Motion-aware temporal median filtering for robust background estimation

Benjamin Laugraud

PhD presentation

March 03, 2020

Montefiore Institute
University of Liège
Belgium

Introduction to background estimation

What is background estimation?

Definition

Given an input **video sequence**, depicting the same scene at different times, **background estimation (BE)** consists in generating a **model of the scene background**, free of the **foreground elements** occluding it.

Example: Recovering a road without cars

$\overset{\text{BE}}{\longmapsto}$

Important note

In this work, the **expected** background model is actually a **background image**!

Static vs. dynamic background

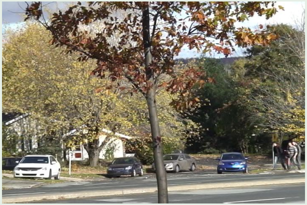
Static background

Composed of the background elements **remaining motionless** throughout the input video sequence.

Dynamic background

Composed of the background elements subject to small movements induced by **external factors** and/or containing **varying sub-elements**.

Examples



external factor: wind

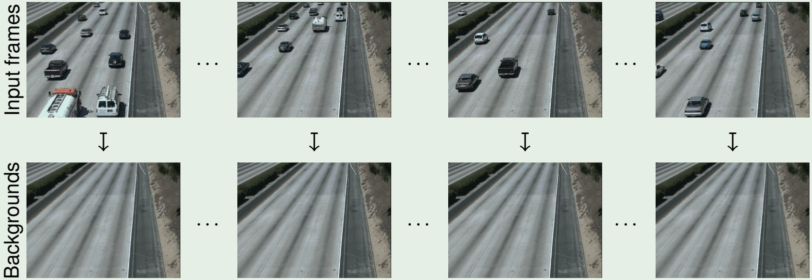
sub-element: ads

Note: Any background is accepted as long as it remains **consistent**!

Objective

Generating, **after each frame**, an image that estimates the **background of the frame**.

Example: Recovering a road without cars



The offline temporal median filter (TMF)

- Simplest and intuitive method $\rightarrow$ **baseline**.
- **Strong assumption** $\rightarrow$ background observable more than 50% in each position.

Given a grayscale video sequence $\gamma = F^1 F^2 \cdots F^T$ with $T \in \mathbb{N}^{>1}$:

$$\boxed{B_{x,y} = \text{median}\left(\{\{I^1_{x,y}, I^2_{x,y}, \ldots, I^T_{x,y}\}\}\right), \quad \forall (x,y) \in \Phi,}$$

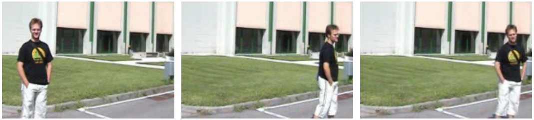
How to deal with colors?

Marginal median $\rightarrow$ classical median per color component.

Symbols

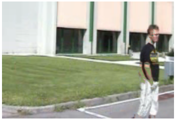
- $B_{x,y} \rightarrow$ background value at position (x,y)
- $I^t_{x,y} \rightarrow$ pixel intensity at position (x,y) in frame F^t
- $\Phi \rightarrow$ image domain
- $\{\{.\}\} \rightarrow$ multiset notation (note that $\{\{40, 42, 42\}\} \neq \{40, 42, 42\} = \{40, 42\}$)

TMF: Toy example

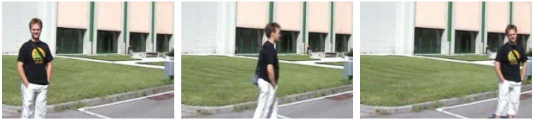


Three input frames

$\xrightarrow{\text{TMF}}$



TMF fails ✘



Three input frames

$\xrightarrow{\text{TMF}}$

TMF succeeds ✔

Applications & challenges

- **Several applications** → video surveillance, tracking, couting, compression, etc.

Computational photography

Input frame(s)

LaBGen output

- Not as easy as it looks → there are **several challenges**! (Jodoin et al. 2017)
  - **Very short** seq. → difficult with **motion detection** and/or **long training**.
  - **Very long** seq. → mix **different challenges** for testing **versatility**.

Challenge: Cluttered sequences

- Several pixels depict foreground elements **more than 50% of the time**.
- The background is **not in the most redundant temporal information**.

