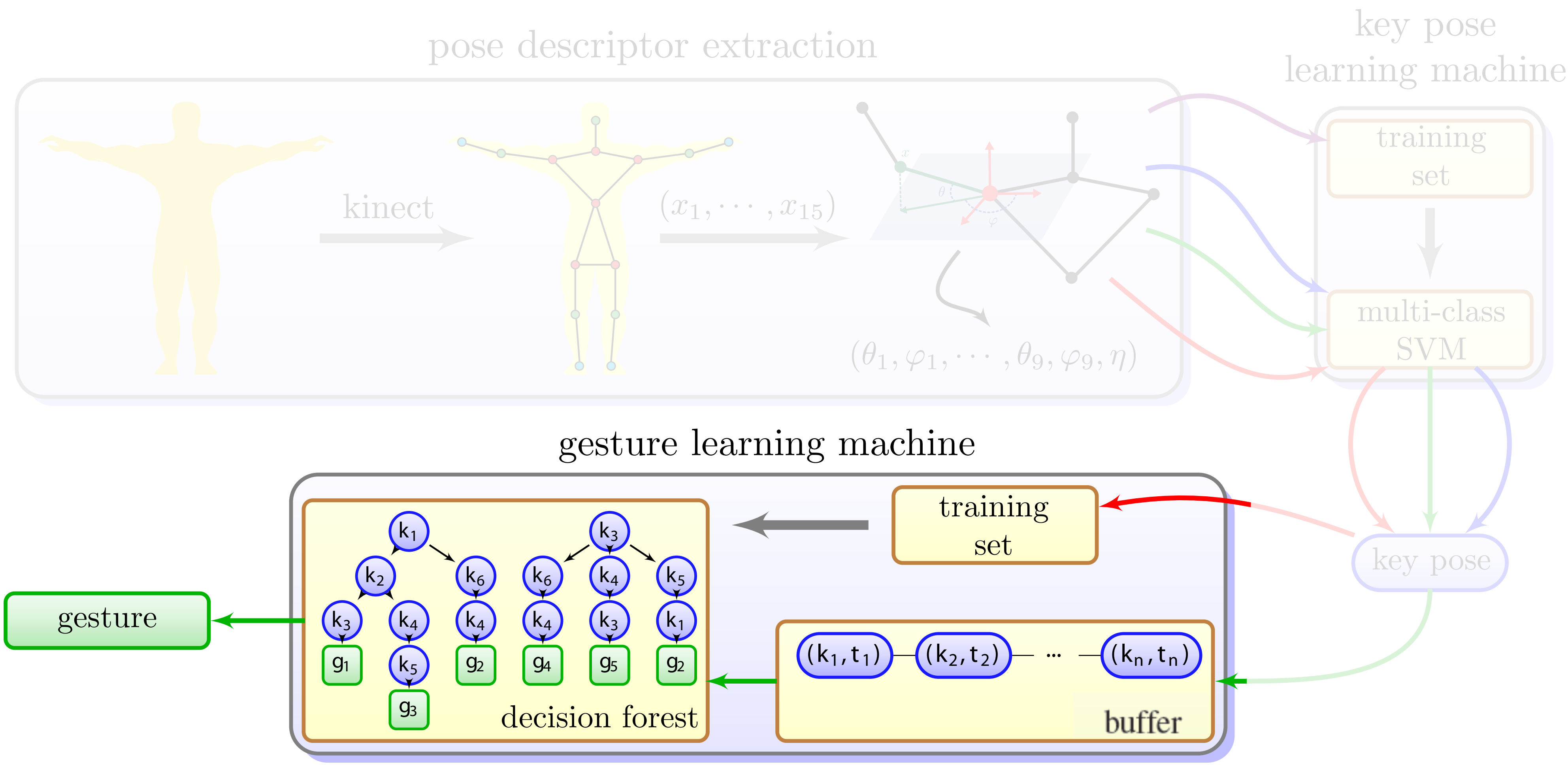
Overview



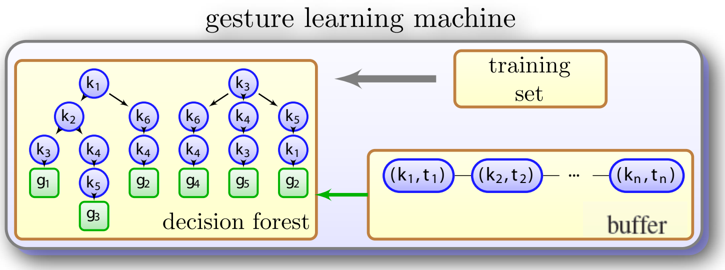
Miranda et al., 2012

Real-time gesture recognition from depth data through key poses learning and decision forests

Overview


pose descriptor extraction
key pose
learning machine
kinect
$(x_1, \cdots, x_{15})$
training
set
multi-class
SVM
$(\theta_1, \varphi_1, \cdots, \theta_9, \varphi_9, \eta)$
gesture learning machine
training
set
key pose
gesture
decision forest
buffer
(k_1, t_1)
(k_2, t_2)
(k_n, t_n)

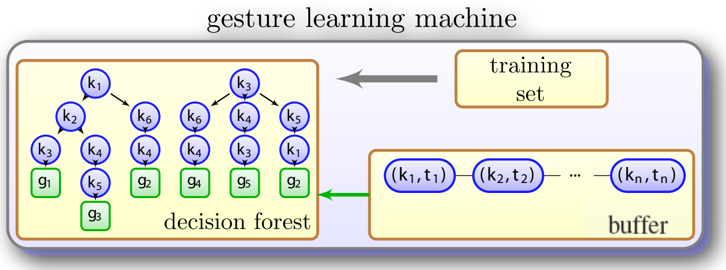
Gestures as key pose sequences



Gesture representation: $g = \{k_1, k_2, \cdots, k_{n_g}\}, \quad k_i \in \mathcal{K}.$

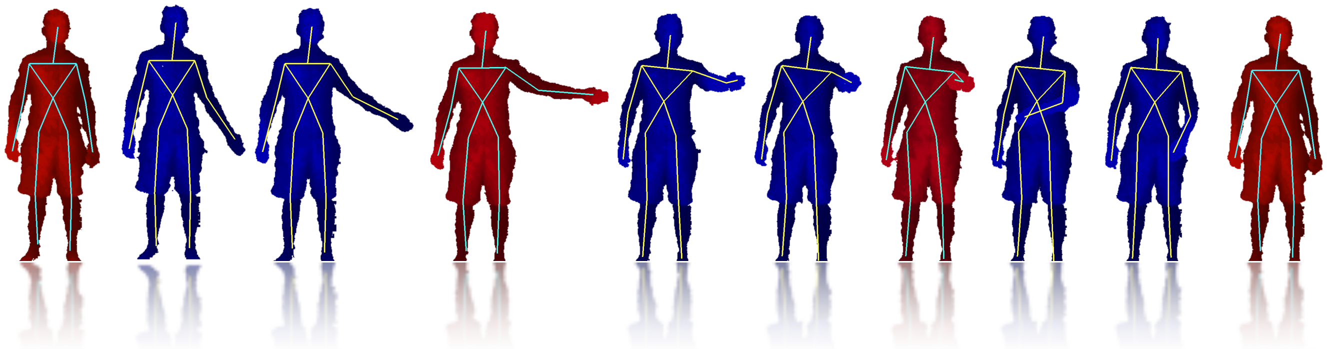
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Gestures as key pose sequences

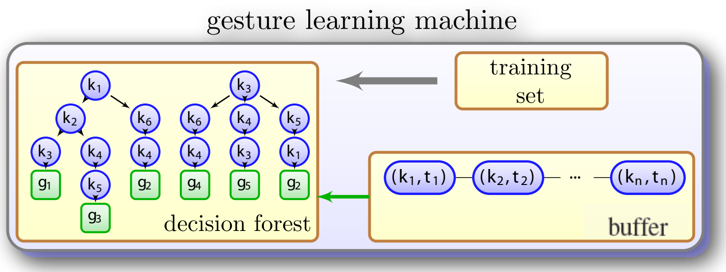


Gesture representation: $g = \{k_1, k_2, \cdots, k_{n_g}\}, \quad k_i \in \mathcal{K}.$

Training session:

