EVOLUTIONARY BAGS OF SPACE-TIME FEATURES FOR HUMAN ANALYSIS

Víctor Ponce López

PhD candidate in Mathematics and Computer Science

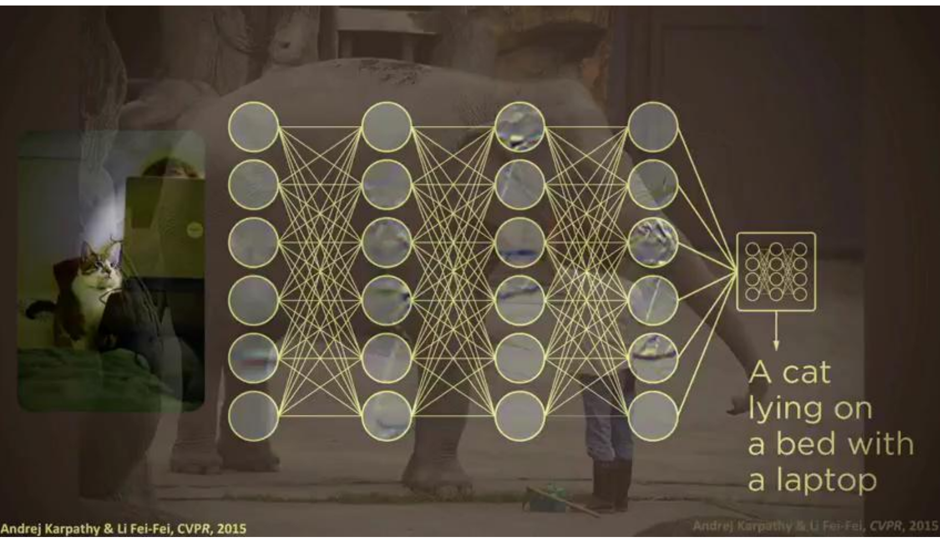
Advisors
Dr. Sergio Escalera Guerrero
Dr. Xavier Baró Solé
Dr. Hugo Jair Escalante

Faculty of Mathematics

Barcelona, June 2^{md}, 2016

Motivation

Humans are experts on recognizing objects and events in the world.



We have taught machines how to perform as closest human-like learning as we know at the moment.

Motivation

Abstraction to human behavioural cues
from multimodal data description.

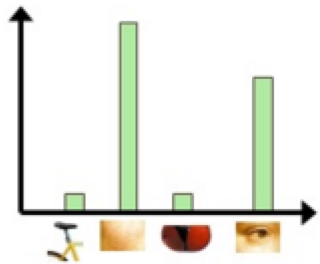
Hypothesis: Language is misconstrued if it is not seen as a unity of speech and gesture.

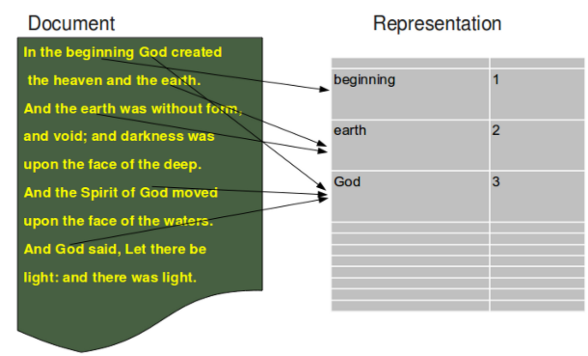
D. McNeill (2012). How Language Began: Gesture and Speech in Human Evolution. Cambridge University Press.

Motivation

The composition of parts is what forms the whole object.

Object — **Bag of 'words'**





http://vision.stanford.edu

http://www.python-course.eu

Motivation

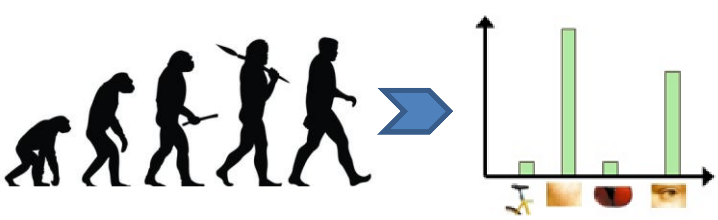
Evolving BoW representations.

- ✓ Biologically inspired.
- ✓ Highly domain-adaptive.
- ✓ Parallel processing.

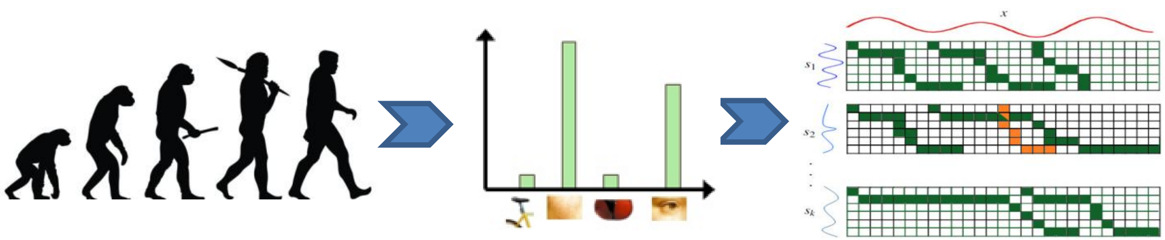
Goals

- **G1:** Define BoVW-weighting schemes representations of objects in data by means of **genetic programming** (**GP**) optimization.
- **G2:** Learn **multimodal BoVW** for recognizing gestures.
- **G3:** Gesture detection through **dynamic programming** and **generative models**.
- **G4**: Learning **Bag of Sub-Gestures** via **evolutionary computation**.

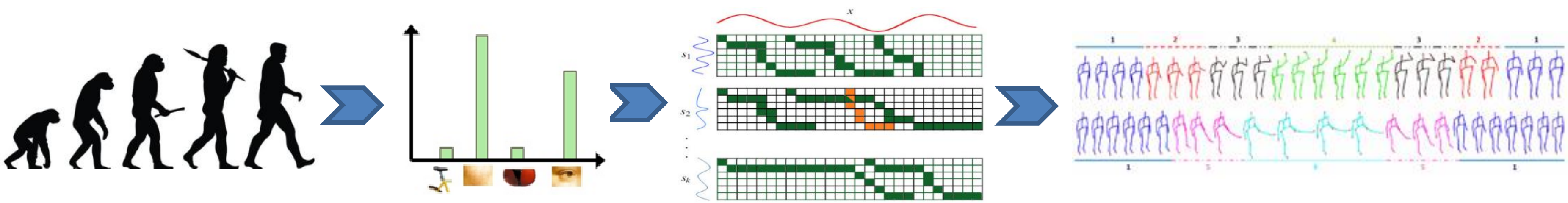
Evolving Visual Representations



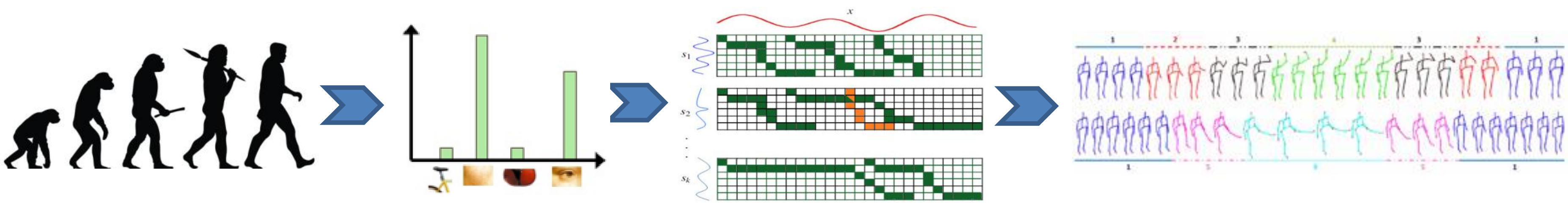
Learning Spatio-Temporal Representations



Evolving Dynamic Representations



Conclusions



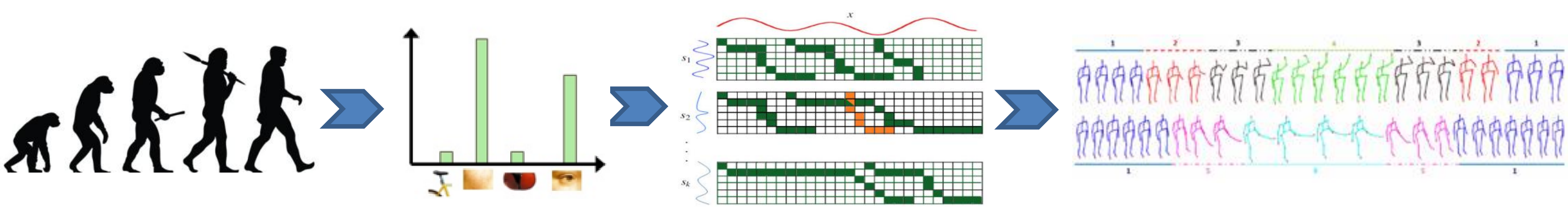
State of the Art

Evolving Visual Representations

Learning Spatio-Temporal Representations

Evolving Dynamic Representations

Conclusions



Evolving Visual Representations

Term-Weighting Schemes

- From text mining and information retrieval, the BoW representations aim at mapping documents into a vectorial space that captures information about the semantics and content of documents:

$$\mathbf{d}_i = \langle x_{i,1}, \ldots, x_{i,|V|} \rangle$$

$x_{i,j}$: Scalar that indicates the importance of the term t_j for describing the content of the i^{th} document [1].

V : Set of different words in the corpus; vocabulary.

- The way of estimating $x_{i,j}$ is given by the so called *term-weighting schemes* (TWS) [2]:
  - *TDR: term-document relevance* (local information):
    - *term-frequency* (*TF*) is the most common, which indicates the number of times a term occurs.
  - *TR*: term relevance (global information):
    - *Inverse-document-frequency* (*IDF*), which penalizes terms occurring frequently across the whole corpus.

[1] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In ICCV, volume 2, pages 1470–1477.
[2] Debole, F. and Sebastiani, F. (2003). Supervised term-weighting for automated text categorization. In SAC, pages 784–788, New York, NY, USA. ACM.

Towards BoVW

- In CV, a visual word is a prototypical visual pattern that summarizes the information of visual descriptors [1] extracted from training images: (3D)HOG, HOF, SIFT, PLS, Voxel reconstructions, CNN [2]:
  - ❑ An image is decomposed into a set of patches obtained from spatial sampling or detecting points, clustered and represented by a vector indicating the importance of visual words for describing its content.
- Effectiveness of BoVW representations depends on a number of factors:
  - ❑ Detection of interest points, choice of the visual descriptors, clustering, and the learning algorithm.
- Great advances have been obtained for incorporating spatio-temporal information [3].

[1] Zhang, J., Marszablek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV, 73(2):213–238.
[2] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. ICCV 2015.
[3] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In CVPR, pages 1–8.