Input frame(s)

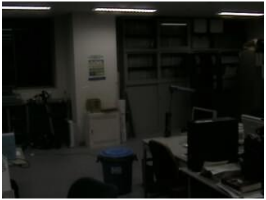
Expected output

Typical failure: ghosts

- **Lighting conditions evolve** over time.
- Several backgrounds are possible.


Input frame(s)


Typical failure: inconsistent


Expected output 1

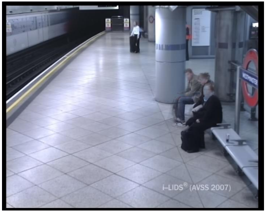

Expected output 2

Challenge: Intermittent motion

- **Foreground** elements **stopping** to move.
- **Background** elements **starting** to move.

Input frame(s)

Expected output



Typical failure: ghosts

Challenge: Jitter

- BE methods should be robust against **small camera jitter**.
- The background is **also in motion** $\rightarrow$ **compensate** camera motion.

Input frame(s)

Expected output

Typical fail.: incons. & blur

Challenge: Background motion

- Difficult for BE methods with **strong stationarity assumptions**.
- When **several backgrounds** are possible → the final one must be **consistent**.
- Ghost free and consistent backgrounds can be **smoothed**.


Input frame(s)

Expected output

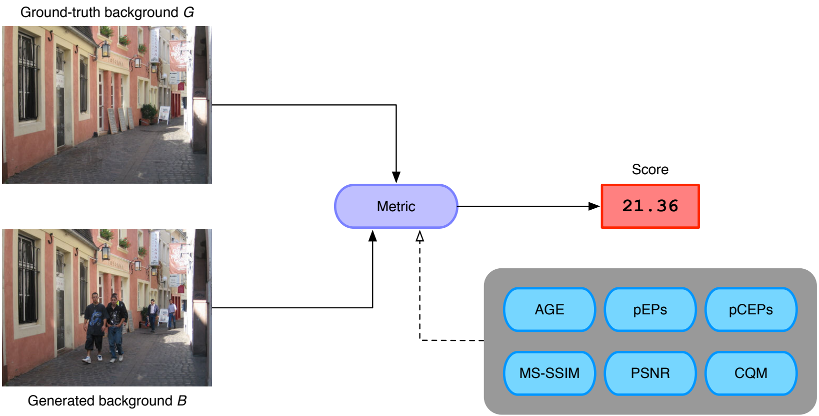

Typical failure: inconsistent


Input frame(s)


Expected output


Typical failure: smoothed

Performance evaluation (offline)



- **Not a classification** problem!
- **IQA** metrics.
- The **ones used in BE** have been suggested by Maddalena et al. 2015b.

Metric: Percentage of error pixels (pEPs)

$$E_{x,y} = \mathbb{1}_{\mathbb{N}>20}\left(\left|B_{x,y} - G_{x,y}\right|\right).$$

$$\text{pEPs (\%)} = \frac{100}{W \times H} \cdot \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} E_{x,y}.$$

Background image B

Ground truth G

Error map E

Symbols

- W / H $\rightarrow$ width / height
- $B_{x,y}$ / $G_{x,y}$ $\rightarrow$ background / ground-truth intensity at position (x, y)

Metric: Color image quality measure (CQM)

Observation

There are **more rods** (*i.e.*, light sensors) **than cones** (*i.e.*, color sensors) in the eye

Assumption

Scores should be determined on luminance Y and chrominance U & V components and **combined according to the proportions of rods and cones** (Yalman et al. 2013).

Per-component PSNRs weighted by the **proportions of rods and cones**:

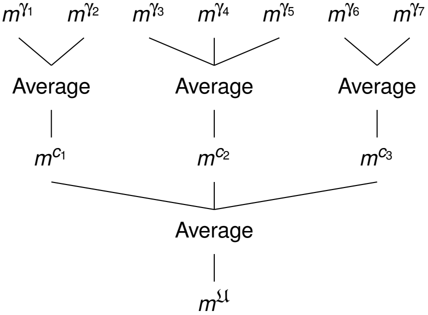
$$\begin{aligned} \text{PSNR (dB)} &= 10 \cdot \log_{10}\left(\frac{255^2}{\text{MSE}}\right), \\ \text{CQM (dB)} &= (0.9449 \cdot \text{PSNR}_\text{Y}) + 0.0551 \cdot \left(\frac{\text{PSNR}_\text{U} + \text{PSNR}_\text{V}}{2}\right). \end{aligned}$$

We apply metrics on **public datasets** (*i.e.*, collections of **video sequences** with **GT**).