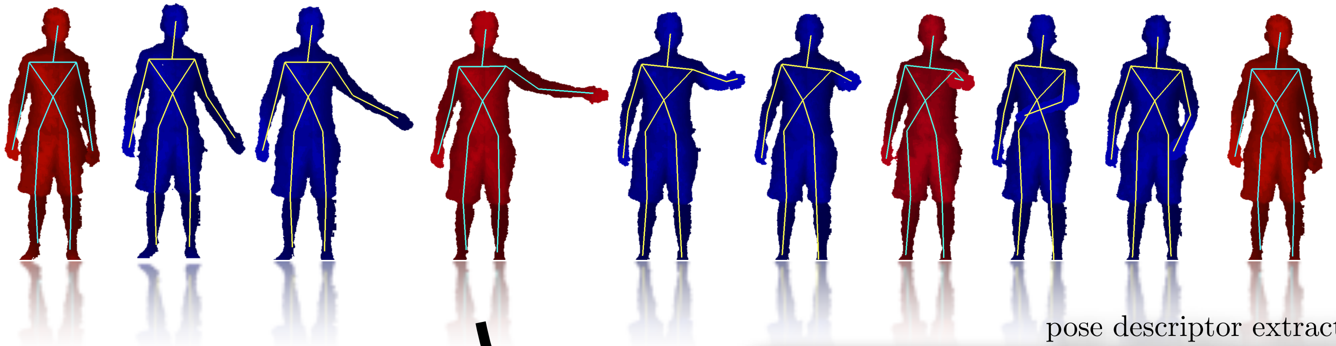


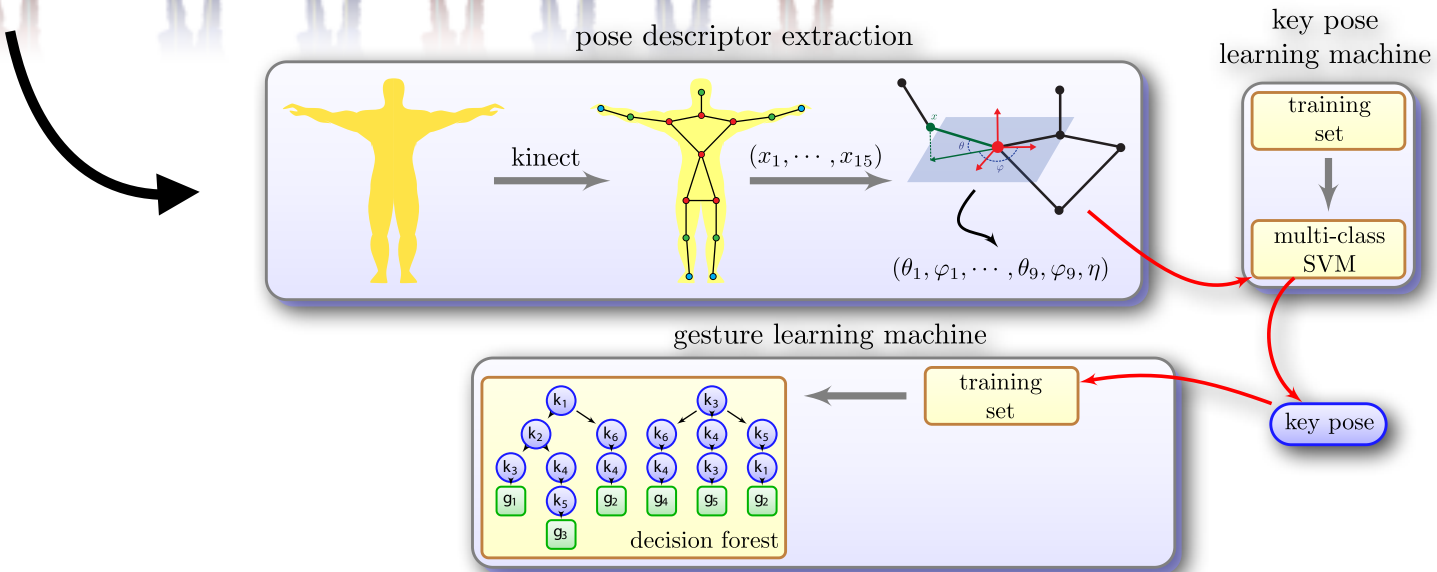
Gestures as key pose sequences



Gesture representation: $g = \{k_1, k_2, \cdots, k_{n_g}\}, \quad k_i \in \mathcal{K}.$

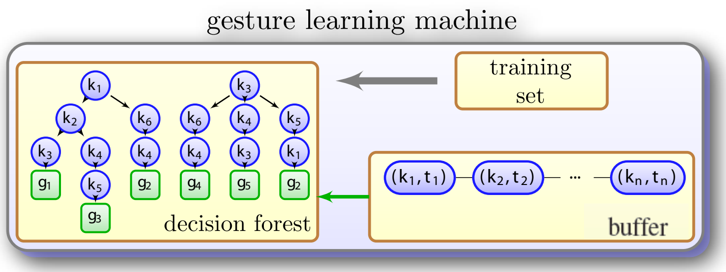
Training session:





Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Decision Forests

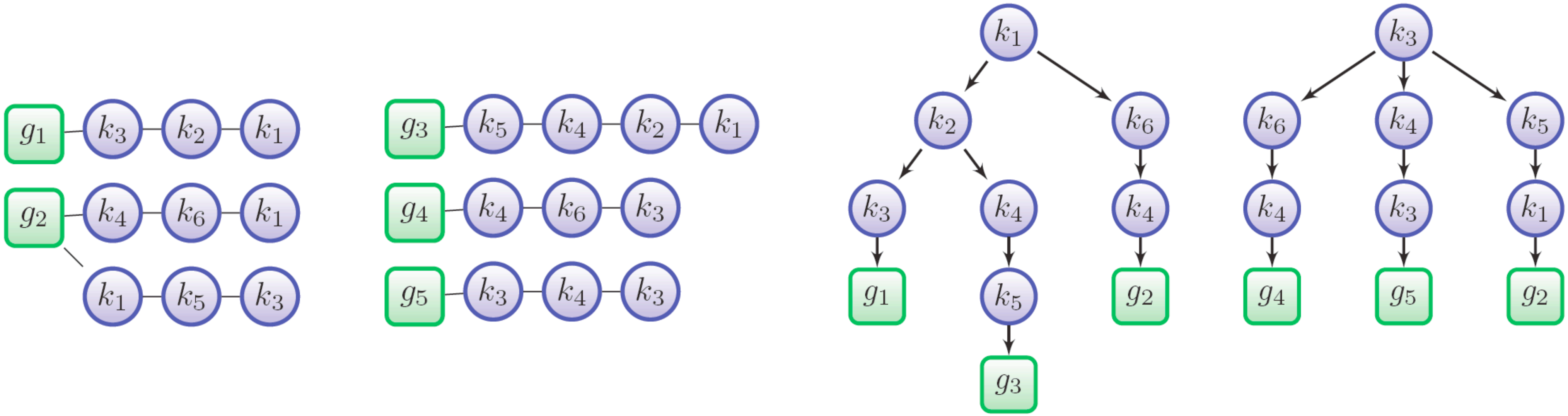


Each node represents a key pose

One tree per key pose

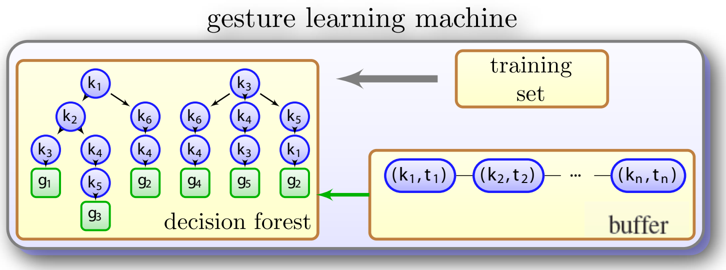
Each root-leaf path represents a gesture stored back-to-front

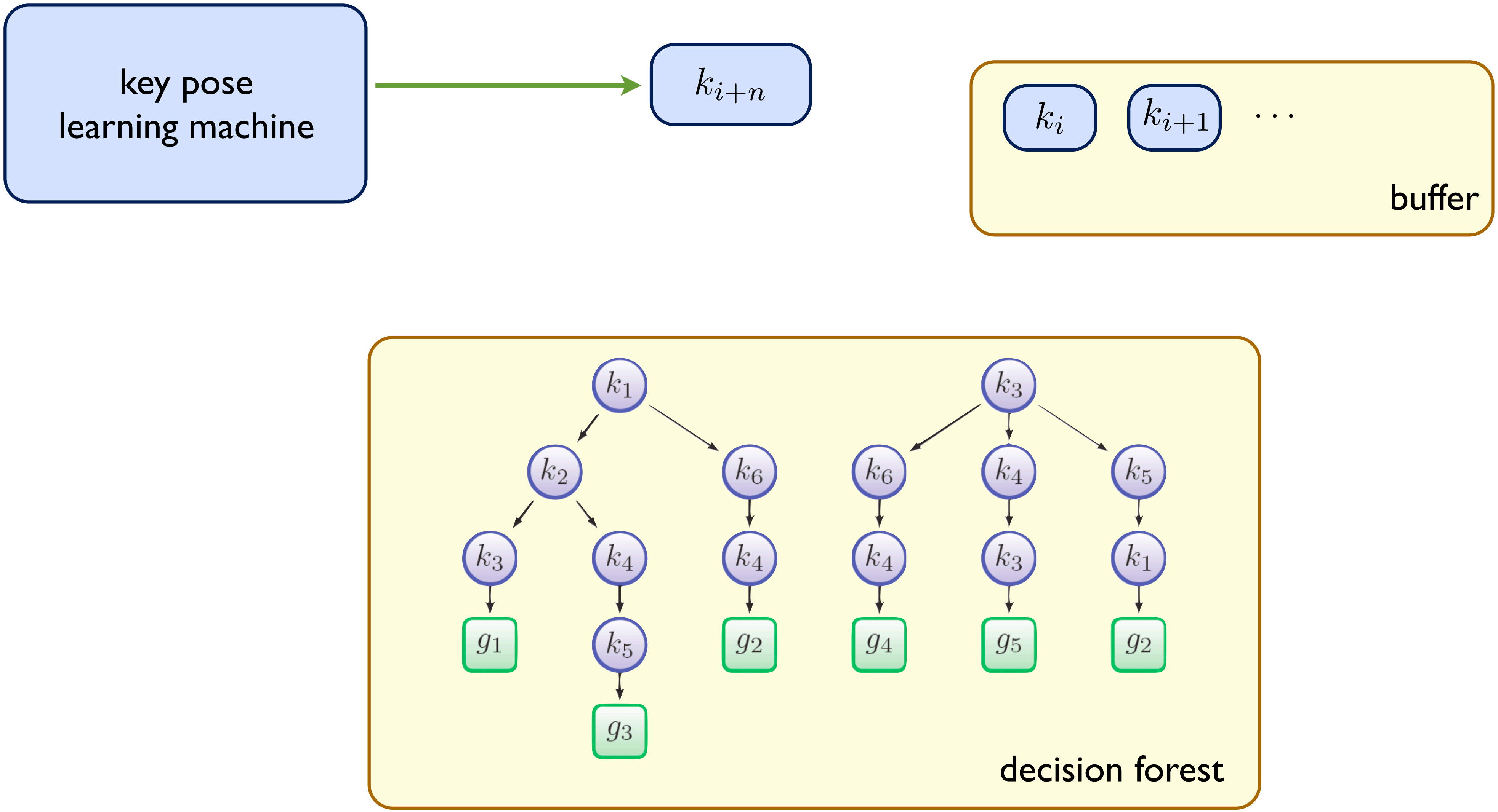
Two paths may represent the same gesture



Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

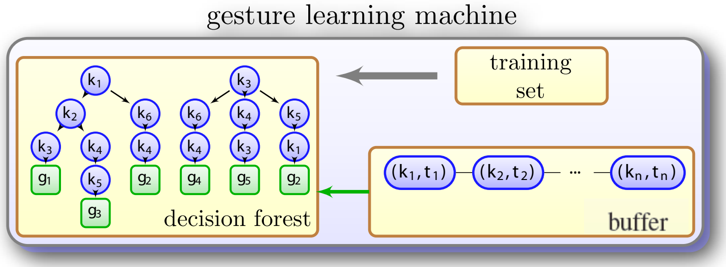
Real-time gesture recognition

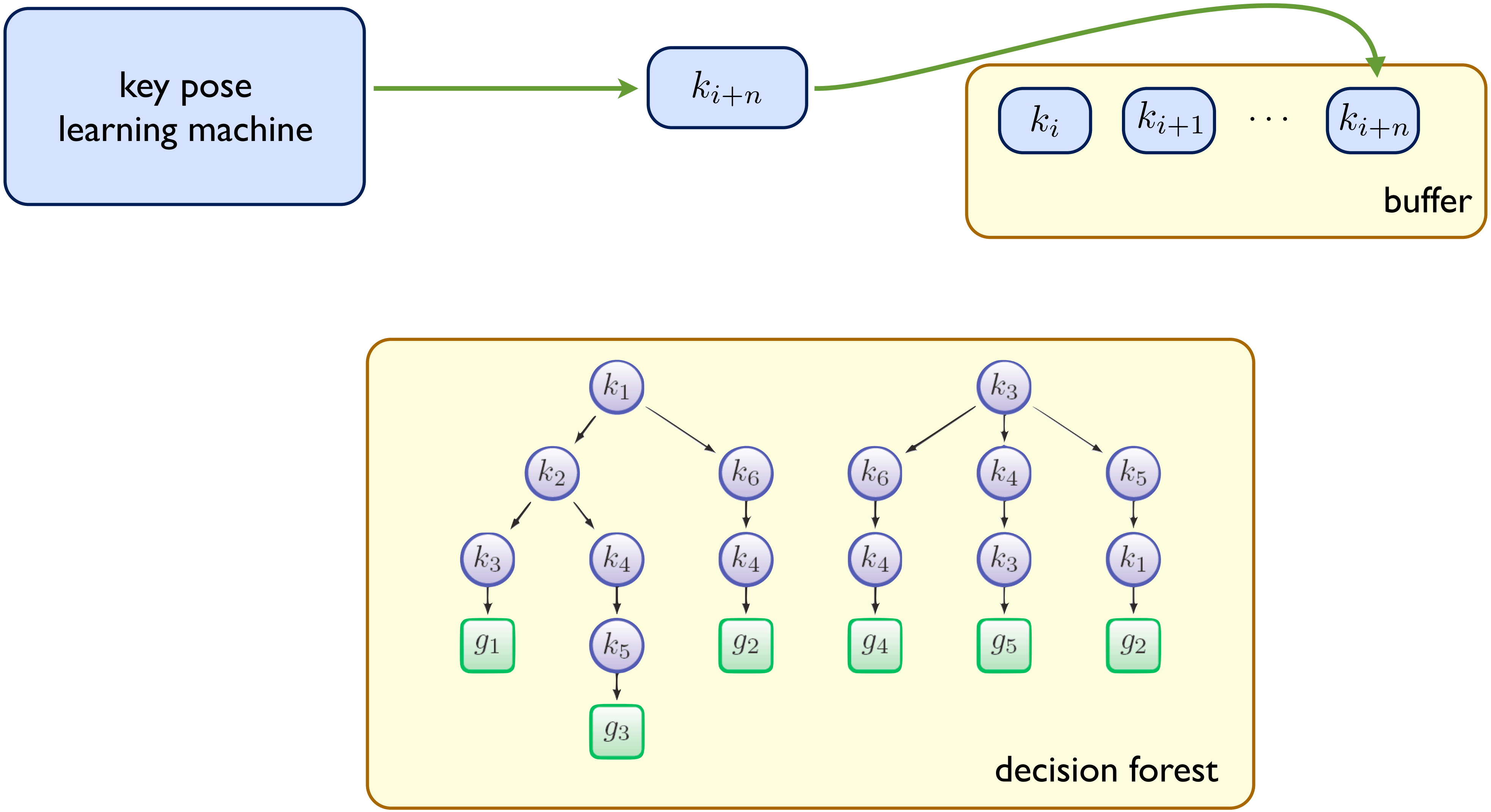




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

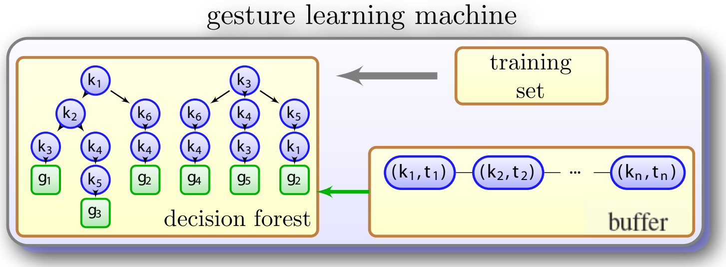
Real-time gesture recognition

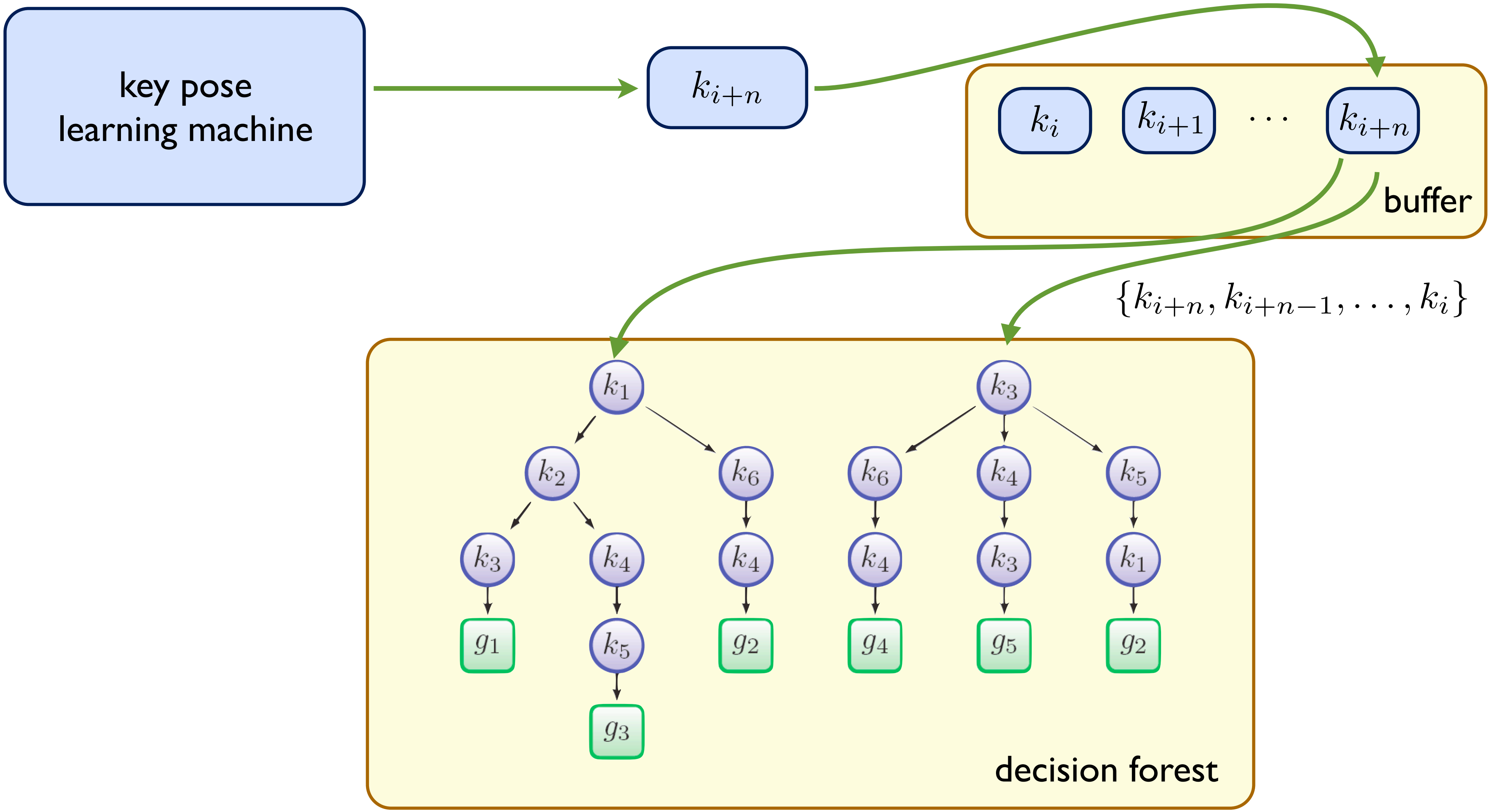




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