Evolutionary Computation

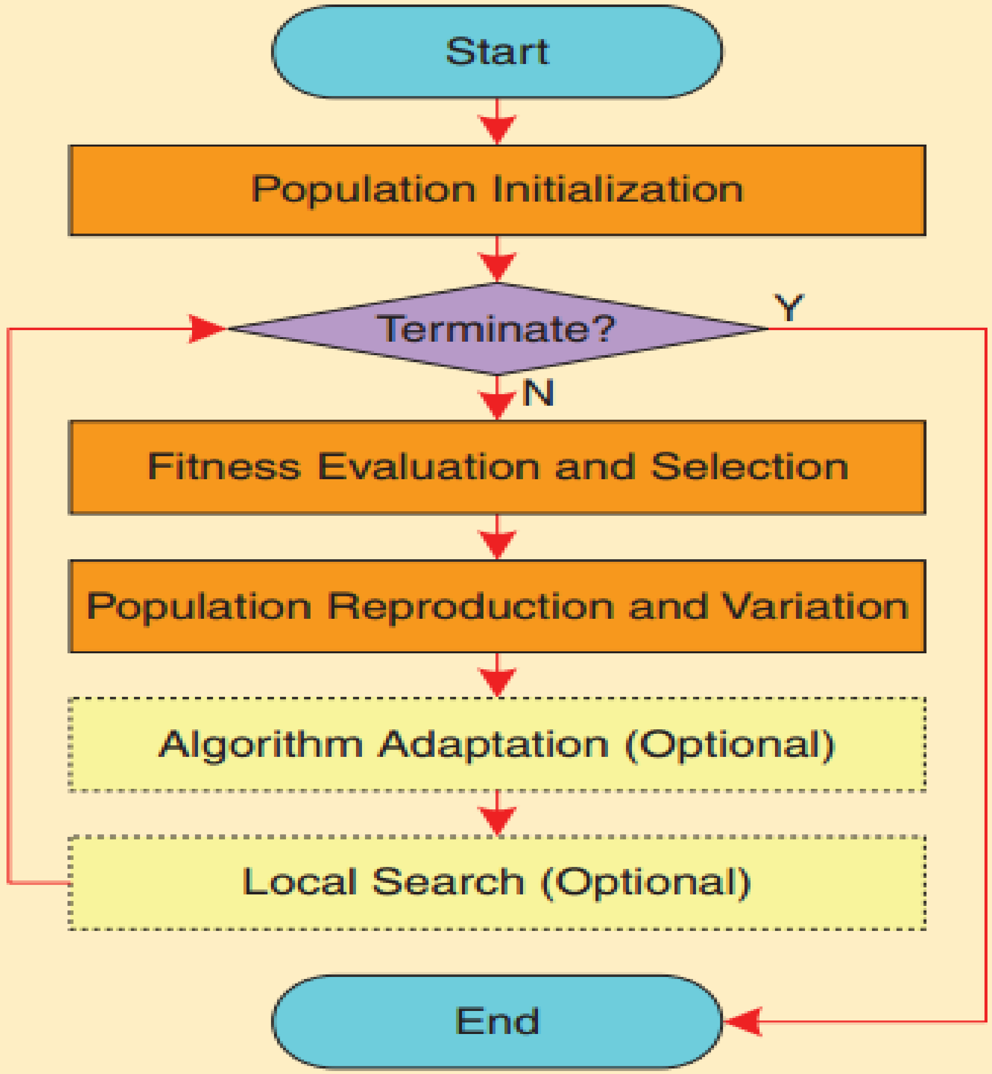
- Evolutionary algorithms (EA) have a long tradition in computer vision:

  - Genetic Algorithms (GA) was proposed as a search heuristic that mimics the process of natural selection [1] for generating useful solutions to optimization and search problems [2].

  - In Genetic Programming (GP), nonlinear and complex data structures are used to represent solutions, such as evolving interest-point detectors [3] for action recognition [4,5].

[1] J. H. Holland. University of Michigan Press, Ann Arbor. 1975. *Adaptation in Natural and Artificial Systems*.
[2] D. E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 1989.
[3] Trujillo, L. and Olague, G. (2006). Synthesis of interest point detectors through genetic programming. In GECCO, pages 887–894, New York, NY, USA. ACM.
[4] Liu, L. and Shao, L. (2013). Learning discriminative representations from rgb-d data. In IJCAI.
[5] Liu, L., Shao, L., and Rockett, P. (2012). Genetic programming-evolved spatio-temporal descriptor for human action recognition. In BMVC, pages 18.1–18.12.

Evolutionary Computation



[1] J. Zhang, Z.-H. Zhan, Y. Lin, N. Chen, Y.-J. Gong, J.-H. Zhong, H.-S.-H. Chung, Y. Li, and Y.-H. Shi. (2011). Evolutionary Computation Meets Machine Learning: A Survey. IEEE Computational Intelligence Magazine, vol. 6, nº 4, pp. 68-75.

Evolving TWS

- TWS with EA has been studied within information retrieval, text categorization and image representation.

  - Exploring supervised TWS has not been deeply studied [1]:

    - GP algorithms to learn weighting schemes by combining a set of visual primitives.

    - Applicable for learning spatio-temporal representations.

[1] Tirilly, P., Claveau, V., and Gros, P. (2009). A review of weigthing schemes for bag of visual words image retrieval. Technical report, IRISA.

EA for BoVW

- BoVW is a widely adopted representation for describing the content of images and videos in computer vision problems:
  - ❑ Standard weighting schemes based on term frequency are effective and popular:
    - Histogram that accounts for the occurrences of visual words.
  - ❑ Explore the suitability of alternative TWS for image and video representation.
- Propose an EA capable of automatically learning TWS:
  - ❑ Explore the search space of possible TWS that can be generated by combining a set of primitives with the aim of maximizing the classification/recognition performance:
    - Image categorization.
    - Adult image classification.
    - Insect and bird classification.
    - Places-scene recognition.
    - Gesture and action recognition.

Weighting Schemes

- Weighting schemes used in text mining and information retrieval:

Acr.	Name	Formula	Description
B	Boolean	$x_{i,j} = \mathbf{1}_{\{\#(t_i,d_j)>0\}}$	Prescense/abscense of terms
TF	Term-Frequency	$x_{i,j} = \#(t_i,d_j)$	Frequency of occurrence of terms
TF-IDF	TF - Inverse Doc. Freq.	$x_{i,j} = \#(t_i,d_j) \times \log(\frac{N}{df(t_j)})$	TF penalizing corpus-based frequency
TF-IG	TF - Information Gain	$x_{i,j} = \#(t_i,d_j) \times IG(t_j)$	TF times term information gain
TF-CHI	TF - Chi-square	$x_{i,j} = \#(t_i,d_j) \times CHI(t_j)$	TF times χ^2 term relevance
TF-RF	TF - Relevance Freq.	$x_{i,j} = \#(t_i,d_j) \times \log(2 + \frac{TP}{\max(1,TN)})$	TF times *RF* relevance

$x_{i,j}$: Scalar that indicates the importance of the term t_j for describing the content of the i^{th} document.

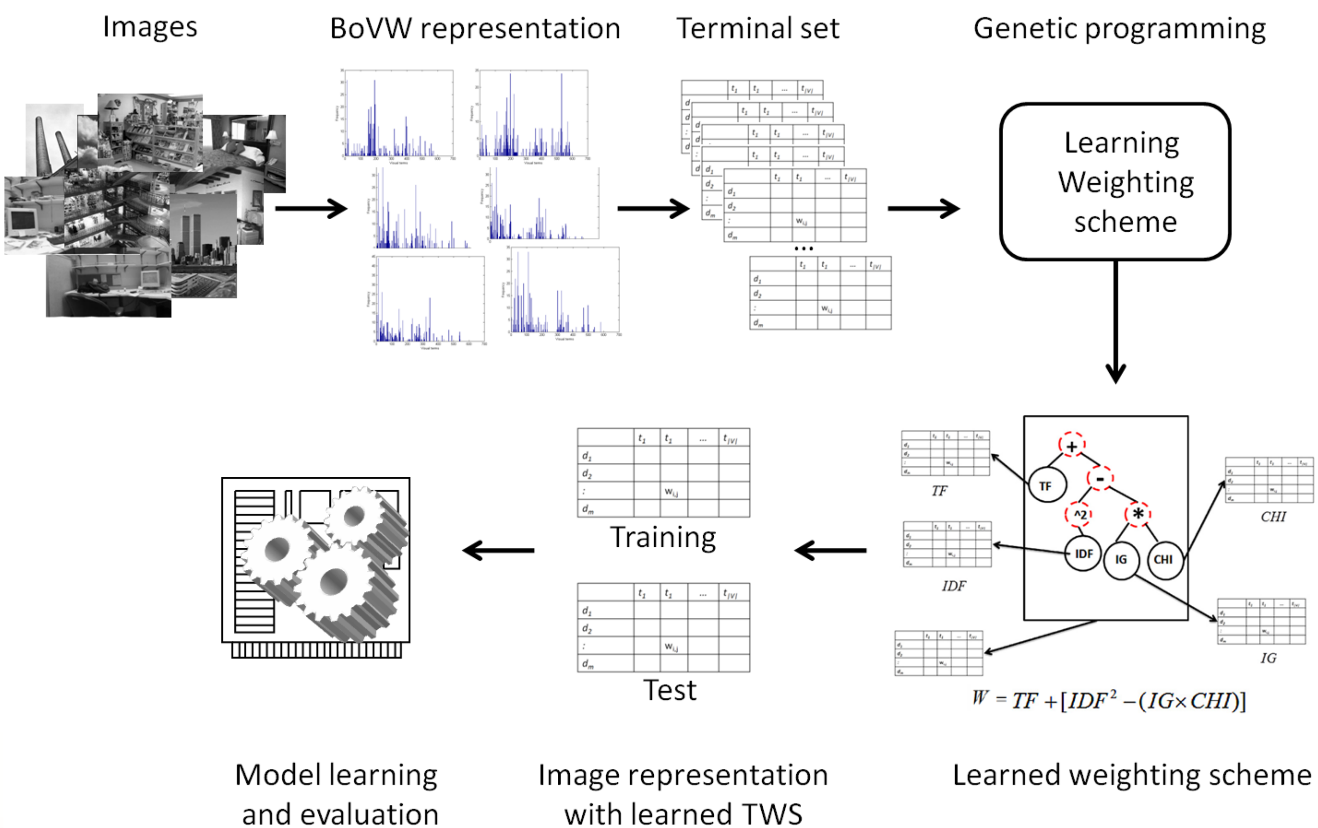
N : Number of documents in training dataset.

$df(t_j)$: Document frequency of the term t_j, i.e., the number of documents in which term t_j occurs.

[1] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Inform. Process. Manag., pages 513–523.
[2] Debole, F. and Sebastiani, F. (2003). Supervised term-weighting for automated text categorization. In Proceedings of the 2003 ACM Symposium on Applied Computing, SAC ’03, pages 784–788, New York, NY, USA. ACM.
[3] Lan, M., Tan, C. L., Su, J., and Lu, Y. (2009). Supervised and traditional term-weighting methods for automatic text categorization. In TPAMI, 31(4):721–735.

GP for learning TWS

- Automatically design weighting schemes by means of EA:
  - ❑ Use Genetic Programming to learn how to combine a set of TR/TDR primitives for every dataset in order to optimize classification performance.