SBI (Maddalena et al. 2015a)		**SBMnet** (Jodoin et al. 2016)	
✘	**14 video** sequences	✔	**79 video** sequences
✘	**1 GT image *G*** for a sequence	✔	**Several GTs G^i** when needed
✔	GT for **all sequences**	✘	GT for **13 sequences**
✘	**Imperfect** GTs	✘	**Imperfect** GTs
✘	**No categories**	✔	**8 categories** (1 per challenge)

Universal and category-specific scores

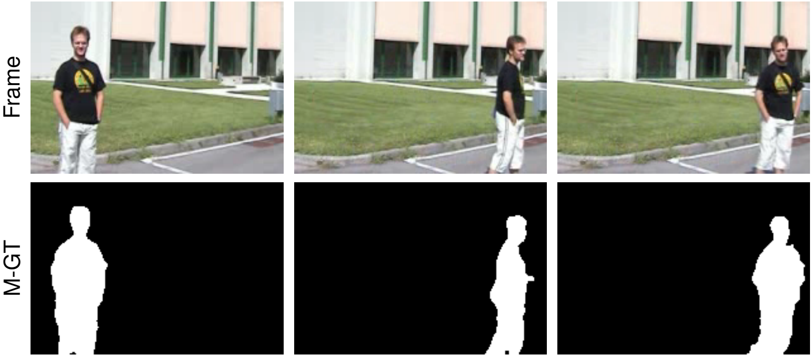
- Given metric m applied to a **video sequence** $\rightarrow$ **scene-specific score** m^{γ}.
- Expected perf. on sequences **captured in similar conditions**.
- **Category-specific score** m^{c} $\rightarrow$ expected perf. on seq. with **specific challenge**.
- **Universal score** $m^{\mathfrak{U}}$ $\rightarrow$ expected perf. on **any video sequence**.



- **Comparing BE methods** $\rightarrow$ **compare scores** as long as it is consistent...
- ...or use **different rankings** proposed by Jodoin et al. 2017.

Median-based background estimation methods leveraging motion detection

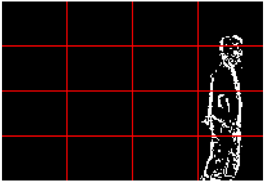
Idea 1: Median filtered by motion



Temporal median filter

- Motion detection could be made by **background subtraction** (BGS) algorithms.
- In practice, they are **not perfect**.
- **Idea:** Aggregate motion classifications in spatial areas.

Frame



Segmentation map

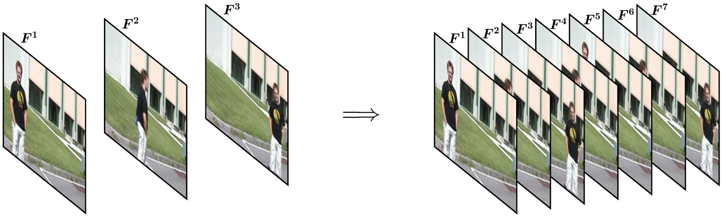
0	0	0	7
0	0	0	30
0	0	0	45
0	0	0	35

Quantities of motion $\in [0, 255]$

Our method called LaBGen is based on these two ideas!

- BGS algorithms can require a **long training**.
- In BE, the **sequences can be short**.
- **Augment input** video sequence in $\mathcal{P}$ **passes**.
- **Odd** passes $\rightarrow$ **forward** & **even** passes $\rightarrow$ **backward** (smooth transitions).

$$\underbrace{F^1 \cdots F^T}_{\gamma} \quad \mapsto \quad \underbrace{F^1 \cdots F^T \left(F^{T-1} \cdots F^1 F^2 \cdots F^T\right)^{\left\lfloor \frac{\mathcal{P}-1}{2} \right\rfloor} \left(F^{T-1} \cdots F^1\right)^{\mathcal{P}-1 \bmod 2} = F^1 \cdots F^{T^\alpha}}_{\gamma^\alpha}.$$

