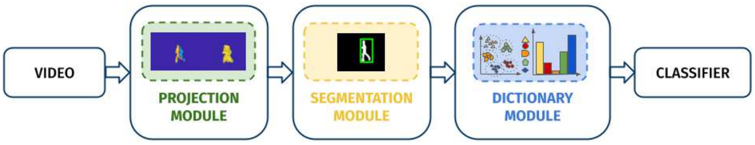
Università di Genova
DIPARTIMENTO INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

Computer Science Workshop
PhD program in Computer Science and Systems Engineering

Leveraging motion information and efficient projection kernels to represent human actions

Elena Nicora, Nicoletta Noceti

From **biological perception systems** able to quickly process huge amount of visual information focusing the attention on relevant part of the scene for a deeper understanding.



To **sustainable Computer Vision** where a single source of information is efficiently computed and exploited to gather details at different granularity about motion occurring in a scene.

Visual Saliency

State or quality of an item by which it stands out from its neighbors. Neuroscience identified saliency mechanism for a number of classes of visual stimuli that include color, orientation, depth and **motion**.

Notion of «center-surround differencing».

Motion detection

Low-level Computer Vision task that has the aim of identifying moving objects in image sequences. Typical first-step in many high-level pipelines in video surveillance, robotics, autonomous driving fields.
Most famous algorithms
(**background subtraction, change detection, optical flow**)
rely on pixel-intensity changes in time.

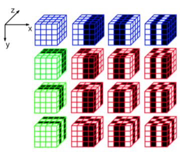
The Gray Code Kernels

Family of filters able to *approximate any desired kernel* that also comes with an **efficient projection scheme**.

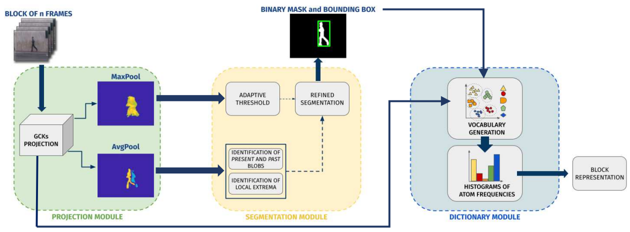
Filtering an image with a set of GCKs means that, after the first convolution, we are able to obtain the same result with a fixed number of summation per pixels (instead of a number of multiplications that depends on the size of the kernel).

Given a set of *M* Gray-Code Kernels and a kernel of size *n*x*n*x*n*:

- ❑ Classical full 3D convolution will require
  M (n³) multiplications per each pixel
- ❑ Separable full convoltions will require
  M (3n) operations per each pixel
- ❑ GCKs projection scheme will require
  one full convolution + (M-1)2 summations per pixel



Representation pipeline



Videos are processed in blocks:

1. **Projection Module** is responsible for the computation of the GCKs projections. Information included in the bank of results are then represented globally by means of two pooling strategies (max and average pooling).
2. **Segmentation Module** combines spatio-temporal information gathered from MaxPool and AvgPool (location, temporal development and direction of motion) and produces a segmentation mask of the moving object.
3. **Dictionary Module** takes in input spatio-temporal projections of the detected area in order to:
   a) Generate a dictionary of common features [only for the training set]
   b) Express each block as a histogram of atom frequencies

Experimental results

Human Action Classification

Weizmann Dataset full-body movement, fixed camera
SVM model + Gaussian Kernel

	Accuracy
Bend	88.63
Jumping Jack	97.75
Jump	69.23
Jump on the spot	82.48
Run	77.04
Side run	59.22
Skip	54.20
Walk	92.68
Wave (one hand)	92.52
Wave (two hands)	95.95
Mean accuracy	**80.97**

Salient motion detection and segmentation

SegTrackv2 dataset includes videos representing challenging Computer Vision scenarios (motion blur, complex deformations, occlusions and slow motion). Video can be categorized wrt camera movement.
We compared the ability of detecting and segmenting salient motion in a video of our pipeline with respect to classical motion detection video:

- **Gunnar-Farneback** Optical Flow
- **SIFT** based Optical Flow
- **Background subtraction**

Table below shows invariance with respect to camera motion achieving the best results both at pixel level and at bounding box level.

		Segmentation IoU				Bounding Box IoU			
		Farneback	SIFT	BS	GCKs	Farneback	SIFT	BS	GCKs
Fixed camera	birdfall	0.469	0.403	0.459	0.222	0.319	0.503	0.288	0.297
	worm	0.036	0.195	0.165	0.146	0.073	0.166	0.064	0.105
	hummingbird	0.419	0.437	0.643	0.341	0.759	0.756	0.870	0.669
	frog	0.361	0.506	0.53	0.243	0.469	0.578	0.549	0.48
	Mean IoU	0.321	0.385	**0.449**	0.238	0.405	**0.500**	0.442	0.387
Handheld camera	bird of paradise	0.293	0.406	0.193	0.286	0.494	0.560	0.551	0.633
	bmx	0.327	0.286	0.271	0.374	0.33	0.514	0.219	0.715
	penguin	0.119	0.278	0.069	0.202	0.557	0.568	0.588	0.566
	parachute	0.023	0.33	0.059	0.28	0.038	0.374	0.046	0.31
	Mean IoU	0.190	**0.325**	0.148	0.285	0.354	0.504	0.351	**0.556**
Dynamic camera	cheetah	0.029	0.069	0.151	0.336	0.097	0.17	0.098	0.4
	drift	0.011	0.005	0.136	0.419	0.16	0.16	0.183	0.348
	monkey	0.035	0.01	0.07	0.063	0.087	0.086	0.087	0.086
	monkeydog	0.048	0.069	0.068	0.114	0.165	0.188	0.142	0.199
	soldier	0.023	0.022	0.083	0.23	0.111	0.116	0.103	0.202
	girl	0.045	0.066	0.064	0.25	0.158	0.231	0.162	0.274
	Mean IoU	0.031	0.040	0.095	**0.235**	0.129	0.158	0.129	**0.251**
	Overall	0.159	0.220	0.211	**0.250**	0.272	0.355	0.282	**0.377**

[1] Nicora, Elena, et al. "The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions." *Scientific Data* 7.1 (2020): 1-15.
[2] Ben-Artzi, Gil, et al. "The gray-code filter kernels." *IEEE transactions on pattern analysis and machine intelligence* 29.3 (2007): 382-393.
[3] Gao, Dashan, et al. "The discriminant center-surround hypothesis for bottom-up saliency." Advances in neural information processing systems 20 (2007): 497-504.
[4] Cong, Runmin, et al. "Review of visual saliency detection with comprehensive information." *IEEE Transactions on circuits and Systems for Video Technology* 29.10 (2018): 2941-2959.
[5] Farneb¨ack, G.: Two-frame motion estimation based on polynomial expansion. In: Scan-dinavian conference on Image analysis. pp. 363–370. Springer (2003)
[6] Zhuo T., et al.: Unsupervised online videoobject segmentation with motion property understanding. IEEE Transactions on ImageProcessing29, 237–249 (2019)
[7] Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In:Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR2004. vol. 2, pp. 28–31. IEEE (2004)

Elena Nicora
elena.nicora@dibris.unige.it

MaLGa center – DIBRIS
Università degli Studi di Genova

UniGe | MaLGa

CSW