Function set

Operator	Name	Arity
$+$	Addition	2
$-$	Substraction	2
$*$	Product	2
$/$	Division (protected)	2
$\log_2 x$	Logarithm b-2	1
$\sqrt{x}$	Square root	1
x^2	Square power	1

Data - still images

Caltech-101 [1]

Adult image filtering [2]

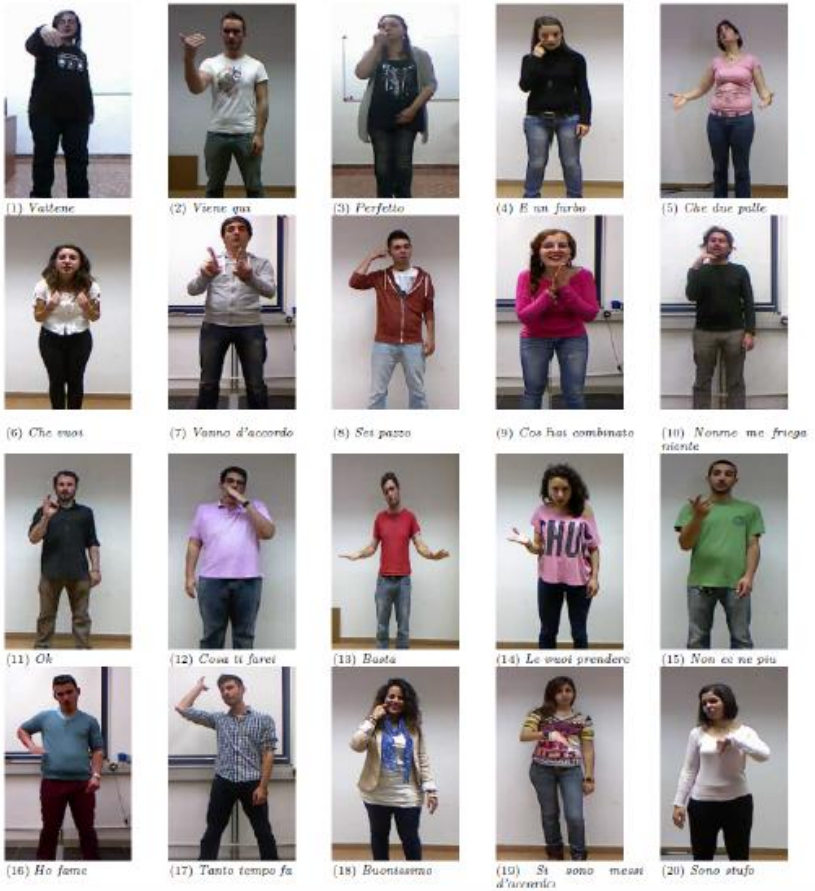
Birds and butterflies

Scene recognition [3]

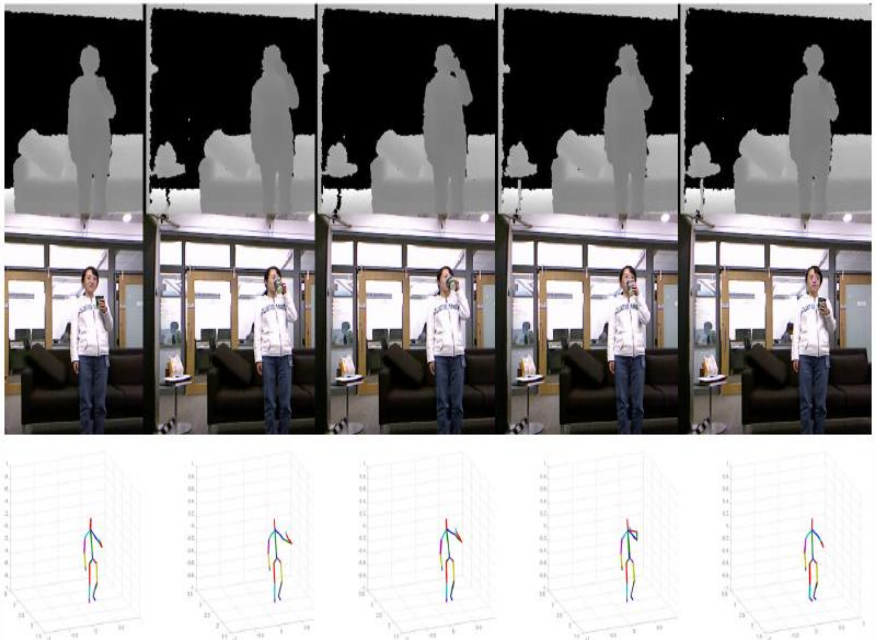
[1] Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories.In IEEE Proc. CVPRW.
[2] Deselaers, T., Pimenidis, L., and Ney, H. (2008). Bag of visual words for adult image classification and filtering. In ICPR.
[3] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVIP, pages 2169–2178.

Data - sequences

Montalbano [1]



MSRDaily3D [2]



[1] Escalera, S., Baró, X., Gonzalez, J., Bautista, M. A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H. J., Shotton, J., and Guyon, I. (2014). ChaLearn looking at people challenge 2014: Dataset and results. In ECCVW.
[2] Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In CVPR, pages 1290–1297.

Experimental settings

- The same evaluation protocol for every dataset:
  - PHOW [1] (Pyramid Histogram Of Visual Words) features used as visual descriptors.
  - Training partitions used both to obtain the visual vocabulary and to learn the term-weighting schemes with GP.
    - Learned the weighting schemes by using subsets of the training sets.

Image Categorization					
Dataset	**Classes**	**|V|**	**# Train**	**# Test**	**images|terms**
Caltech-tiny	5	12000	75	75	15|12000
Caltech-102 (15)	101	12000	1530	1530	165|3000
Caltech-102 (30)	101	12000	3060	3060	330|3000
Birds	6	400	540	60	540|400
Butterflies	7	400	552	67	552|400
Action recognition					
Dataset	**Classes**	**|V|**	**# Train**	**# Test**	**im.|terms**
MSRDaily3D	12	600	192	48	192|600
Gesture recognition					
Dataset	**Classes**	**|V|**	**# Train**	**# Test**	**im.|terms**
Montalbano	20	1000	6850	3579	2055|600
Scene recognition					
Dataset	**Classes**	**|V|**	**# Train**	**# Test**	**im.|terms**
15 Scenes	15	12000	1475	3010	1475|2000
Pornographic image filtering					
Dataset	**Classes**	**|V|**	**# Train**	**# Test**	**im.|terms**
Adult	5	12000	6808	1702	6808|2000

[1] Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classifcation using random forests and ferns. In ICCV.

Experimental settings

- The same evaluation protocol for every dataset:
  - ❑ PHOW [1] (Pyramid Histogram Of Visual Words) features used as visual descriptors.
  - ❑ Training partitions used both to obtain the visual vocabulary and to learn the term-weighting schemes with GP.
    - Learned the weighting schemes by using subsets of the training sets.
  - ❑ Fitness goal: maximize the F1-measure under 5−fold cross validation.
    - Training and test images are represented with the winner weighting schemes.
  - ❑ Learning from training images and performance of the model evaluated in test images.
  - ❑ Reported the average and standard deviation performance of 5 runs of the GP.
  - ❑ Run in all cases for 50 generations with a population of 500 individuals [2].
  - ❑ Default values were used for the remainder of GP parameters:
    - Generational selection mechanism with elitism.
    - Lexictour parent selection [3].
    - Crossover and mutation probabilities.

[1] Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classifcation using random forests and ferns. In ICCV.
[2] Langdon, W. B. and Poli, R. (2001). Foundations of Genetic Programming. Springer.
[3] Luke, S. and Panait, L. (2002). Lexicographic parsimony pressure. In Proceedings of GECCO, pages 829–836.

Results

- Results obtained by the different weighting schemes (traditional, alternative-supervised and learned) in all of the considered datasets:
  - ❑ Average f1−measure performance in the test partitions.

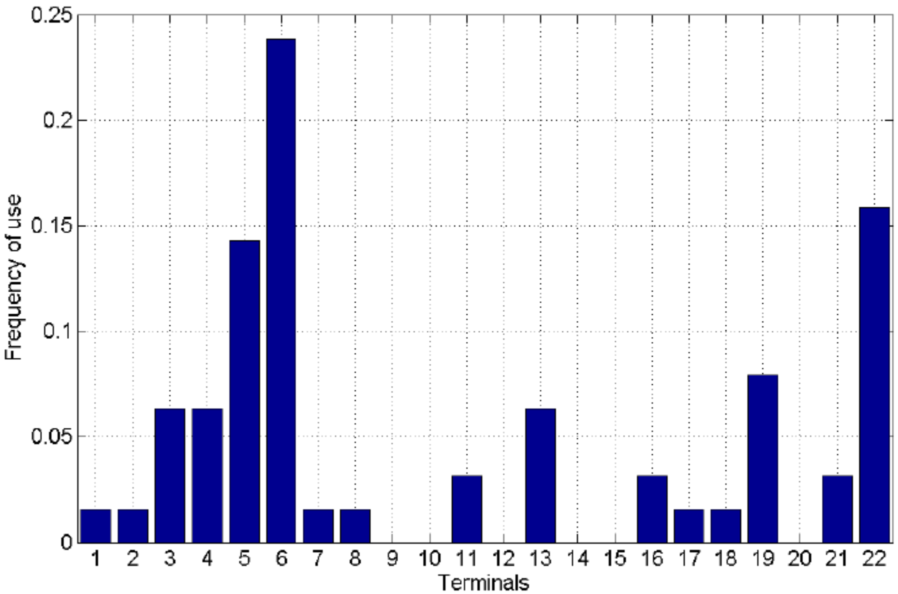
	Traditional			Alternative-supervised			Learned
Dataset / TWS	**TF (baseline)⋆**	**BoI.⋆**	**TF-IDF⋆**	**TF-RF⋆ [81]**	**TF-CHI⋆ [33]**	**TF-IG⋆ [33]**	**GP (ours)**
Tiny	85.65	84.01	76.72	85.65	78.85	80.49	**90.75±1.56**
101-15	52.26	58.43	48.08	52.30	52.00	51.43	**61.05±1.12**
101-30	56.61	59.28	49.95	56.68	54.63	52.03	**63.04±1.02**
Birds	44.68	48.53	30.55	44.68	44.6	43.95	**52.95±5.11**
Butterflies	26.07	41.44	20.45	26.07	26.08	26.75	**42.12±3.07**
Adult	52.53	58.35	55.39	52.53	46.39	47.23	**62.68±2.08**
15 scenes	59.12	61.26	56.51	59.12	55.02	55.07	**63.43±0.16**
Montalbano	88.55	86.46	88.49	88.55	88.5	88.58	**88.79±0.12**
MSRDaily3D	75.22±4.2	68.0±6.22	74.72±4.47	75.058±3.9	73.94±5.65	73.77±4.9	**76.01±4.01**
Average	54.34±22.06	56.91±18.78	50.81±22.38	54.33±22.04	52.46±21.04	52.51±21.11	**61.45±18.67**

ID	Dataset	Learned TWS	Formula
1	Caltech101-15	sqrt((sqrt(RF× TF)+log2(RF× TF)))	$\sqrt{\sqrt{W_{22}} + \log 2(W_{22})}$
2	Birds	log2((FMeas × (CHI × log2(TF × RF))))	$\log 2(W_{16} \times (W_3 \times \log 2(W_{22}))$
3	MSRDaily3D	((TF × FN) × sqrt(T))	$((W_6 \times W_{11}) \times \log 2(\sqrt{W_{22}}))$
4	Adult	(sqrt(IDF)×D)	$(\sqrt{W_5} \times D)$
5	Montalbano	log2(log2(CHI))× sqrt(IDF)	$(\log 2(\log 2(W_3)) \times \sqrt{W_5})$
6	15-Scenes	log2((ProbR + TF × RF))	$\log 2(W_{19} + W_{22})$

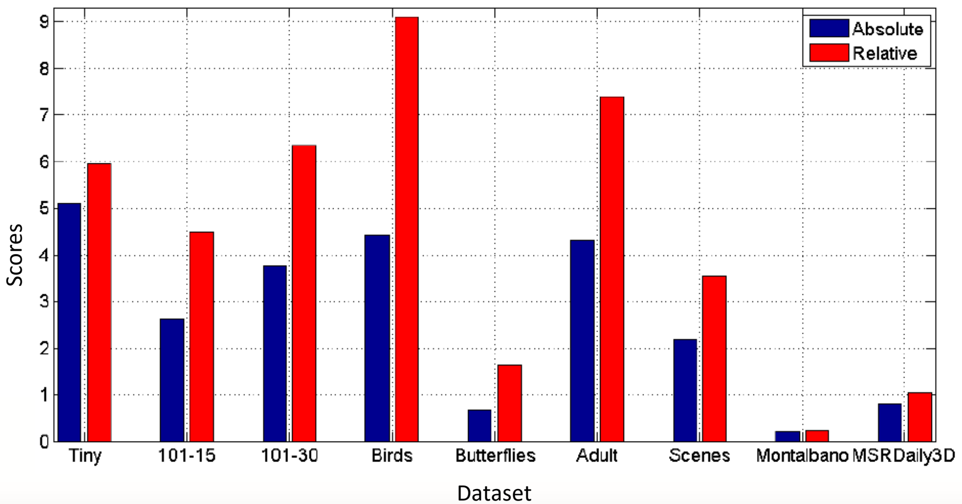
Results

- Range of improvement of the proposed method over the best traditional/alternative weighting scheme per dataset in terms of absolute and relative differences.