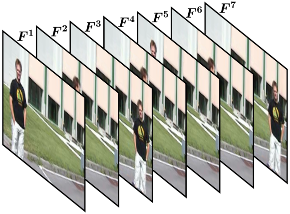


Input video sequence γ

Augmented sequence γ^α ($\mathcal{P} = 3$)

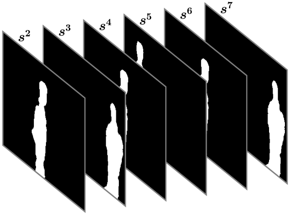
- **Objective:** Information to filter out **foreground pixels** from median buffers.
- Motion detection made by **any BGS algorithm** $\mathcal{A}$.

$$s^t_{x,y} = \begin{cases} 1 & \text{if } p^t_{x,y} \in \text{foreground}, \\ 0 & \text{if } p^t_{x,y} \in \text{background}. \end{cases}$$



Augmented sequence γ^α
($\mathcal{P} = 3$)

$\Longrightarrow$

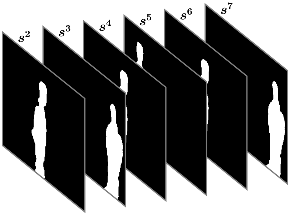


Segmentation maps s^t
($\mathcal{A} =$ M-GT)

Step 3: Estimation step

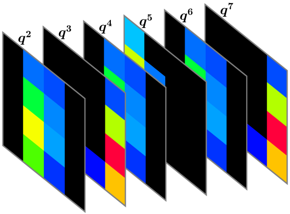
- **Objective:** Estimate a **quantity of motion $\boldsymbol{q_i^t}$** per patch f_i^t.
- Domain Φ divided into $\mathcal{E}^2$ **rectangular spatial areas $\boldsymbol{\Psi_i}$** of $\approx W/\mathcal{E} \times H/\mathcal{E}$ px.

$$q_i^t = \sum_{(x,y)\in\Psi_i} s_{x,y}^t ,$$



Segmentation maps s^t
($\mathcal{A} =$ M-GT)

$\Longrightarrow$

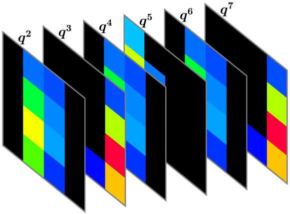


Quantities of motion q^t
($\mathcal{E} = 4$)

- **Objective:** Select in each Ψ_i the $\mathcal{S}$ patches f_i^t with the **smallest QOMs $\boldsymbol{q_i^t}$**.
- Builds iteratively, for each Ψ_i, a **submultiset of patches $\boldsymbol{\Omega_i}$**:

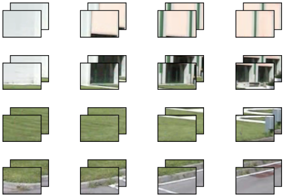
$$\Omega_i^t = \begin{cases} \emptyset & \text{for } t = 1, \\ \Omega_i^{t-1} \uplus \left\{\left\{ f_i^t \right\}\right\} & \text{for } t = 2, 3, \ldots, \mathcal{S}+1, \\ \Omega_i^{t-1} \uplus \left\{\left\{ f_i^t \right\}\right\} \setminus \left\{\left\{ f_i^\beta \right\}\right\} & \text{if } t > \mathcal{S}+1 \wedge q_i^t \leq q_i^\beta, \\ \Omega_i^{t-1} & \text{otherwise,} \end{cases}$$

$$\beta = \min \underset{\left\{t' \middle| f_i^{t'} \in \Omega_i^{t-1}\right\}}{\arg\max}\; q_i^{t'}, \qquad \Omega_i = \Omega_i^{T^\alpha}.$$



Quantities of motion q^t ($\mathcal{E} = 4$)

$\Longrightarrow$



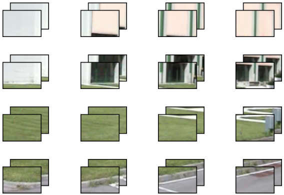
Submultisets Ω_i ($\mathcal{S} = 2$)

Step 5: Generation step

- **Objective:** Generating the **background image *B*** (or background images B^t).
- **Combining the patches** selected in the different submultisets Ω_i.
- Combination performed with a **pixel-wise median filter**.