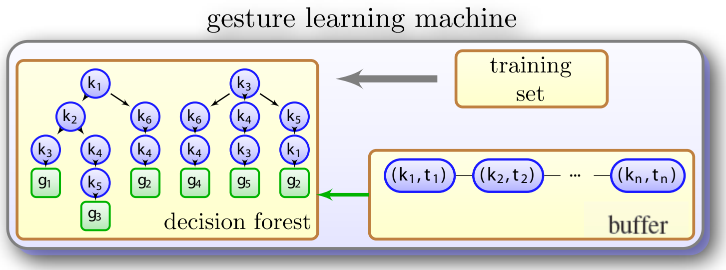
Real-time gesture recognition

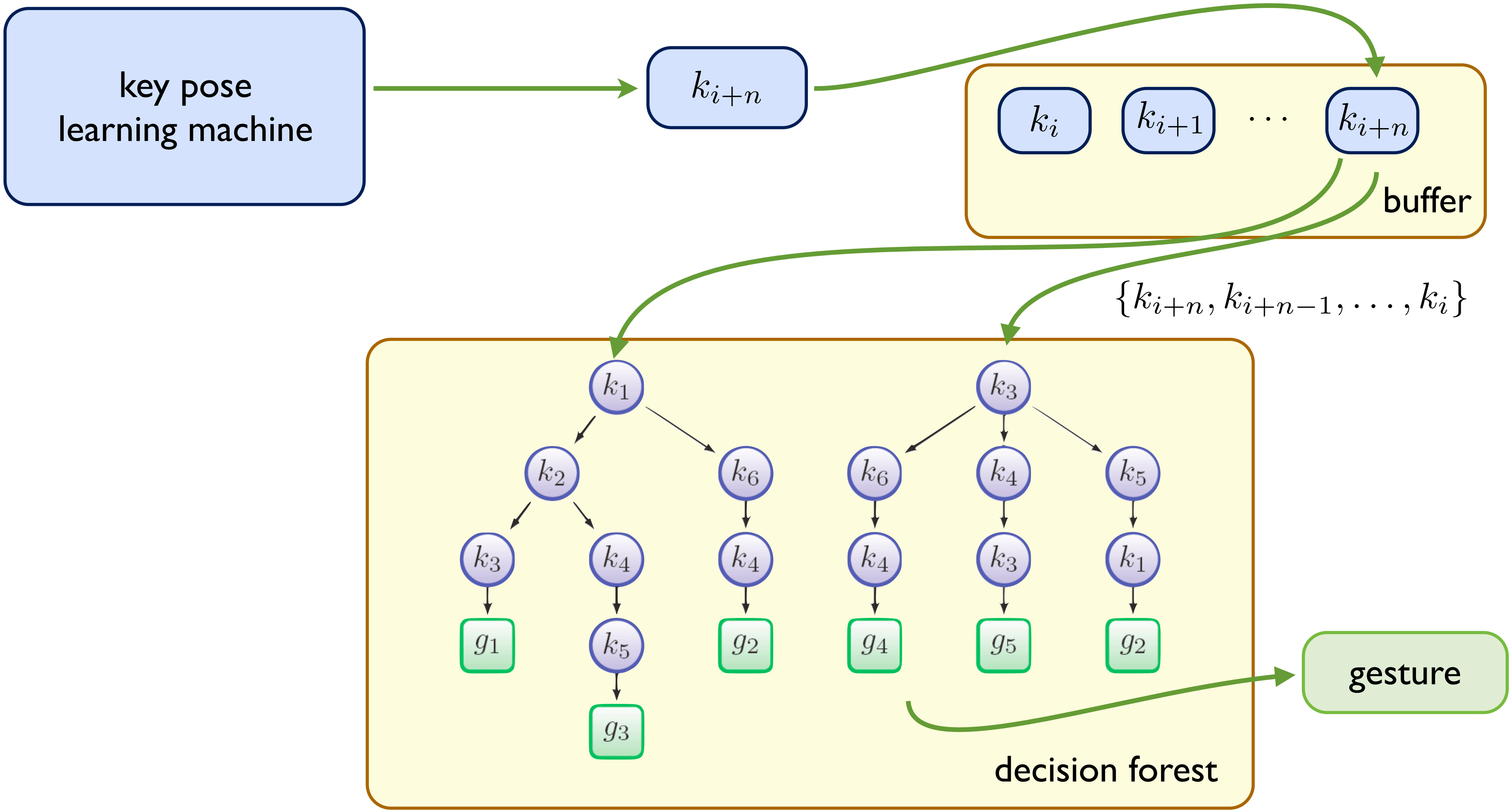




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

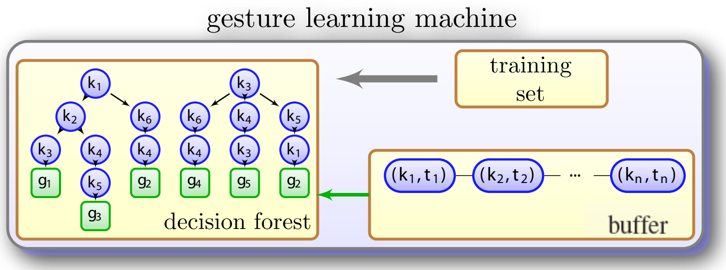
Real-time gesture recognition

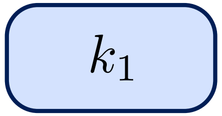


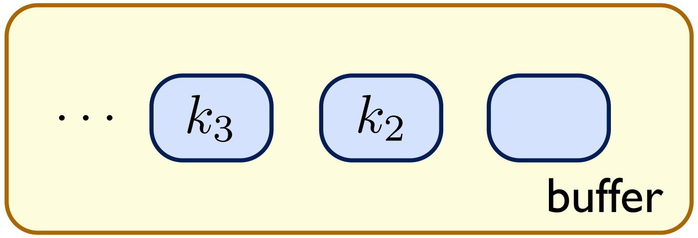


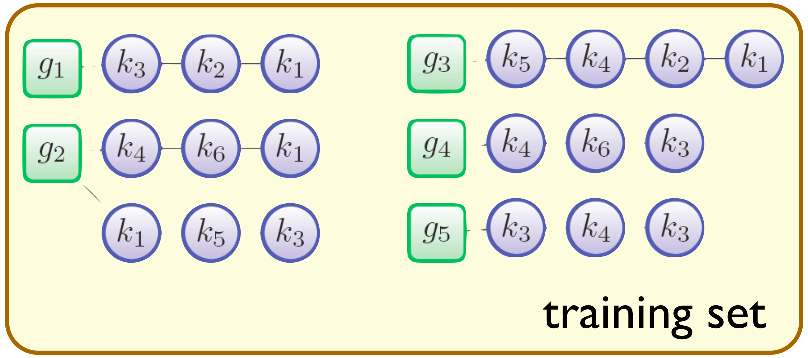
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