Frequency of appearance of terminals into the solutions found by the GP.



Absolute and relative improvements for the different datasets, taking as reference the best traditional/alternative weighting scheme for each dataset.



GP for learning TWS

Terminal set Representation

Variable	Meaning
W_1	N, Constant matrix, number of training documents.
W_2	$\|V\|$, Constant matrix, number of terms.
W_3	CHI, Matrix containing in each row the vector of χ^2 weights for the terms.
W_4	IG, Matrix containing in each row the vector of information gain weights for the terms.
W_5	$TF \times IDF$, Matrix with the TF-IDF term-weighting scheme.
W_6	TF, Matrix containing the TF term-weighting scheme.
W_7	FGT, Matrix containing in each row the global term-frequency for all terms.
W_8	TP, Matrix containing in each row the vector of true positives for all terms.
W_9	FP, Matrix containing in each row the vector of false positives.
W_{10}	TN, Matrix containing in each row the vector of true negatives.
W_{11}	FN, Matrix containing in each row the vector of false negatives.
W_{12}	$Accuracy$, Matrix where each row contains the accuracy obtained when using the term as classifier.
W_{13}	$Accuracy\_Balance$, Matrix containing the AC_Balance each (term, class).
W_{14}	Bi-normal separation, BNS, An array that contains the value for each BNS per (term, class).
W_{15}	$DFreq$, Document frequency matrix containing the value for each (term, class).
W_{16}	$FMeasure$, F-Measure matrix containing the value for each (term, class).
W_{17}	$OddsRatio$, An array containing the OddsRatio term-weighting.
W_{18}	$Power$, Matrix containing the Power value for each (term, class).
W_{19}	$ProbabilityRatio$, Matrix containing the ProbabilityRatio each (term, class).
W_{20}	Max_Term, Matrix containing the vector with the highest repetition for each term.
W_{21}	RF, Matrix containing the RF vector.
W_{22}	$TF \times RF$, Matrix containing TF-RF.

Learning Spatio-temporal Representations

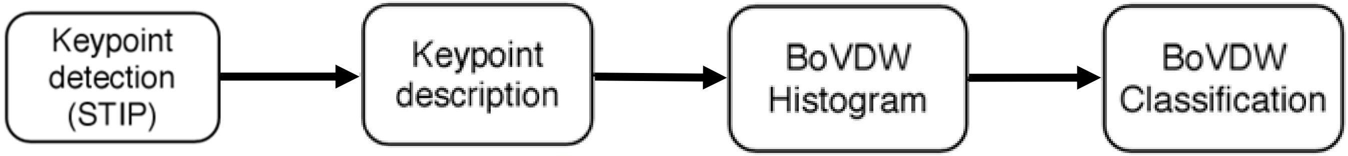
Multimodal representations

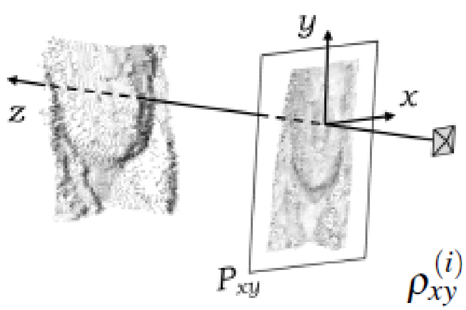
- Most approaches are based on classic computer vision techniques applied to RGB data [1]. However, extracting discriminative information from standard image sequences is sometimes unreliable.
  - ❑ Compact multi-modal devices allow 3D partial information to be obtained from the scene [2]: New descriptors combining RGB plus Depth (RGB-D) for HCI Apps:
    - Inferring pixel label probabilities from learned offsets of depth features [3].
- As an extension of BoVW, these approaches attempt to benefit from the multimodal fusion of visual and depth features.
- This information has been particularly exploited for human gesture recognition, body segmentation and tracking.

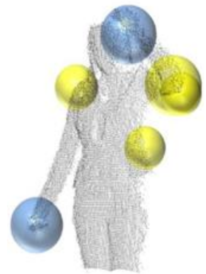
[1] Tirilly, P., Claveau, V., and Gros, P. (2009). A review of weigthing schemes for bag of visual words image retrieval. Technical report, IRISA.
[2] HD. Yang, S. L. (2007). Reconstruction of 3d human body pose from stereo image sequences based on top-down learning. Pattern Recognition, 40(11):3120–3131.
[3] Shotton, J. et al. (2011). Real-time human pose recognition in parts from single depth images. In CVPR, pages 1297–1304.

Gesture Representation

- BoVDW approach: Merging RGB + Depth information by means of late fusion.

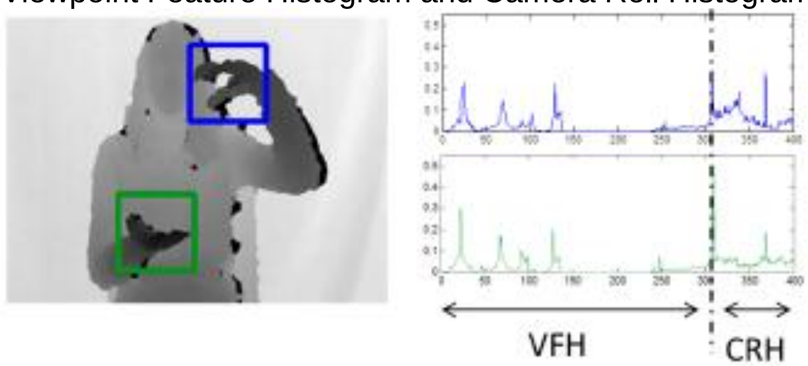


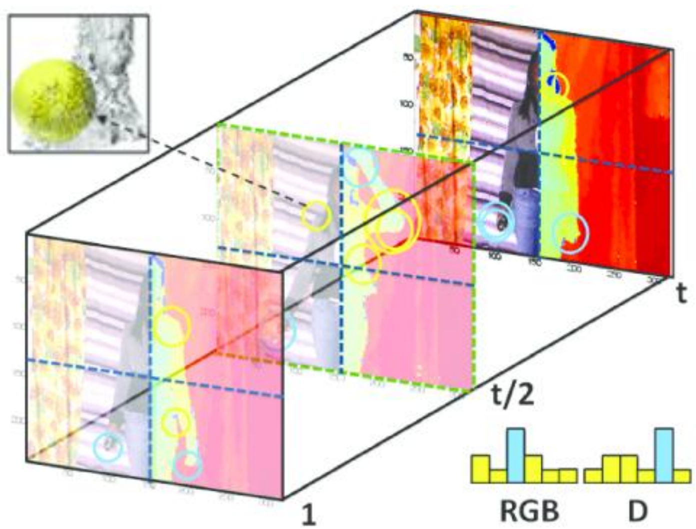




$$\rho_{xy}^{(i)} = ||\rho^{(i)}|| \cdot \sin(\phi)$$

Viewpoint Feature Histogram and Camera Roll Histogram





$$d^F = 1 - \sum_i \min(h^F_{model}(i), h^F_{query}(i))$$

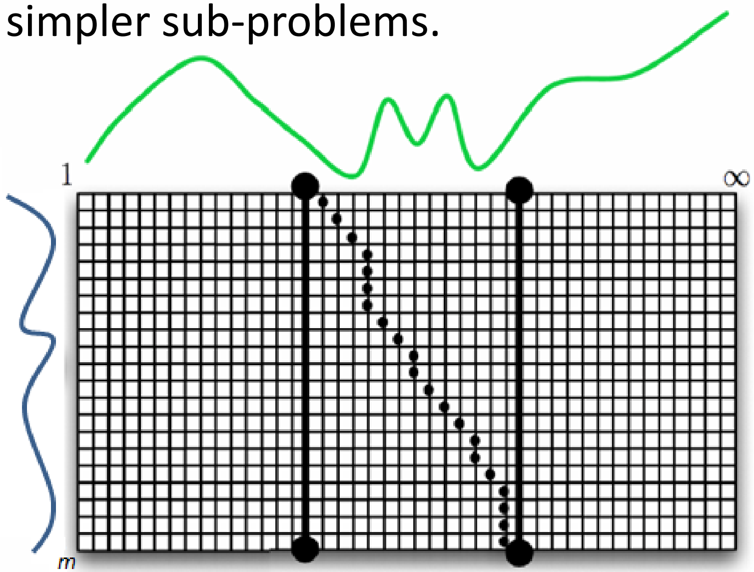
$$d_{hist} = (1-\varsigma)d^{RGB} + \varsigma d^D$$

[1] Laptev, I. (2005). On space-time interest points. In IJCV, 64(2-3):107–123.
[2] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. CVPR, 2:886–893.
[3] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In CVPR, pp. 1–8.
[4] Rusu, R., Bradski, G., Thibaux, R., and Hsu, J. (2010). Fast 3d recognition and pose using the viewpoint feature histogram. In IROS, pp. 2155 –2162.

Probabilistic Dynamic Programming

- In the context of gesture recognition, it is common the use of methods based on *dynamic programming*, which breaks down a complex problem into a collection of simpler sub-problems.

$$M(i,j) = d(i,j) + \min\{M(i-1,j-1), M(i-1,j), M(i,j-1)\}$$



$$\Omega = \{v_1, v_2, \ldots, v_\tau\}$$

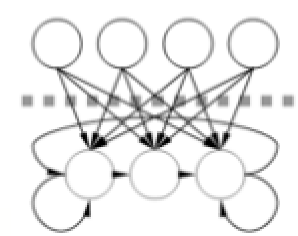
$$DTW(M) = \min_{\Omega}\left\{\frac{M(v_\tau)}{\tau}\right\}$$

$\flat \in [1, .., \infty]$ s.t. $M(m,\flat) < \theta$ $\Longleftrightarrow$

for $i = 1 : m$ **do**
  for $j = 1 : \infty$ **do**
    $M(i,j) = \infty$
end end
for $j = 1 : \infty$ **do**
  $M(0,j) = 0$
end

if $M(m,j) < \theta$ **then**
  $\Omega = \{\arg\min_{x' \in \mathcal{N}(x)} M(x')\}$
  return
end