online:
$$B^t_{x,y} = \text{median}\left(\left\{\left\{ I^{t'}_{x,y} \,\middle|\, \exists i : (x,y) \in \Psi_i \wedge f^{t'}_i \in \Omega^t_i \right\}\right\}\right),$$

offline:
$$B_{x,y} = \text{median}\left(\left\{\left\{ I^{t}_{x,y} \,\middle|\, \exists i : (x,y) \in \Psi_i \wedge f^{t}_i \in \Omega_i \right\}\right\}\right).$$



Submultisets Ω_i ($\mathcal{S} = 2$)

$\Longrightarrow$

Background image B

Universal performance on SBI

- Minimizing **universal pEPs** w. r. t. 12 BGS $\mathcal{A}$ (Barnich et al. 2011; Elgammal et al. 2000; Goyat et al. 2006; Heikkilä et al. 2006; Hofmann et al. 2012; Maddalena et al. 2008; Manzanera et al. 2004; Stauffer et al. 1999; Wren et al. 1997; Zivkovic 2004; St-Charles et al. 2015).
- Regardless of $\mathcal{A} \rightarrow$ LaBGen **outperforms the TMF**.
- Surprise! The **frame difference** leads to the **best universal** performance.

Parameter values				Universal pEPs $\downarrow$
$\mathcal{A}$	$\mathcal{P}$	$\mathcal{E}$	$\mathcal{S}$	
F. diff.	**29**	**4**	**57**	**1.3972**
GMM	17	2	41	2.4494
SuBSENSE	17	3	3	2.6674
PBAS	1	2	1	3.4566
VuMeter	1	2	27	3.9232
KDE	5	2	177	4.3616
AGMM	29	50	7	5.1848
LBP	1	1	39	5.2596
ViBe	23	∞	7	5.8089
Σ-Δ	3	2	7	6.1998
Pfinder	1	2	23	7.3764
SOBS	29	∞	9	9.4441
Temporal median filter				14.0500

Good visual results

CAVIAR2

CaVignal

HumanBody2

IBMtest2

Scene-specific performances on SBI

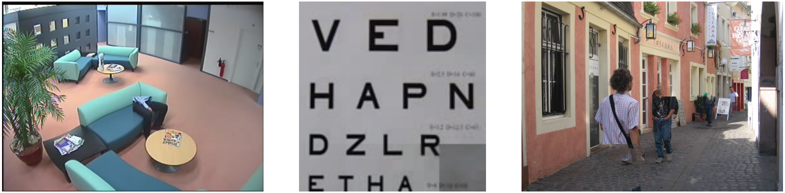
- Minimizing **scene-specific pEPs** with all **parameters free**.
- **No consensus** on which BGS $\mathcal{A}$ is the best but **F. diff. chosen for 5 seq**.
- **Good results** with $\mathcal{A} =$ **F. diff.** and $(\mathcal{P}, \mathcal{E}, \mathcal{S})$ free (mean Δ pEPs $\approx 0.03\%$).

	Parameter values				Scene-specific
Sequence	$\mathcal{A}$	$\mathcal{P}$	$\mathcal{E}$	$\mathcal{S}$	**pEPs** $\downarrow$
Board	KDE	1	5	99	0.3201
Candela_m1.10	Σ-Δ	1	4	1	0.0000
CAVIAR1	KDE	3	17	149	0.0661
CAVIAR2	F. diff.	1	13	5	0.0000
CaVignal	F. diff.	1	2	1	0.0147
Foliage	F. diff.	1	1	1	0.0000
Hall&Monitor	SOBS	1	15	111	0.0178
HighwayI	GMM	1	1	37	0.0612
HighwayII	AGMM	1	5	123	0.0143
HumanBody2	PBAS	23	5	51	0.0521
IBMtest2	Pfinder	3	5	71	0.0000
People&Foliage	F. diff.	1	3	1	0.0013
Snellen	F. diff.	1	1	1	0.0048
Toscana	SuBSENSE	3	13	7	0.4850

The frame difference provides the best BGS for LaBGen!

Universal vs. scene-specific (visual comparison)

Universal



$\Downarrow$

Scene-specific

Enough degrees of freedom for correct backgrounds on a small dataset!

LaBGen-P: A pixel-level variant of LaBGen

- LaBGen subject to the **patch effect** (*i.e.*, discontinuities between patches).
- The captain obvious' idea: **Do not use patches**, but...
- ...**insufficient information at pixel level** $\rightarrow$ quantities of motion.