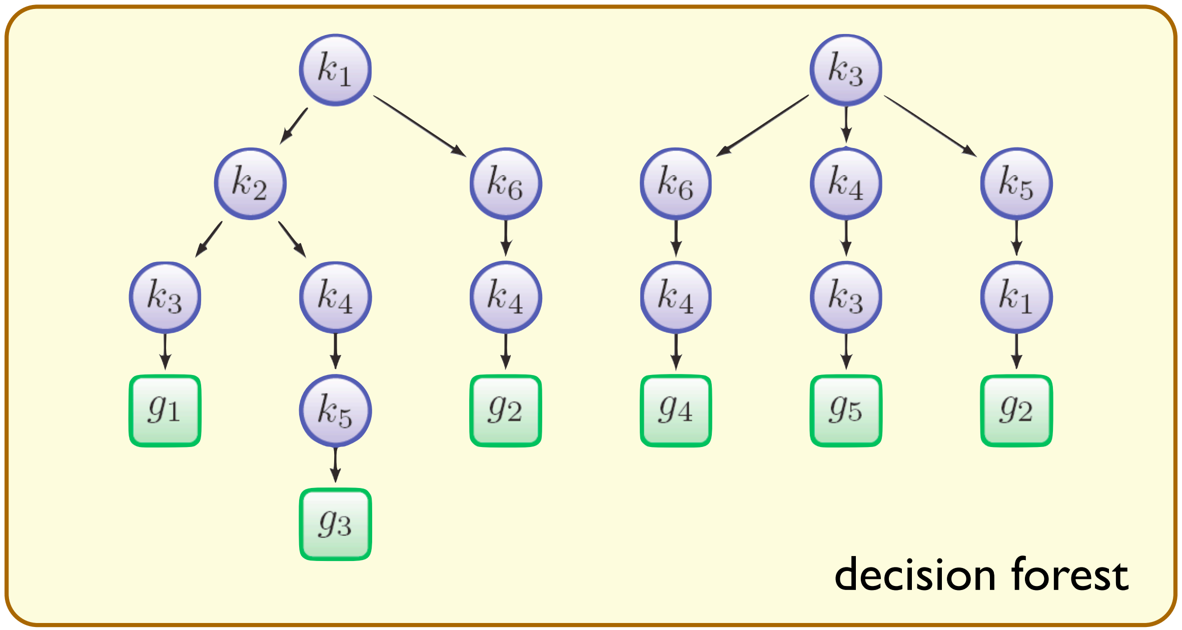
Real-time gesture recognition: Example





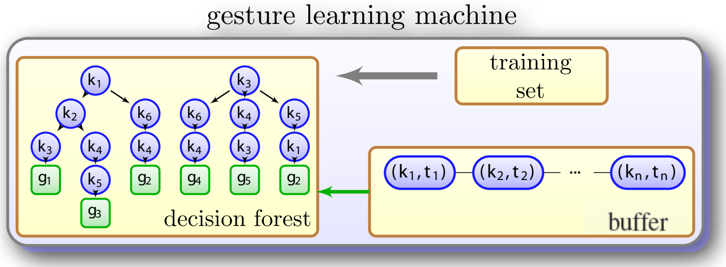


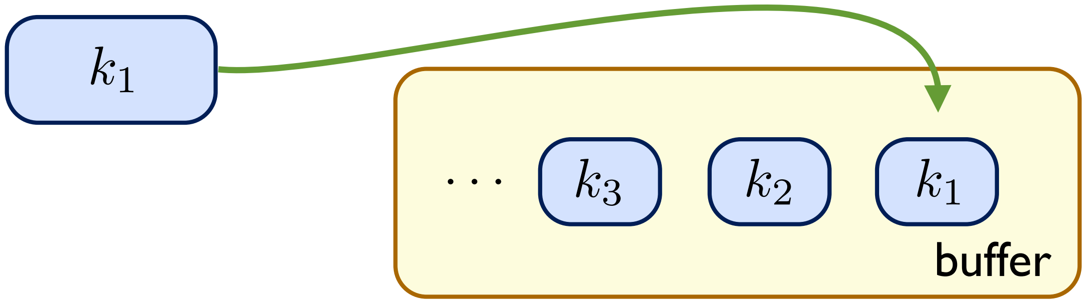


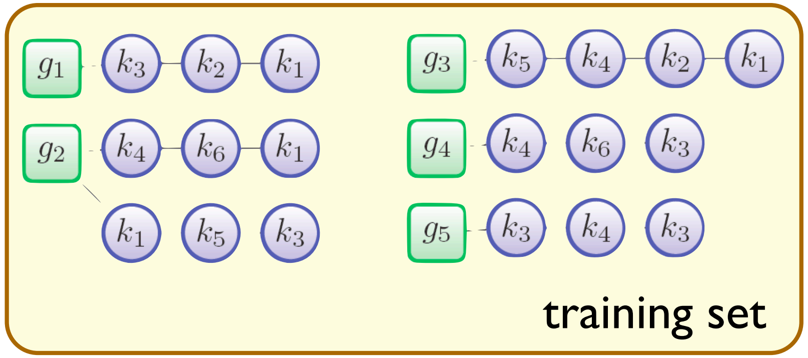


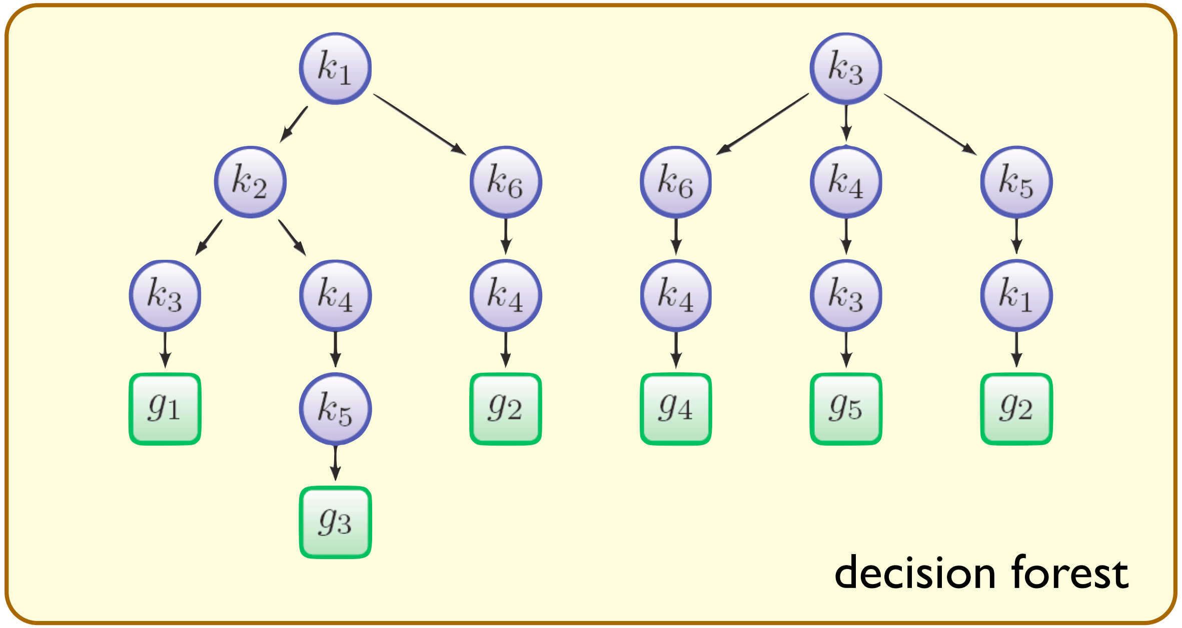
Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Real-time gesture recognition: Example