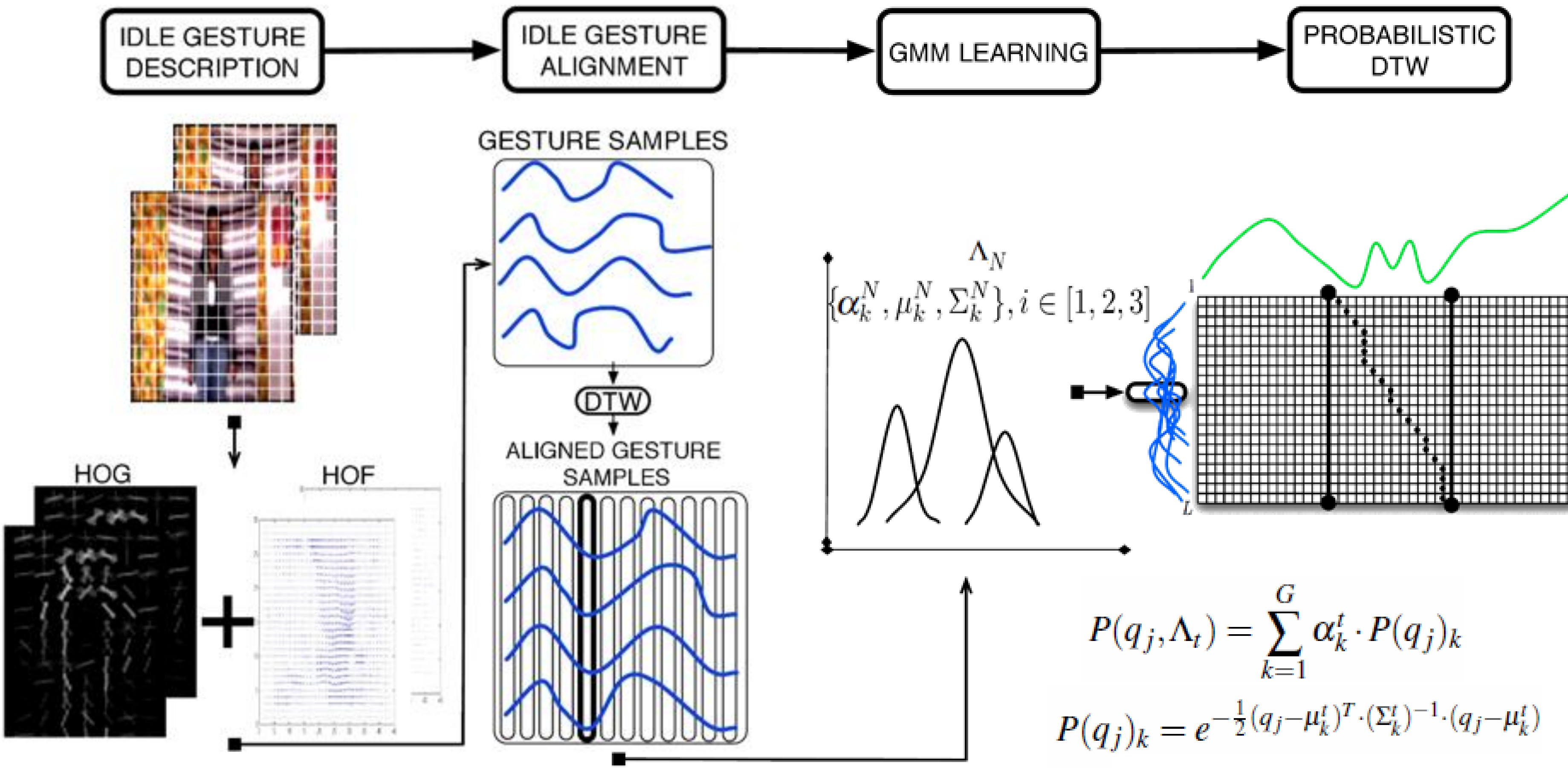
- Generative models allow to deal with the high variability due to environmental conditions among different domains:
  - ❑ Wide range of human pose configurations, influence of background, continuity of human movements, spontaneity of human actions, speed, appearance of unexpected objects, illumination changes, partial/self occlusions, different points of view…



[1] Parizeau, M. and Plamondon, R. (1990). A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification. IEEE TPAMI, 12(7).
[2] Reyes, M., Dominguez, G., and Escalera, S. (2011). Feature weighting in dynamic time warping for gesture recognition in depth data. ICCV Workshops, pp. 1182–1188.
[3] Zhou, F., De la Torre Frade, F., and Hodgins , J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE TPAMI, 35(3):582–596.

Gesture Segmentation

- Generative models learned in PDTW handle the variance present in data.



$$P(q_j,\Lambda_t) = \sum_{k=1}^{G} \alpha_k^t \cdot P(q_j)_k$$

$$P(q_j)_k = e^{-\frac{1}{2}(q_j-\mu_k^t)^T \cdot (\Sigma_k^t)^{-1} \cdot (q_j-\mu_k^t)}$$

$$p(\widetilde{F}_t) = \sum_{k=1}^{G} \alpha_k^t \cdot e^{-\frac{1}{2}(x-\mu_k^t)^T \cdot (\Sigma_k^t)^{-1} \cdot (x-\mu_k^t)}$$

$$D(q_j,\Lambda_t) = \exp^{-P(q_j,\Lambda_t)}$$

[1] Yu, Guoshen (2012). "Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity". *IEEE Transactions on Image Processing* 21 (5): 2481–2499.

Data - CGD

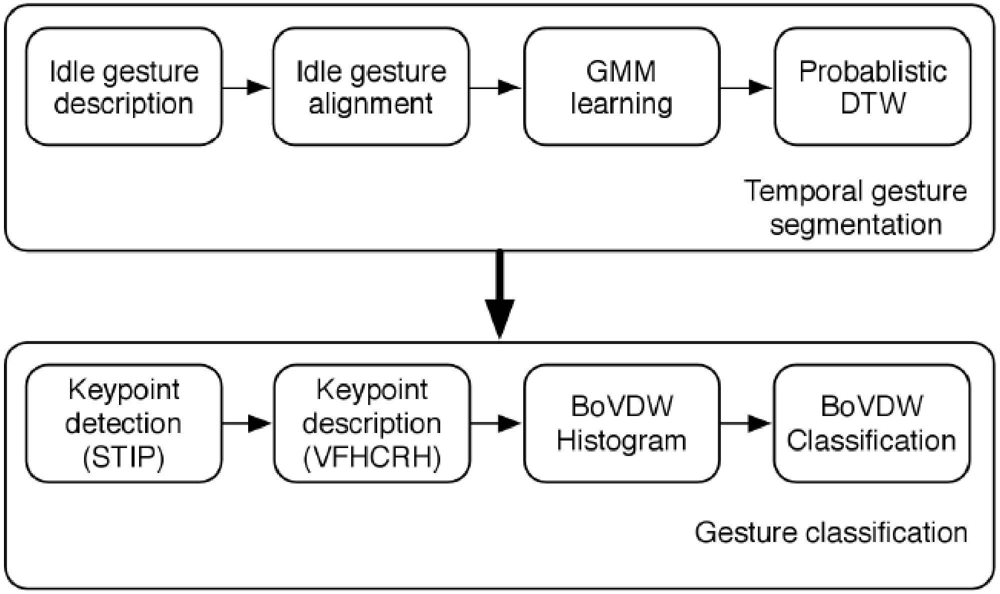
- Challenge Gesture Dataset (CGD) [1] of 50,000 gesture video sequences.
  - ❑ Single user in front of a fixed camera.
  - ❑ Images are captured by the Kinect™ device providing both RGB and depth images.
  - ❑ 20 development batches with a manually tagged gesture segmentation:
    - 100 recorded gestures grouped in sequences of 1-5 gestures performed by the same user.
    - Different lexicon of 8-15 unique gestures and just one training sample per gesture is provided, categorized in 9 (10) classes:
      - Body language gestures (scratching your head, crossing your arms, etc.).
      - Gesticulations (performed to accompany speech).
      - Illustrators (like Italian gestures).
      - Emblems (like Indian Mudras).
      - Signs (from sign languages for the deaf).
      - Signals (diving signals, mashalling signals to guide machinery or vehicle, etc.).
      - Actions (like drinking or writing).
      - Pantomimes (gestures made to mimic actions).
      - Dance postures.

      - **~1800 Idle gesture** in between (for temporal segmentation).

[1] Guyon, I., Athitsos, V., Jangyodsuk, P., and Escalante, H. J. (2014). The chalearn gesture dataset (CGD 2011). Machine Vision and Applications, 25(8):1929–1951.

PDTW and BoVDW

- Address the problem of continuous human gesture recognition:
  - ❑ Recognize idle (or reference) gestures performed between gestures.
    - Gesture Segmentation: Probability-based Dynamic Time Warping (PDTW).
    - Gesture Representation: Bag of Visual and Depth Words (BoVDW).



  - ❑ The experiments are performed using the public dataset provided by the ChaLearn Gesture Challenge [1].
  - ❑ Standard BoVW model and early fusion are compared to the proposed late fusion.

[1] Guyon, I., Athitsos, V., Jangyodsuk, P., and Escalante, H. J. (2014). The chalearn gesture dataset (CGD 2011). Machine Vision and Applications, 25(8):1929–1951.

Temporal Segmentation

- Each idle gesture sequence is described using a grid approach of HOGHOF descriptor, and a random projection for reducing dimensionality.
- Ten-fold cross validation strategy using 180 idle gestures as validation data:
  - ❑ Chosen DTW cost threshold θ by maximizing the overlap.
  - ❑ Chosen Gaussian components *G* for the GMM by means of 10-fold CV.
  - ❑ Baum-Welch algorithm for training an HMM:
    - Vocabulary computed using *k*-means over idle gestures.
    - Empirically set hidden states.
- Recognition is performed with temporal sliding windows of different wide sizes, based on the idle gesture samples length variability.

	Overlap.	Acc.
Probability-based DTW	**0.3908± 0.0211**	**0.6781 ± 0.0239**
Euclidean DTW	0.3003 ± 0.0302	0.6043 ± 0.0321
HMM	0.2851 ± 0.0432	0.5328 ± 0.0519

Temporal Segmentation

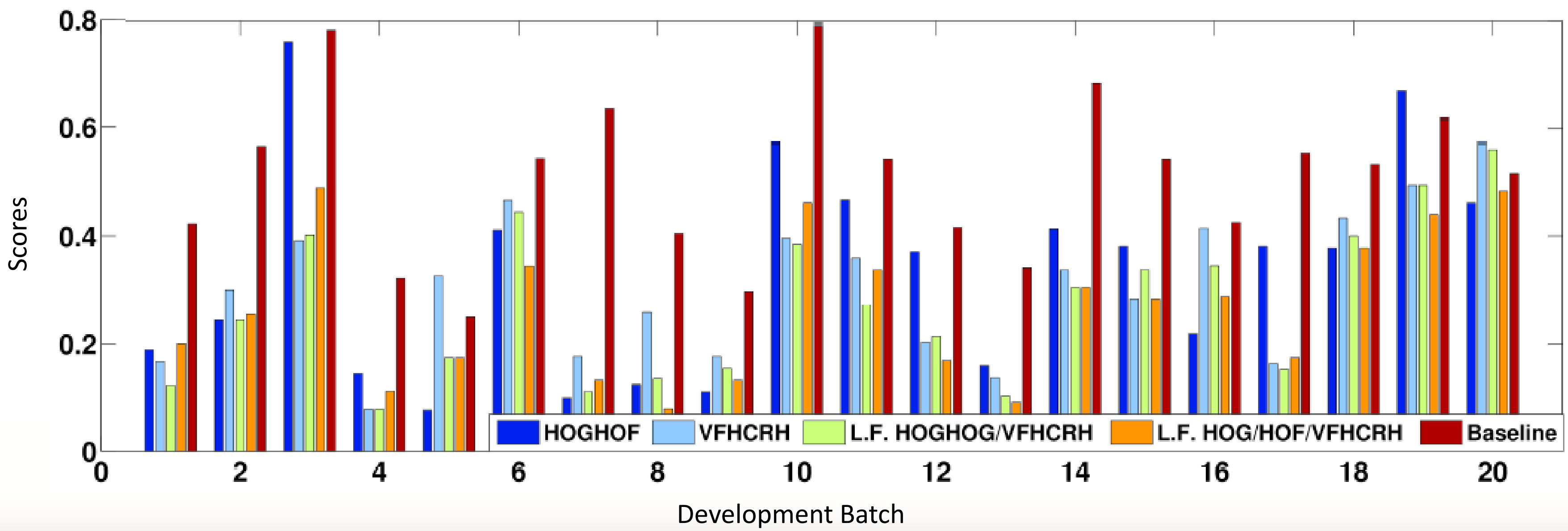
BoVDW classification

- Empirically set V, $b_u \times b_\upsilon \times b_\varpi$, and ς.
- Mean Levenshtein Distance (MLD) over all gesture sequences.