LaBGen

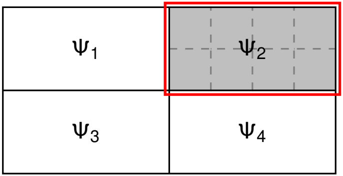
LaBGen-P

Ground truth

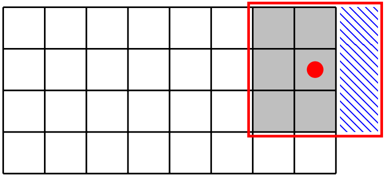
Idea

Select pixels by taking into account the motion information available on the **spatial neighborhood**.

Pixel-level quantities of motion



Positions considered in **LaBGen**



Positions considered in **LaBGen-P**

$$q_{x,y}^{t} = \sum_{(x',y') \in \Psi_{x,y}} m_{x',y'}^{t},$$

$$\Psi_{x,y} = \left\{ (x',y') \;\middle|\; \begin{array}{l} \max\left(x - \lfloor \mathcal{W}/2 \rfloor, 0\right) \leq x' \leq \min\left(x + \lfloor \mathcal{W}/2 \rfloor, W-1\right) \wedge \\ \max\left(y - \lfloor \mathcal{W}/2 \rfloor, 0\right) \leq y' \leq \min\left(y + \lfloor \mathcal{W}/2 \rfloor, H-1\right) \end{array} \right\},$$

$$\mathcal{W} = 1 + 2 \cdot \left\lfloor \frac{\min(W,H)}{2\mathcal{E}} \right\rfloor .$$

Results on SBMnet with the frame difference

Method	Rank across cat. $R_{<}^{AR^{C}} \downarrow$
MSCL	**1**
SPMD	2
BEWiS	3
LaBGen	4
LaBGen-P	5
TMF	6

30 methods in the rankings

Category	Cat. rank $R_{<}^{AR^{c}} \downarrow$	
	LaBGen-P	**LaBGen**
Basic	**7**	8
Intermittent Motion	**5**	10
Clutter	**13**	15
Jitter	**5**	7
Illumination Changes	10	**5**
Background Motion	**13**	**13**
Very Long	19	**10**
Very Short	11	**9**

−

- Ranked **after LaBGen**
- Worse than LaBGen in **3 categories**
- **Illumination** → No mechanisms
- **Jitter** → No mechanisms
- **Intermittent** → F. diff. unsuited
- **Very Short** → F. diff. unsuited

+

- LaBGen(-P) is **state of the art**
- **FPS:** 126 LaBGen-P & 1312 LaBGen
- **Simpler** and **faster** than competitors (*e.g.*, MSCL: 0.6 FPS)
- Improves LaBGen in **3 categories**

? We have **no explanations** for the bad ranks for ***Very Long***!

LaBGen

$\Downarrow$

LaBGen-P

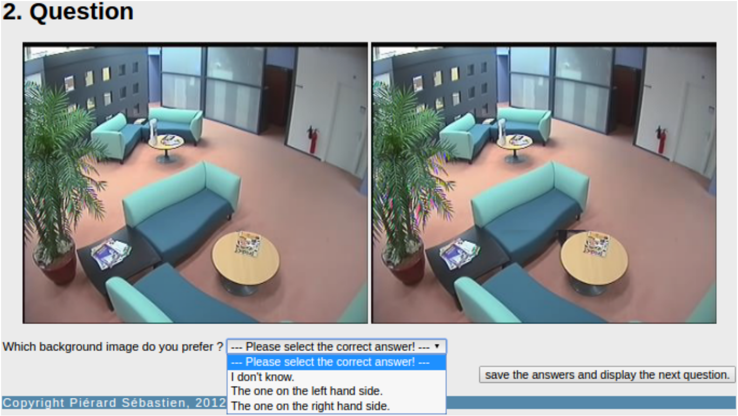
Although not perfect, we prefer LaBGen-P to LaBGen...

Subjective evaluation: Questions

1. Video for which we would like to define a background image

video/Candela_m1.10.m4v

- We had **35 volunteers**.
- One question for each of the **79 SBMnet video sequences**.



- Most people **undecided for 38 sequences** over 79.
- **LaBGen-P** preferred for **26 sequences** & **LaBGen** preferred for **15 sequences**.