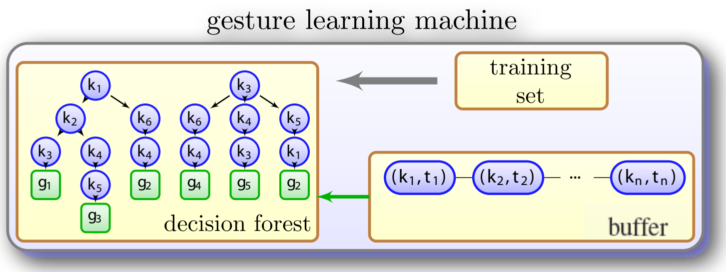


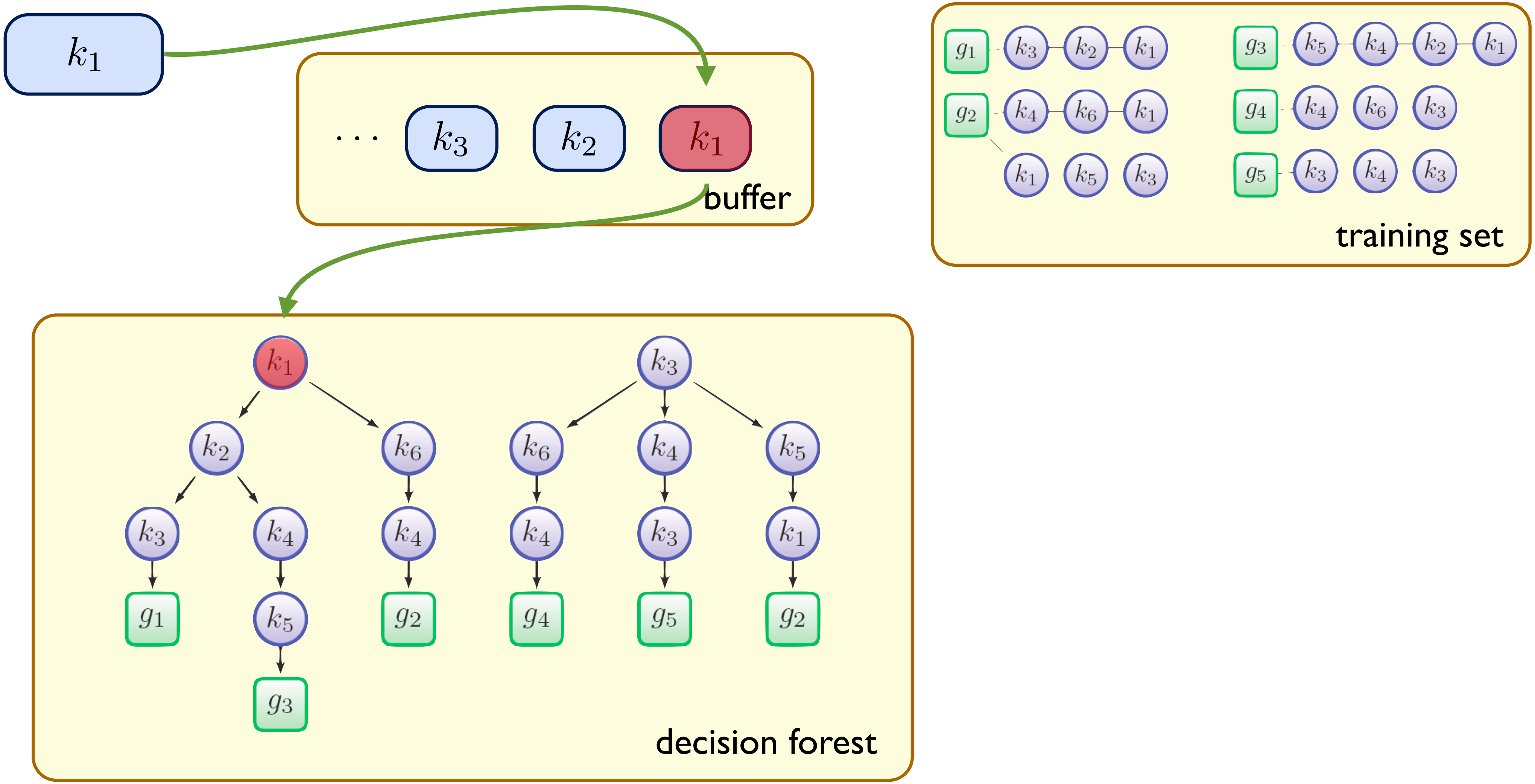




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

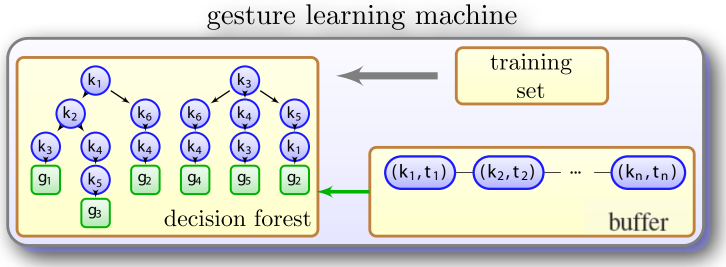
Real-time gesture recognition: Example

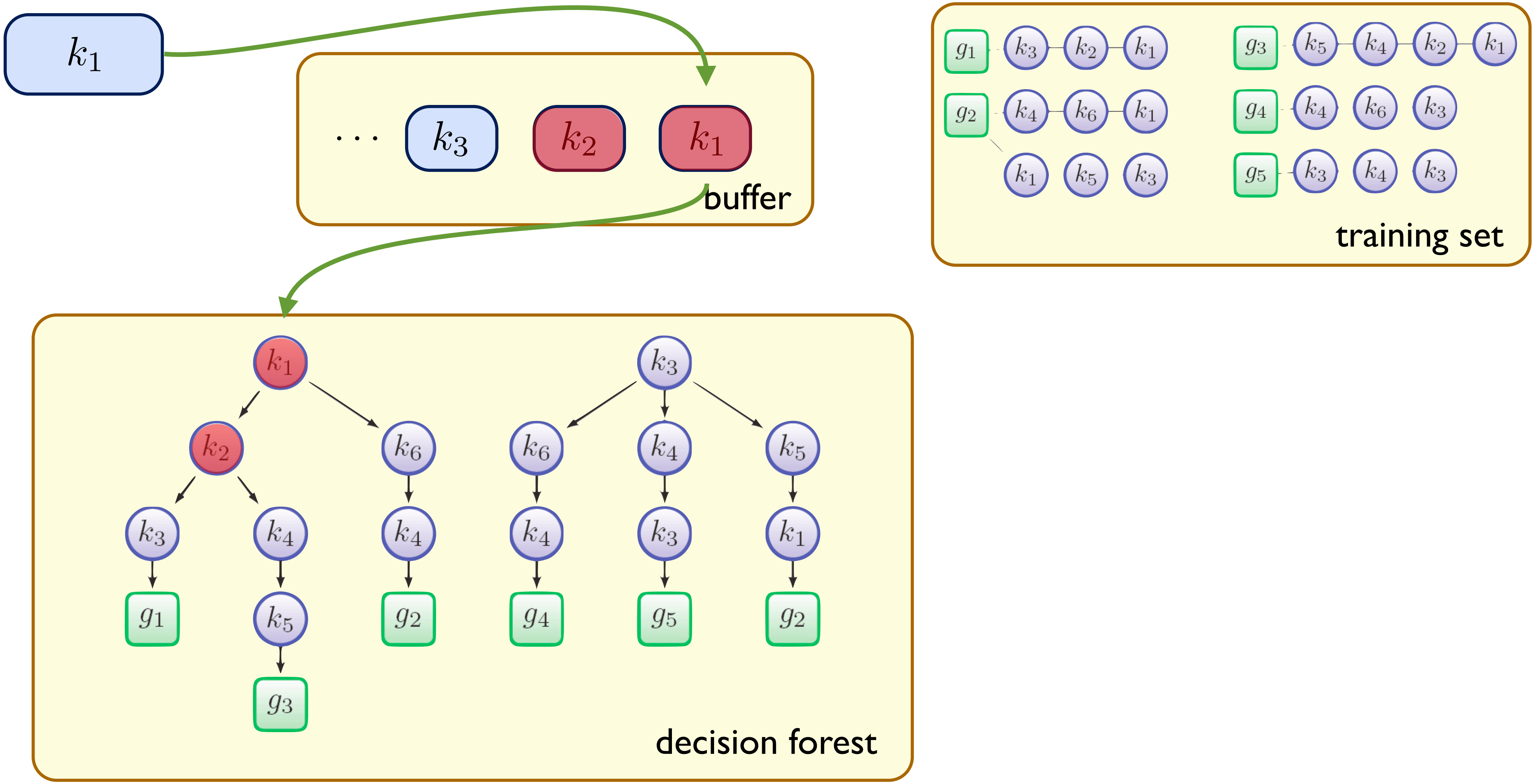




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

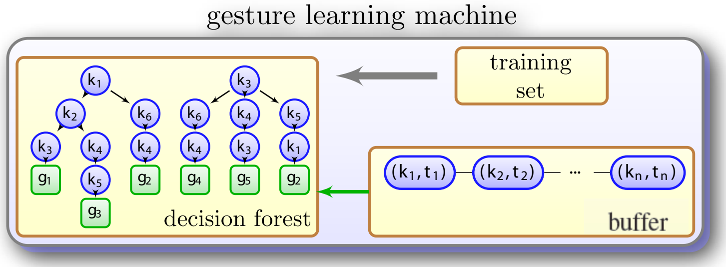
Real-time gesture recognition: Example

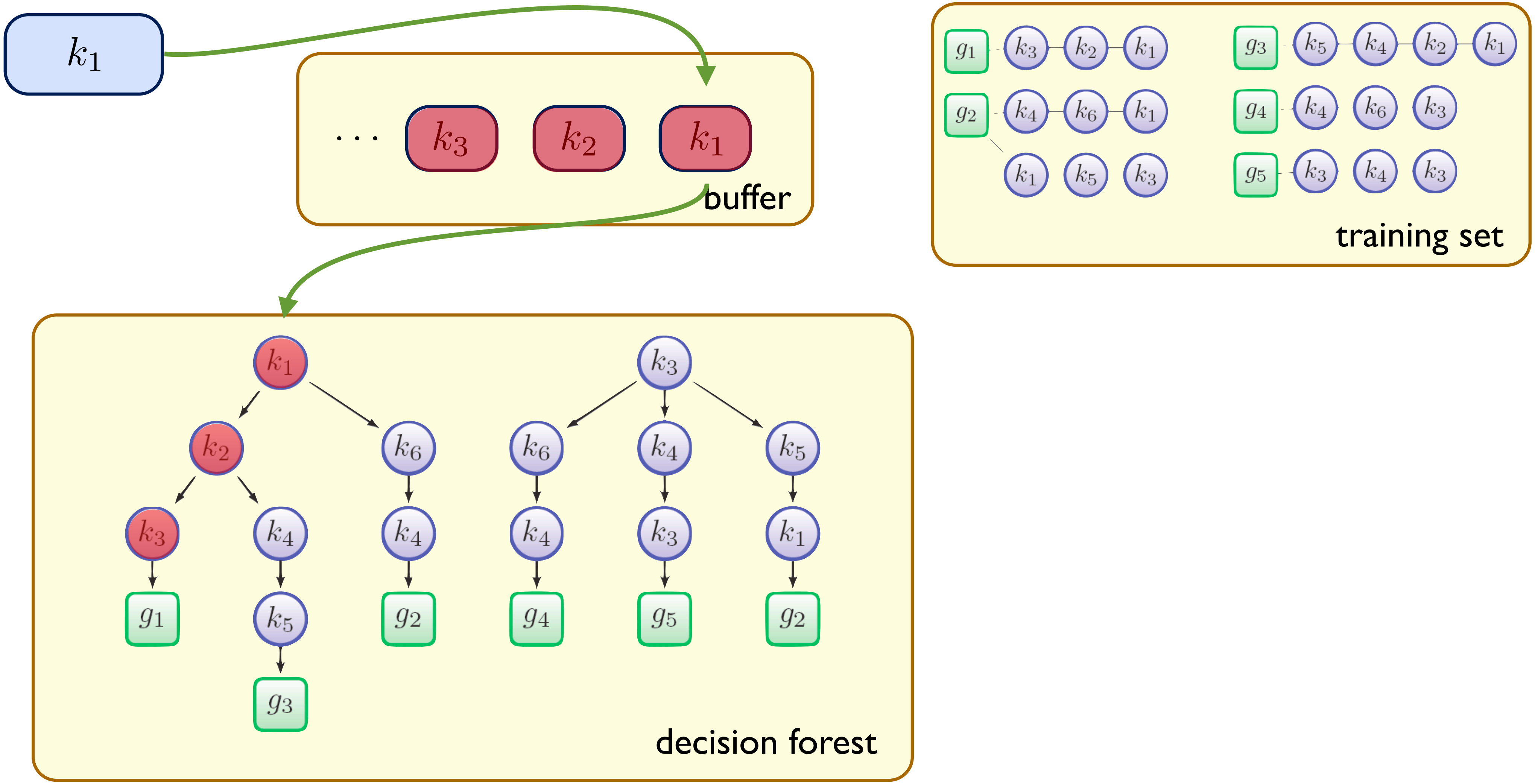




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

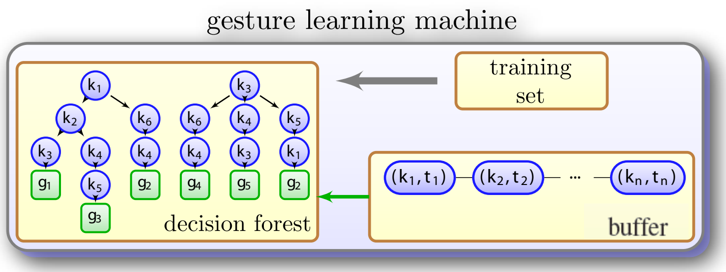
Real-time gesture recognition: Example

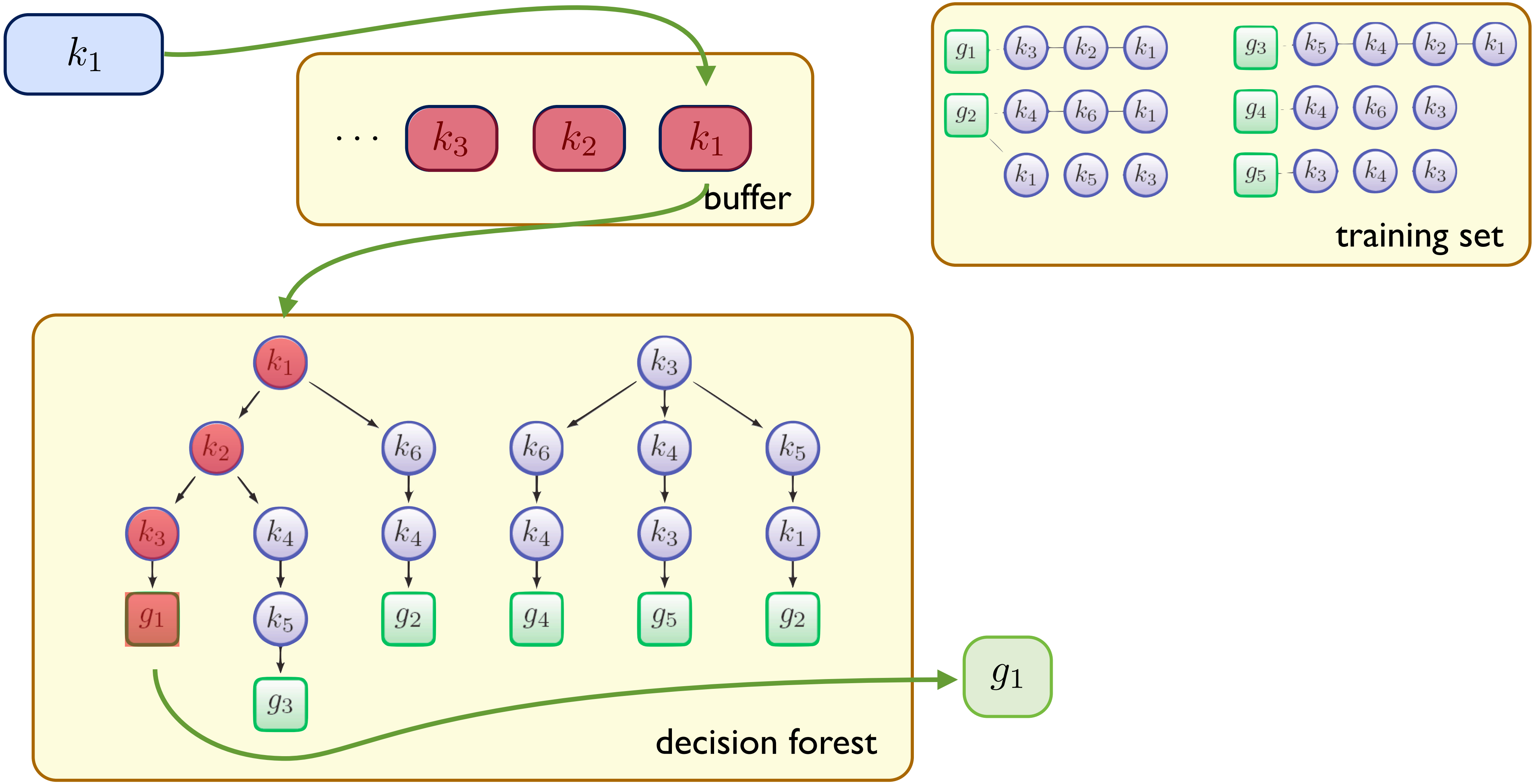




Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

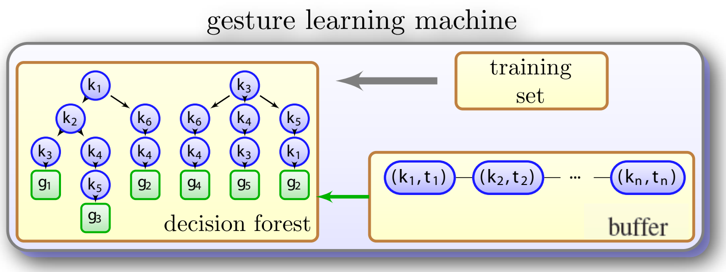
Real-time gesture recognition: Example