RGB desc.	MLD	Depth desc.	MLD
HOG	0.3452	VFH	0.4021
HOF	0.4144	VFHCRH	**0.3064**
HOGHOF	**0.3314**		

- Late fusion of best descriptors HOG, HOF and VFHCRH:

2-LF MLD: 0.2714 ; **3-LF MLD: 0.2662**



[1] Guyon, I., Athitsos, V., Jangyodsuk, P., and Escalante, H. J. (2014). The chalearn gesture dataset (CGD 2011). Machine Vision and Applications, 25(8):1929–1951.

Evolving Dynamic Representations

Gesture & Action recognition

- Landmark tasks of the so called Looking at People field: the visual analysis of humans [1].
- Exponential growth on research, with a variety of methods proposed from the nineties, since the release of low-cost multimodal sensors [2].
- Traditional methods were based on temporal templates [3], sequence alignment or statistical sequential modeling. They approach the problem in a *holistic* way.
  - ❑ Inspiration of part-based techniques: dynamic-poselets [4], sub-gestures [5].
- Evolutionary algorithms have been also developed for key-frame extraction [6]:

Bag of Key Poses (BoKP)	Bag of Sub-Gestures (BoSG)
Learn subsets of frames	Learn spatio-temporal units
Class-specific key-poses	Inter-class subgestures (shared primitives)

[1] Moeslund, T., Hilton, T., Kruger, A., and Sigal, V. (2011). Visual Analysis of Humans, Looking at People. Springer.
[2] Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. Trans. on SMC-C, 37(3):311–324.
[3] Bobick, A. and Davis, J. (2001). The recognition of human movement using temporal templates. IEEE TPAMI, 23(3):257–267.
[4] Wang, L., Qiao, Y., , and Tang, X. (2014). Video action detection with relational dynamic-poselets. In ECCV.
[5] Malgireddy, et al. (2011). A shared parameter model for gesture and sub-gesture analysis. In Combinatorial Image Analysis, vol. 6636, pp. 483–493.
[6] Chaaraoui, A. and Florez-Revuelta, F. (2014). Adaptive human action recognition with an evolving bag of key poses. IEEE TAMD, 6(2):139–152.

Bag of Sub-Gestures

- Gesture & Action recognition, two widely studied tasks and topics in computer vision:
  - ❑ Attempt to capture and recognize whole gestures (in a holistic approach).
  - ❑ Classical approaches are based on DTW [1] and HMM [2].
- Recent research is moving towards approaches that model the problem in terms of gesture primitives (or subgestures) [3–5]:
  - ❑ The underlying assumption is that whole gestures are composed by primitives:
    - Shared or not among gestures from different categories.
  - ❑ The hypothesis is that learning with primitives leads to better recognition performance.
    - How to define/learn subgestures and how to perform inference models are still open questions.
- Describe a novel approach using subgesture modelling:
  - ❑ Learn subgestures by searching for temporal patterns that improve performance.
  - ❑ EA with ad-hoc variations operators suitable for learning primitives.
  - ❑ Learning and inference referred to DTW and HMM using subgestures.

[1] Bobick, A. and Wilson, A. (1997). A state-based approach to the representation and recognition of gestures. IEEE TPAMI, 19(12):1325–1337.
[2] Wilson, A. and Bobick, A. (1999). Parametric hidden markov models for gesture recognition. IEEE TPAMI, 21(9):884–900.
[3] Li, K., Hu, J., and Fu, Y. (2012). Modeling complex temporal composition of actionlets for activity prediction. ECCV, vol. 7572, pages 286–299.
[4] Malgireddy, M. R., Nwogu, I., Ghosh, S., and Govindaraju, V. (2011). A shared parameter model for gesture and sub-gesture analysis. In CIA, vol. 6636, pp. 483–493.
[5] Wang, L., Qiao, Y., , and Tang, X. (2014b). Video action detection with relational dynamic-poselets. In ECCV.

Training subgestures

- Goal: find a subgesture set $S = \{s_1, s_2, \dots, s_k\}$ from $X^T = \{x_1^T, x_2^T, \dots, x_n^T\}$ that maximizes recognition performances given a particular recognition method.

Given a **Population**, the whole continuous **training** and **validation sequences** → Decode each new unique and valid individual → number of k clusters and number of segments → Aligned Temporal Clustering: **Subgestures**

$k, segments \leftarrow decode(I);$ $X^T_{seg} \leftarrow getDataPartitions(X^T, segments);$ $S \leftarrow k\text{-}meansDTW(X^T_{seg}, k)$

Dynamic Programming

Subgesture Dissimilarities

$\mathbf{D} \leftarrow getDissimilarities(S);$[1]

Class Representants

$R \leftarrow getResizedClassModels(X^T, S);$

1D Subgesture Models

$M \leftarrow getUpdatedCosts(R, S);$

1D validation sequences

ω ← Model parameters

Generative Model

1D Training sequences

$D^T \leftarrow getUpdatedCosts(X^T, S);$

1D validation sequences

$D^V \leftarrow getUpdatedCosts(X^V, S);$[1]

Learn class-HMM

$M \leftarrow learnGM(D^T, \omega);$

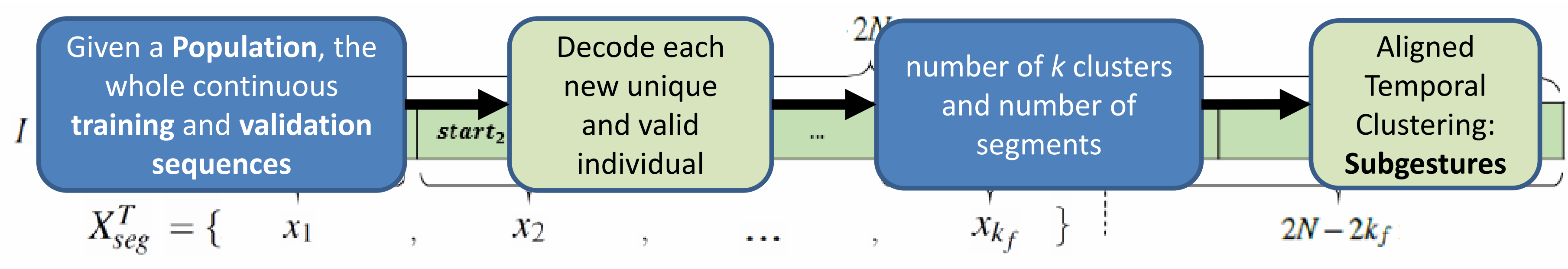
ω ← Model parameters

Score, ω^* ← evaluation function $g(D^V, \omega)$

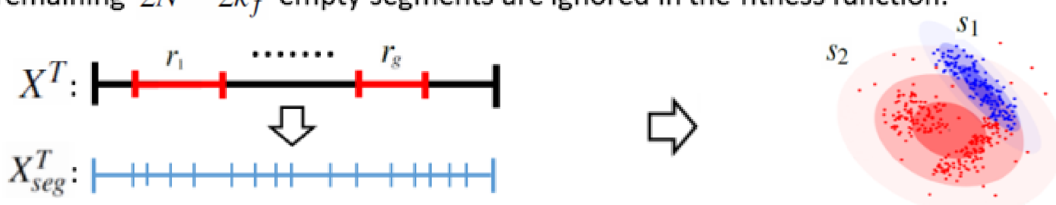
[1] Two-sided column blocks can be computed in parallel.

Evolutionary optimization

Given a **Population**, the whole continuous **training** and **validation sequences** → Decode each new unique and valid individual → number of *k* clusters and number of segments → Aligned Temporal Clustering: **Subgestures**



- Initially, there is a probability p_s of randomly selecting each pair-wise segment.
- Constraints for the frame length of segments and number of clusters:
  - ❑ Segments length within the range $[n_{min}, n_{max}]$[1].
  - ❑ The number of generated segments is $k_f \leq N$, each one in the range $[k_0, k_f]$, such that $k_0 \leq k \leq k_f$:
    - That is, the number of clusters allowed is set depending on the generated segments.
    - The remaining $2N-2k_f$ empty segments are ignored in the fitness function.



- The goal in the fitness function of the GA is to maximize the score given by the evaluation function, so as to obtain a measure of performance of subgesture models.

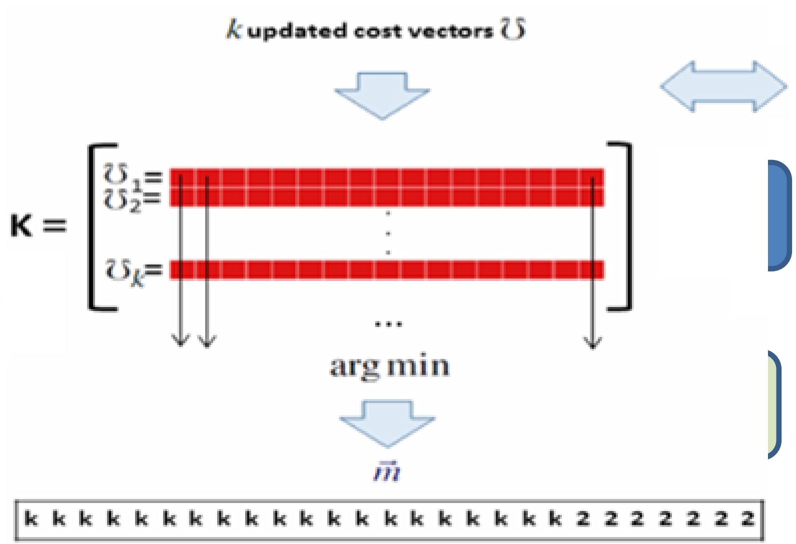
[1] Zhou, F., De la Torre Frade, F., and Hodgins , J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE TPAMI, 35(3):582–596.

Fitness function

Subgesture Dissimilarities

$$w_{ij} = DTW(s_i, s_j) = min_{\Omega}\{\sum_{p=1}^{\tau} d_p, \Omega = < v_1, v_2, \ldots, v_\tau >\}$$

$$\mathbf{D} = \mathbf{W} + \mathbf{W}^{\mathsf{T}}, \quad s.t. \quad \mathbf{W} = \frac{1}{\gamma}\begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1k} \\ w_{21} & w_{22} & \ldots & w_{2k} \\ . & . & \ldots & . \\ w_{k1} & w_{k2} & \ldots & w_{kk} \end{bmatrix}$$



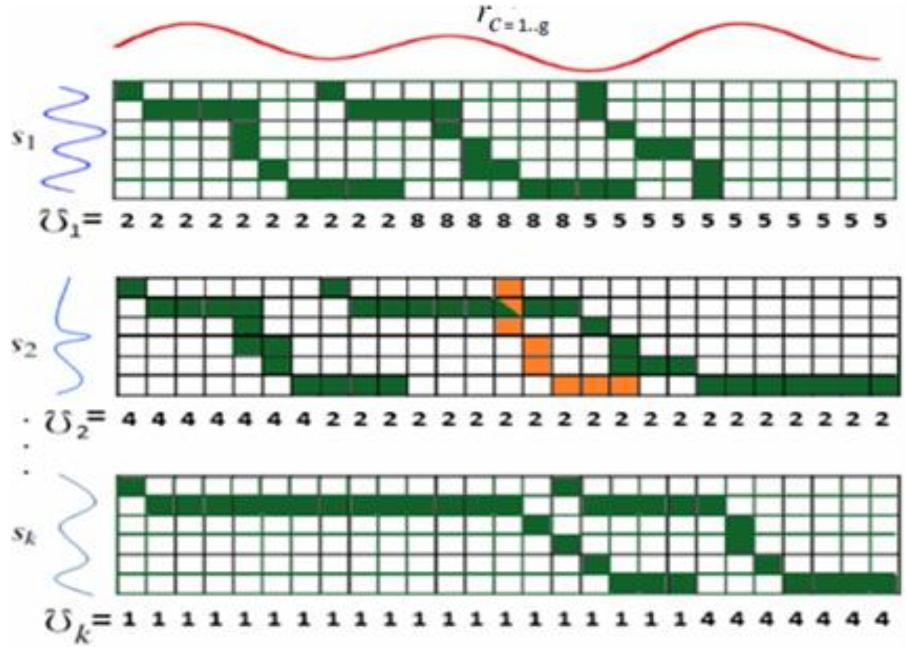
ω ← Model parameters

1D Subgesture Models

1D Training sequences

1D validation sequences

Class Representants



Learn class-HMM

ω ← Model parameters

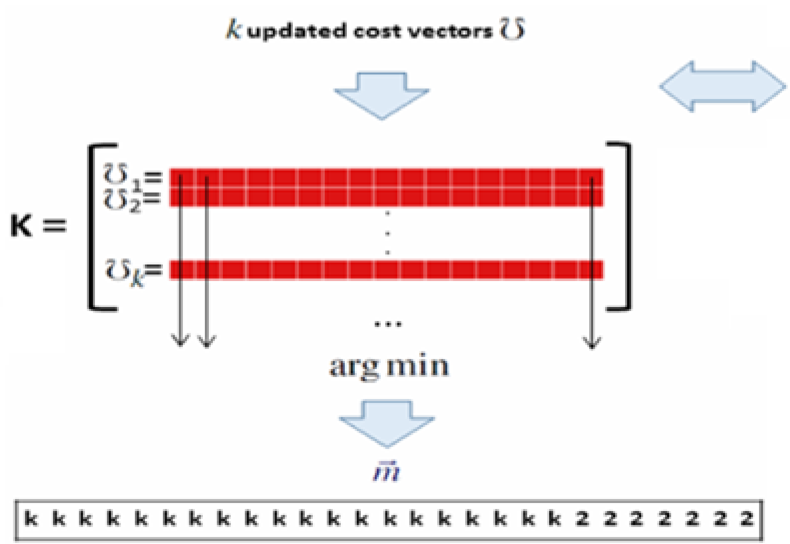
Score, ω^* ← evaluation function $g(D^V, \omega)$

Fitness function

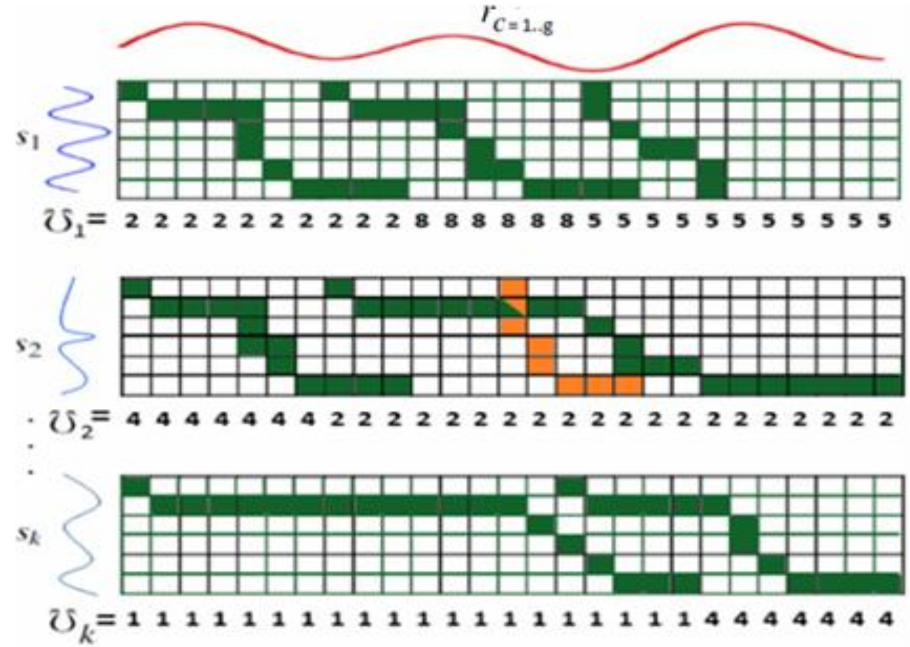
Subgesture Dissimilarities

$$w_{ij} = DTW(s_i, s_j) = min_{\Omega}\{\sum_{p=1}^{\tau} d_p, \Omega = < v_1, v_2, \ldots, v_\tau >\}$$

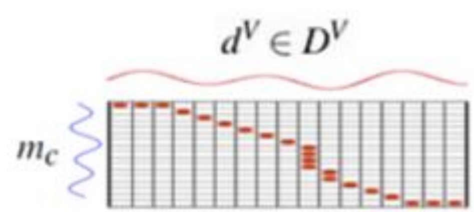
$$\mathbf{D} = \mathbf{W} + \mathbf{W}^{\intercal}, \quad s.t. \quad \mathbf{W} = \frac{1}{\gamma}\begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1k} \\ w_{21} & w_{22} & \ldots & w_{2k} \\ . & . & \ldots & . \\ w_{k1} & w_{k2} & \ldots & w_{kk} \end{bmatrix}$$



Class Representants

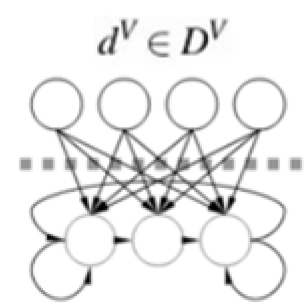


Dynamic Programming: DTW



$$DTW(m_c, d^V) = min_{\Omega}\{\sum_{p=1}^{\tau} \mathrm{D}(i_p, j_p), \Omega = < v_1, v_2, \ldots, v_\tau >\}$$

Generative modelling: HMM learning $m_{c=1..g}$ and inference



1D Subgesture Models

1D Training sequences

1D validation sequences

Learn class-HMM

Score, $\omega^* \leftarrow$ evaluation function $g(D^V, \omega)$

Operators and Evaluation

- Standard genetic operators are considered for *selection*, *crossover*, and *mutation* [1].
  - However, before mutation operator is applied each of the *N* segments has again a random probability p_s either to *add* if it is empty, or to *delete* if it already exists:
  - *Offsprings* require to ensure both that they are evaluable on the next generations and that the new trends caused by genetic modifications are respected:
    - Check and correct the segment boundaries and number of clusters by means of a *repair* function:

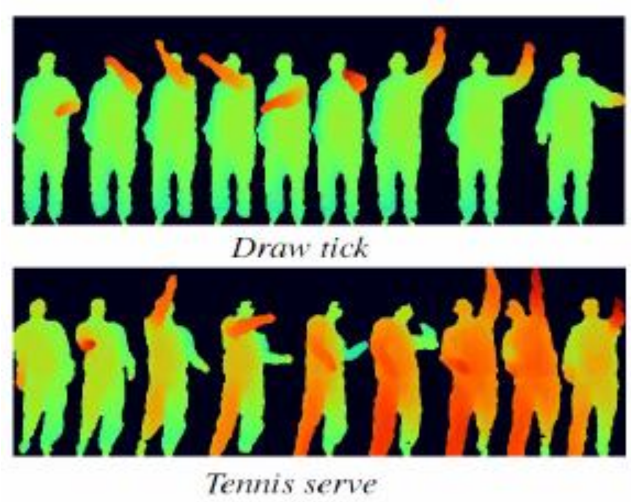
$$p(k) = \frac{k - k_0}{k_f - k_0} \Rightarrow \begin{cases} \text{if } p(k) \leq 1 & \text{increase } k_f \text{ segments} \\ \text{Otherwise} & \text{decrease } k \text{ clusters.} \end{cases}$$

- After learning the class-thresholds $\Theta = \{\theta^{c_1}, \theta^{c_2}, \ldots, \theta^{c_g}\}$, the evaluation function computes the mean score of classifying each sequence given the learned model parameters ω^* :
  - Dynamic programming:
    - Compute classification rate, considering as detections those DTW costs under the class thresholds.
  - Generative models:
    - Compute classification rate, considering as detections those probabilities above the class thresholds.

[1] Goldberg, D. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

Datasets

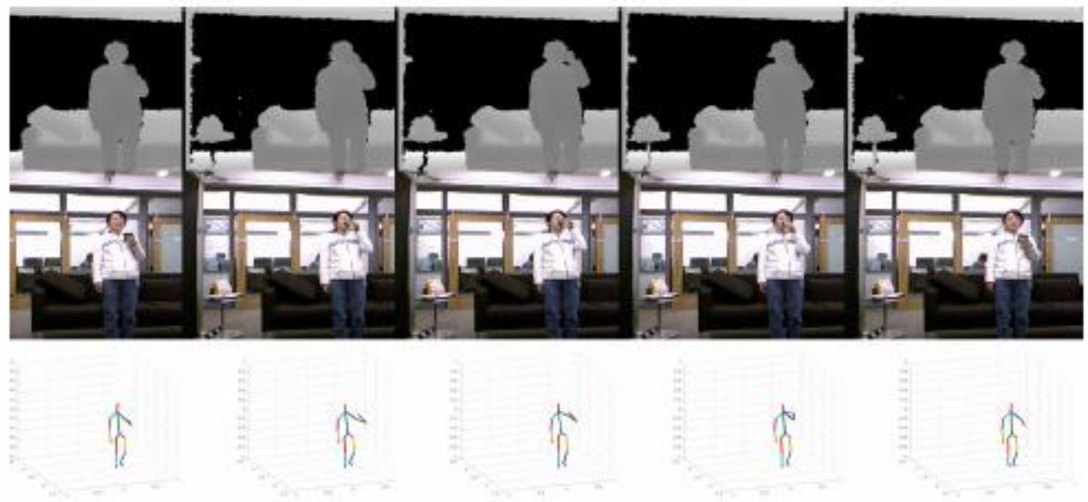
MSRAction3D [1]



- Depth Cuboid Similarity Features (DCSF) [2].
- Depth and skeleton information.

- 20 actions by 16 subjects.
- Half-training (subjects 1,3,5,7,9) / Half-testing (rest of subjects) [3].

MSRDaily3D [2]



- Depth Cuboid Similarity Features (DCSF) [2].
- RGB-D and skeleton information.

- Considered 12 out of 16 actions and half-subject split [2].
- 16 actions of daily activities using 5-fold cross-validation [4].

[1] Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3D points. In CVPRW, pages 9–14.
[2] Xia, L. and Aggarwal, J. K. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In CVPR, pages 2834–2841.
[3] Padilla-López, J. R., Chaaraoui, A. A., and Flórez-Revuelta, F. (2014). A discussion on the validation tests employed to compare human action recognition methods using the MSR action3d dataset. CoRR, abs/1407.7390.
[4] Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In CVPR, pp. 1290–1297.

Settings

- Framework implemented in [1]MATLAB/C++, including GA optimtool [2] and PMTK3 libs.

- Parameters of the method fixed to:
  - $P_s = 0.2$, $n_{min} = 5$ and $n_{max} = 25$ [3], population length $l = 20$ with 2 elitist members for the next generations, and $N = 500$ start-end segments.
  - $k_0 = 3$ minimum number of k clusters within the range $[k_0, k_f]$.
  - Number of iterations of k-meansDTW $\iota = 20$ to smooth its cost $\mathcal{O}(\iota \times k \times n^2)$.
  - Number of thresholds to learn Θ set to $T = 20$.

- DTW baseline consists of direct resizing all sequence examples of the same class with respect to the max-length sequence to get the representants $r_{c=1..g}$.

- HMM baseline splits each sequence into 3 fixed parts of the same length for learning subgesture class-models, where the number of clusters is the half of the total number of resulting segments.

[1] Library publicly available at https://github.com/vponcelo/Subgesture

[2] Goldberg, D. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.
[3] Zhou, F., De la Torre Frade, F., and Hodgins , J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE TPAMI, 35(3):582–596.