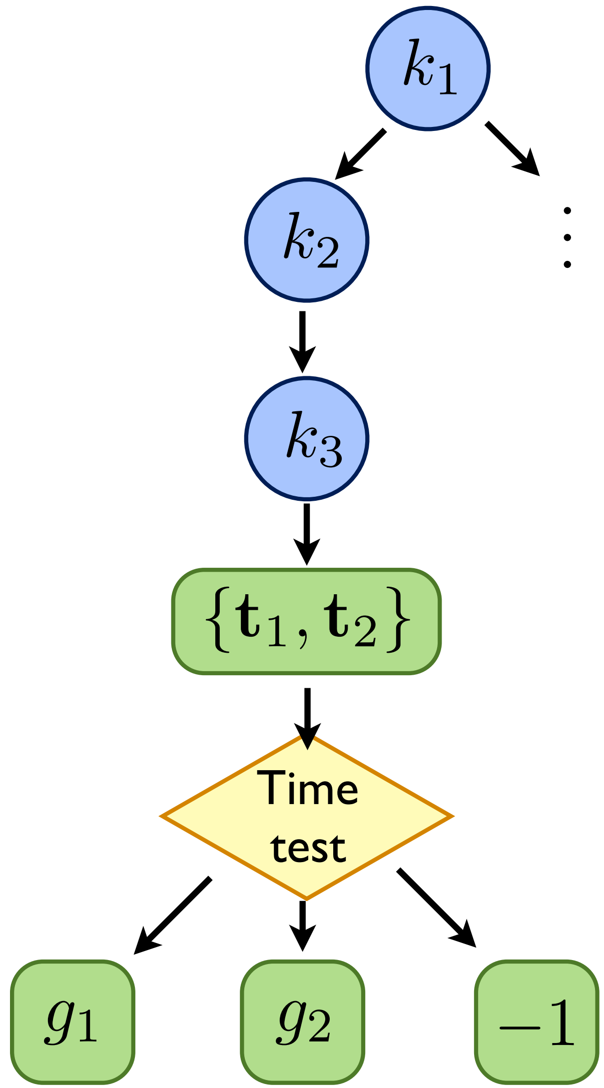


Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Time constraints



Time vector: interval $\mathbf{t} = [t_1, t_2, \cdots, t_{n-1}]$ between consecutive key poses



Time test

> for each time vector t_i found on the leaf
>
> if $\|\mathbf{t}_i - \mathbf{t}\|_\infty > T$
>
> discard t_i
>
> return g_i that minimizes $\|\mathbf{t}_i - \mathbf{t}\|_1$

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

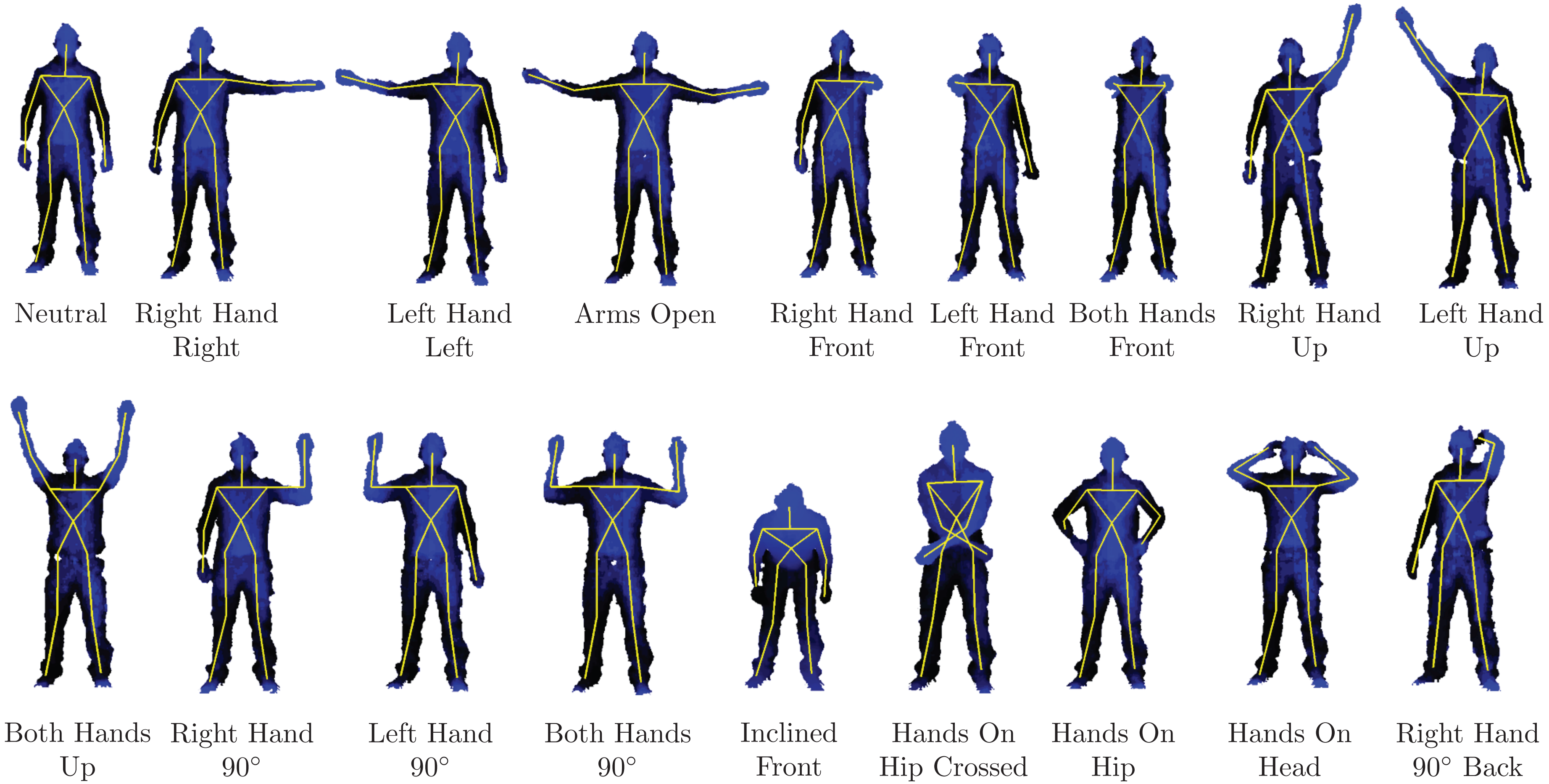
Results

Experiment Setup

One trainer

18 trained key poses (approx. 30 examples per key pose)

10 trained gestures (approx. 10 executions per gesture)



Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Key pose recognition: robustness