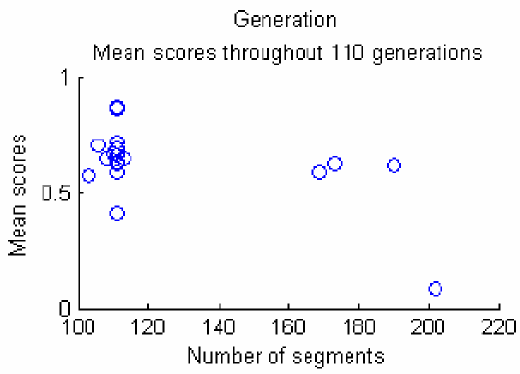
Results

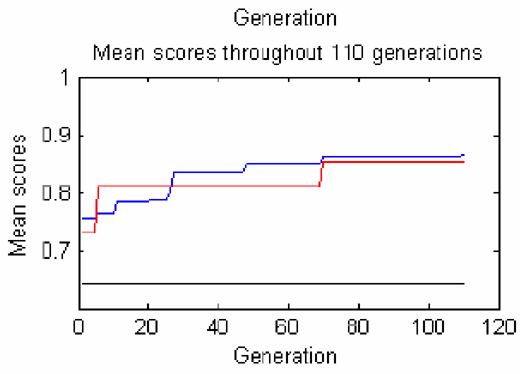
MSRAction3D-HS		MSRDaily3D-CV		MSRDaily3D-HS	
Method	Accuracy	**Method**	Accuracy	**Method**	Accuracy
[168] (LOP+J.)	88.2%	[71] (SOSVM)	68.3%	[167] (LOP)	42.5%
[175] (DCSF)	89.3%	[72] (SMMED)	73.20%	[112] (DTW)	54%
[130] (HOPC)	91.64%	[175] (DCSF)	83.60%	[168] (MKL)	80.0%
[49] (PBR)	92.3%	[175] (DCSF+Skl.)	88.2	[91] (GP)	85.6%
[169] (MMTW)	92.7%	-	-	[168] (LOP+J.)	85.75%
Dynamic Time Warping					
Baseline	85.76%	Baseline	77.36%	Baseline	70.20%
Evolved	90.89%	Evolved	**89.51%**	Evolved	**88.16%**
Hidden Markov Model					
Baseline	70.85%	Baseline	74.62%	Baseline	69.29%
Evolved	**95%**	Evolved	**91.39%**	Evolved	**92.30%**

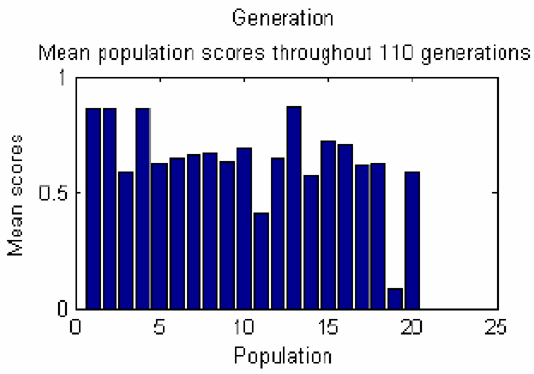
Recognition results in the MSRAction3D and MSRDaily3D datasets for half-split (HS) and cross-validation (CV), for the latter setting we report the 4 results available in published literature.

Results

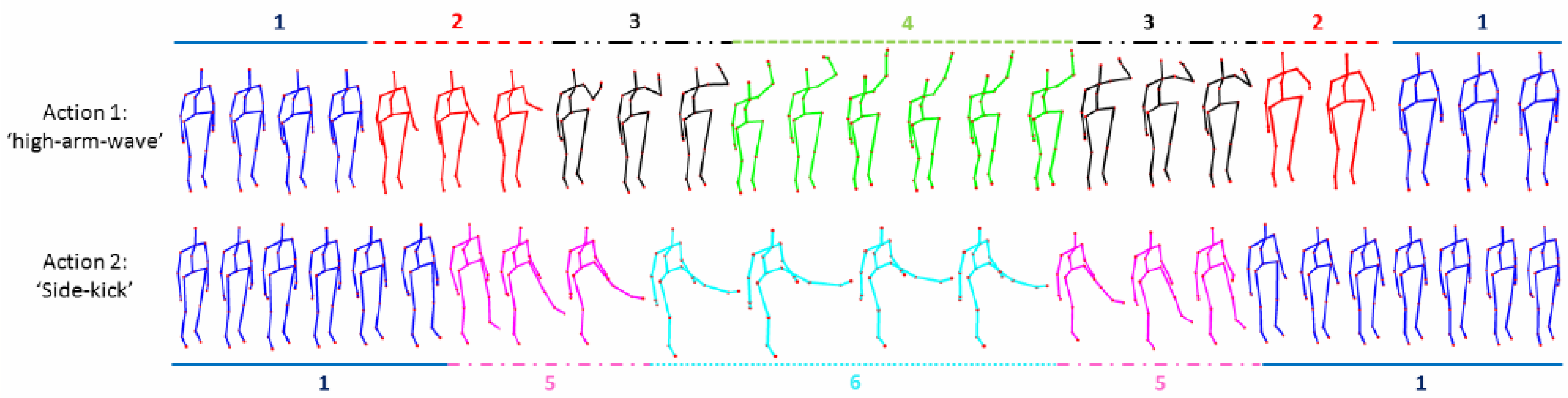
DTW & MSRDaily3D







DTW & MSRAction3D



Conclusions

Conclusion – Part 1

- Little research has been performed on TWS for computer vision. We introduced a novel methodology for learning weighting schemes to boost the performance of classification models relying on the BoVW:
  - ❑ Among traditional and alternative weighting schemes, the Boolean one obtained the highest performance.
  - ❑ Weighting schemes learned with our proposed approach outperformed consistently other weighting schemes in the considered datasets.
    - Schemes learned for some datasets do not generalize well in other datasets.
  - ❑ Among all of the considered terminals, three weighting schemes were used most often by solutions returned by the GP (TF, TF-IDF and TF-RF). However, the way in which the GP combined such primitives resulted in much better performance.

Conclusion – Part 2

- BoVDW approach for human gesture recognition presented using multimodal RGB-D images:
  - ❑ A new depth descriptor VFHCRH has been proposed, outperforming VFH.
  - ❑ Analyzed the effect of late fusion for combining RGB and Depth descriptors, obtaining better performance in comparison to early fusion.

- A Probabilistic-based DTW has been proposed to asses the temporal segmentation of gesture sequences and to be able to deal with multiple deformations present in data:
  - ❑ Different samples of the same gesture category modelled with Gaussian-based probabilistic models, encoding possible deformations.
  - ❑ Define a soft-distance based on the posterior probability of the GMM to embed probabilistic models into the DTW framework.

Conclusion – Part 3

- Introduce a novel approach for learning dynamic gesture primitives for gesture and action recognition.
- Evolutionary computation presents advantages when incorporating notable gesture methodologies based on dynamic programming and generative models in few generations.
- Results suggest that subgesture learning enhances the recognition of traditional techniques.

Future Work – Part 1

- Studying alternative methodologies for learning Term-Weighting Schemes:

  - ❑ Pose the problem as one of learning/optimizing the representation matrix, where other EA could be used.

  - ❑ Learning TWS for other domains, like audio, time series or accelerometer data.

- Explore the use of Genetic Programming frameworks for deep learning-based schemes.

Future Work – Part 2

- Including samples with different points of view of the same gesture class to analyze whether they fit using the proposed approaches.
- The definition of other powerful descriptors to obtain gesture-discriminative features.
- The use of Recurrent Neural Networks and Temporal Convolutions to learn spatio-temporal features using Deep Dynamic Neural Networks (DDNN) [1].

[1] D. Wu, L. Pigou, P.-J. Kindermans, N. Le, L. Shao, J. Dambre, and J.-M. Odobez. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. TPAMI 2016.

Future Work – Part 3

- Explore alternatives to the temporal clustering such as subgesture ranking to speed up the computational costs.
- Immediate work is to use representations learned with deep networks as input features for the method.
- Model subgesture primitives as part of deep dynamic neural networks:
  - ❑ Including them at the inner steps of the global optimization process made by the fitness function of the Evolutionary Algorithm.
  - ❑ Having several independent architectures for training subgestures from different data modalities.
  - ❑ Discovering subgestures through unsupervised deep learning and RNN [1].

[1] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. arXiv:1506.01911, 2016.

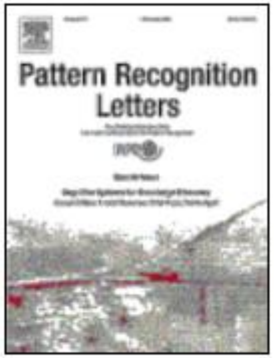
Publications

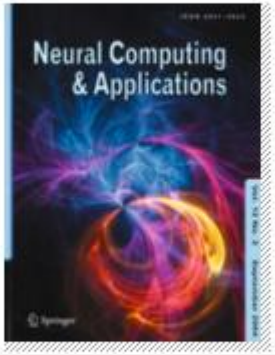
Research period: 2011 – 2016

Citations: 136; h-index: 6; i10-index: 5;
4 JCR journals (3 in Q1); 10 conference & workshop proceedings, 3 non-indexed technical reports.

Detailed info at http://sunai.uoc.edu/~vponcel/publications

sensors





IJCAI

acm ICMI

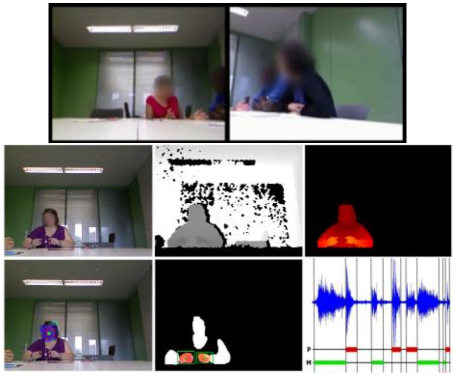
ECCV '14
Zürich, September 6-12th, 2014

Projects

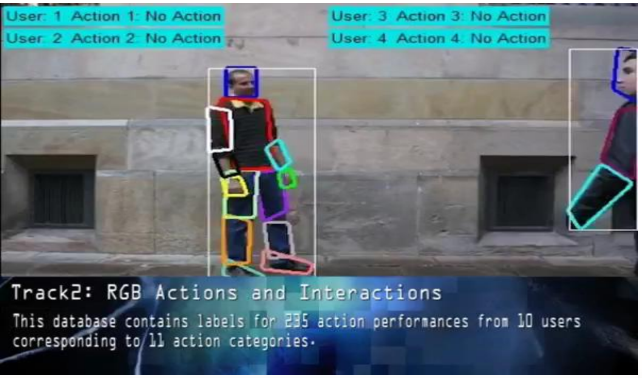
Detailed info at http://sunai.uoc.edu/~vponcel/research

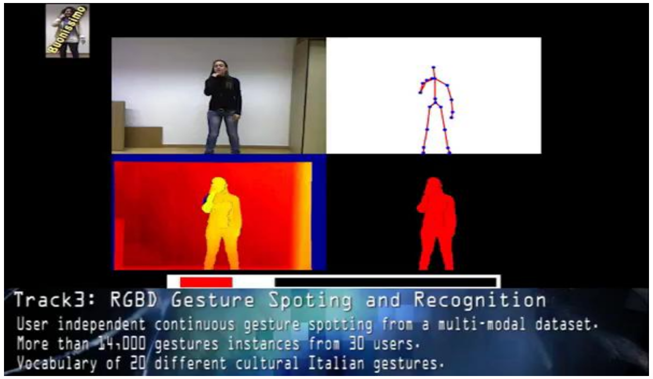
Generalitat de Catalunya
Centre d'Estudis Jurídics i Formació Especialitzada

Applications in Restorative Justice, real-case conversations.



Google
amazon Computer Vision
facebook
Microsoft Research
CHALEARN
NVIDIA
Disney Research





Thank You !