10 inexperienced individuals performed trained key poses 10 times

key pose	id	recognized key poses per user											total
		u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}^1	u_{10}^2	(%)
Neutral	k_1	10	10	10	10	10	10	10	10	10	10	10	**100.00**
Right Hand Right	k_2	10	10	10	10	10	10	10	10	10	10	8	**98.18**
Left Hand Left	k_3	10	10	10	9	10	10	9	10	10	10	10	**98.18**
Arms Open	k_4	10	10	10	7	10	10	10	9	10	7	10	**93.63**
Right Hand Front	k_5	10	10	10	10	10	10	10	10	10	8	7	**95.45**
Left Hand Front	k_6	10	10	9	10	10	10	10	10	10	10	10	**99.09**
Both Hands Front	k_7	10	10	10	10	10	10	10	10	10	10	10	**100.00**
Right Hand Up	k_8	10	10	10	10	10	10	10	10	10	10	10	**100.00**
Left Hand Up	k_9	10	10	10	10	10	9	10	10	10	9	10	**98.18**
Both Hands Up	k_{10}	10	10	10	10	10	10	10	10	10	10	10	**100.00**
Right Hand 90°	k_{11}	10	8	9	10	10	10	10	10	8	10	10	**95.45**
Left Hand 90°	k_{12}	10	10	10	10	10	6	10	10	10	5	10	**91.81**
Both Hands 90°	k_{13}	10	10	10	10	10	10	10	10	10	10	10	**100.00**
Inclined Front	k_{14}	8	10	10	10	10	8	10	10	10	5	7	**89.09**
Hands-on-Hip Crossed	k_{15}	7	8	6	8	8	10	10	10	8	10	8	**84.54**
Hand-On-Hip	k_{16}	10	10	10	10	10	10	10	9	10	10	10	**99.09**
Hands on Head	k_{17}	9	10	10	8	10	10	9	7	10	10	6	**90.00**
Right Hand 90° Back	k_{18}	8	10	9	6	7	7	7	10	10	3	8	**77.27**
total (%)		**95.5**	**97.7**	**96.1**	**93.3**	**97.2**	**94.4**	**97.2**	**97.2**	**97.7**	**87.2**	**91.11**	

Average recognition rate: 94.84%

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Key pose recognition: stability

Out-of-sample tests:

1. Remove 20% of training set data;
2. Compute SVM classifier;
3. Try to classify removed training data.

Results after 10 experiments:

False classifications: <u>4.16%</u>

Unclassified key poses: <u>3.45%</u>

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Key pose recognition

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Gesture recognition

10 inexperienced individuals performed trained gestures 10 times

gesture	id	key pose seq.	rec. rate
Open-Clap	g_1	$k_1,\ k_4,\ k_7$	**99%**
Open Arms	g_2	$k_1,\ k_7,\ k_4$	**96%**
Turn Next Page	g_3	$k_1,\ k_2,\ k_5,\ k_1$ $k_1,\ k_6,\ k_3,\ k_1$	**83%**
Turn Previous Page	g_4	$k_1,\ k_5,\ k_2,\ k_1$ $k_1,\ k_3,\ k_6,\ k_1$	**91%**
Raise Right Arm Laterally	g_5	$k_1,\ k_2,\ k_8$	**80%**
Lower Right Arm Laterally	g_6	$k_8,\ k_2,\ k_1$	**78%**
Good Bye (k_{11} time constraint: 1sec.)	g_7	$k_1,\ k_{11}$	**92%**
Japanese Greeting	g_8	$k_1,\ k_{14},\ k_1$	**100%**
Put Hands Up Front	g_9	$k_1,\ k_5,\ k_{18}$ $k_1,\ k_5,\ k_8$ $k_1,\ k_5,\ k_{11},\ k_8$ $k_1,\ k_8$	**96%**
Put Hands Up Laterally	g_{10}	$k_1,\ k_4,\ k_{10}$	**100%**

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Gesture recognition

Miranda et al., 2012

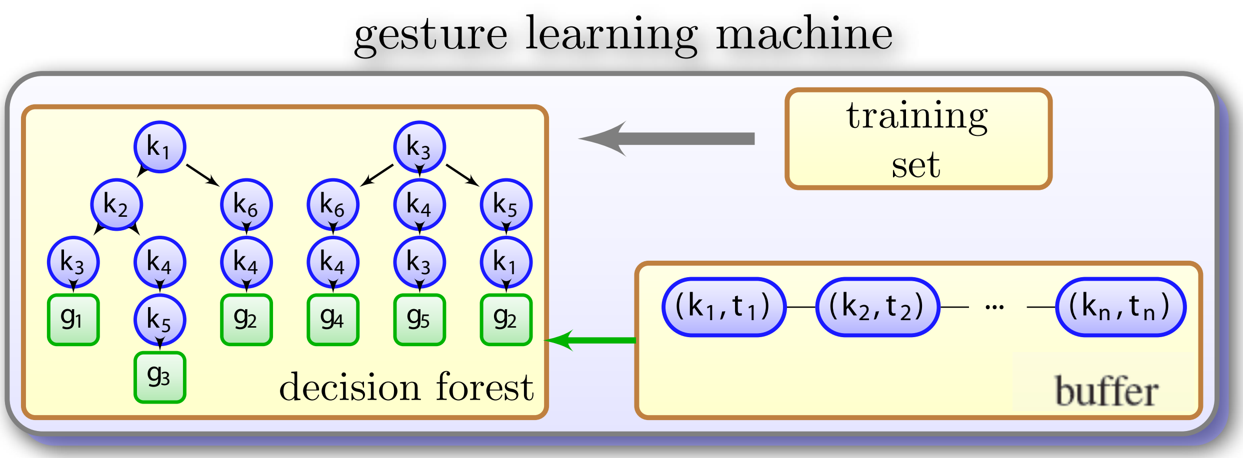
Real-time gesture recognition from depth data through key poses learning and decision forests

Performance

Preprocessing bottleneck: computing SVM classifiers

For a training set of 2,000 key pose examples of 18 classes:
18 functions were computed in 3.9 secs

<u>Negligible performance during training/recognition phases</u>



Usually very low tree depths

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Comparison

Dataset from Li *et al* (2010): 20 gestures, 10 individuals, 3 executions

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Pickup & throw	Side boxing	Pickup & throw

Cross-subject test:

Gesture subset	Li [10]	Vieira [15]	our method
AS1	72.9%	84.7%	**93.5%**
AS2	71.9%	81.3%	**52.0%**
AS3	79.2%	88.4%	**95.4%**
Average	74.7%	84.8%	**80.3%**

Comparison

Dataset from Li *et al* (2010): 20 gestures, 10 individuals, 3 executions

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Pickup & throw	Side boxing	Pickup & throw

Cross-subject test:

Gesture subset	Li [10]	Vieira [15]	our method
AS1	72.9%	84.7%	**93.5%**
AS2	71.9%	81.3%	**52.0%**
AS3	79.2%	88.4%	**95.4%**
Average	74.7%	84.8%	**80.3%**

Delicate gestures

Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

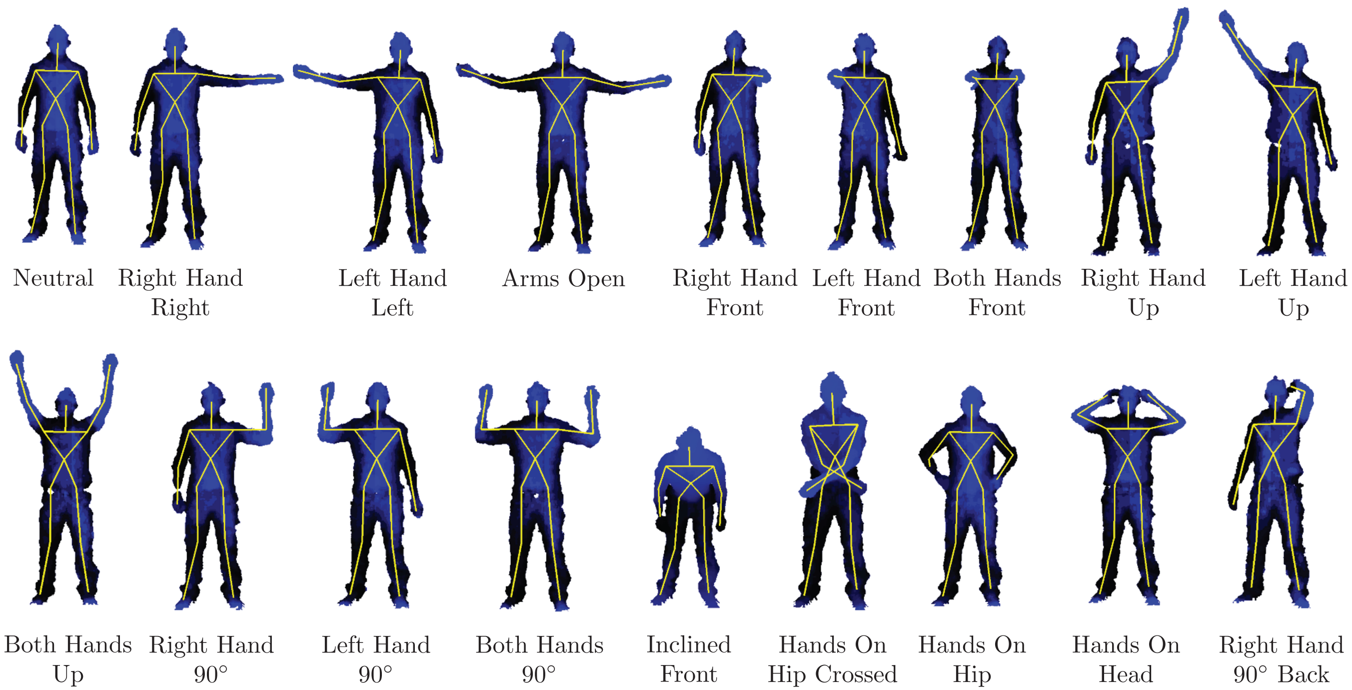
Limitations

- Robustness issues

  Skeleton tracking

  Delicate gestures

- Key pose design not the friendliest solution



Miranda et al., 2012 *Real-time gesture recognition from depth data through key poses learning and decision forests*

Future Work

- ✓ Automatic key pose generation
- ✓ Work on skeleton tracking algorithms (More than 1 Kinect?)
- ✓ Improve time constrained gesture recognition
- ✓ Take into account key pose descriptor periodicity

Thank you for your attention!

Thank you for your attention!

Thank you for your attention!

Questions?