LaBGen-P was preferred by the volunteers for more sequences!

On the importance of a temporally memoryless motion detection

Motivation

- Although it is the worst, the **frame difference is the most useful** in LaBGen.
- **No (obvious) correlation** between **BGS** perf. and **LaBGen** perf.
- It has an **unshared property** $\rightarrow$ it is **temporally memoryless**.
- Temporal **history ignored** $\rightarrow$ never influenced by its **past errors**.

Hypothesis

The **most relevant motion detection** algorithms for LaBGen are **temporally memoryless**.

A simple motion detection memory model

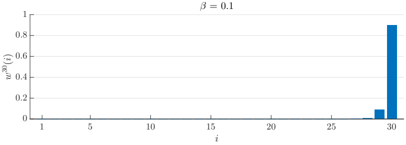
The exponential filter (EF)

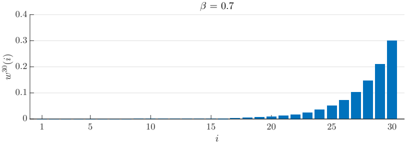
$$\mathcal{B}^t_{x,y} = (1-\beta)\cdot I^t_{x,y} + \beta\cdot\mathcal{B}^{t-1}_{x,y},$$
$$= \sum_{i=1}^{t} w^t(i)\cdot I^i_{x,y}.$$

$$w^t(i) = \begin{cases} \beta^{t-1} & \text{if } i = 1, \\ (1-\beta)\cdot\beta^{t-i} & \text{if } i > 1. \end{cases}$$

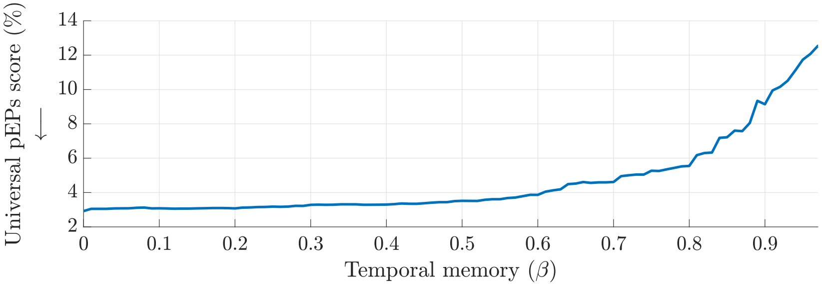
- **Infinite** memory.
- **Oldest** intensities have **insignificant weights**.

- The parameter $\beta \in [0,1]$ enables to **tune the amount of memory**.
- $\beta = 0 \Leftrightarrow$ **frame difference** (no memory).





Measure the impact of **temporal memory** on the LaBGen **universal performance**.

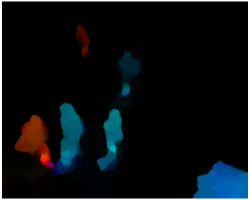


Remove temporal memory!

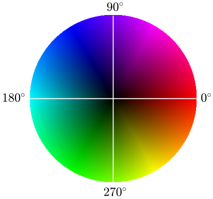
- Other **temporally memoryless** algorithms!
- Determine a **vector field $\mathbf{v}^t$** known as **optical flow**.
- **For each pixel** in F^t gives the **displacement vector** to F^{t+1}.

Image F^t

Image F^{t+1}



Optical flow $\mathbf{v}^t$



Color wheel

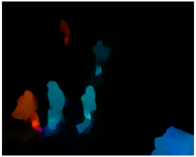
From LaBGen to LaBGen-OF

Modifying the motion step

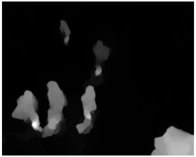
$$s_{x,y}^{t} = \begin{cases} 1 \text{ (foreground)} & \text{if } n_2\left(\mathbf{v}^t(x,y)\right) > \tau, \\ 0 \text{ (background)} & \text{otherwise.} \end{cases} \qquad n_2\left(\mathbf{v}^t(x,y)\right) = \frac{\left\|\mathbf{v}^t(x,y)\right\|_2}{\max\limits_{(x,y)\in\Phi} \left\|\mathbf{v}^t(x,y)\right\|_2}$$

- **Hard threshold** on magnitudes τ.
- $n_2(.)$ aims at reducing the **sensibility to video scaling** without changing τ.

A frame F^t



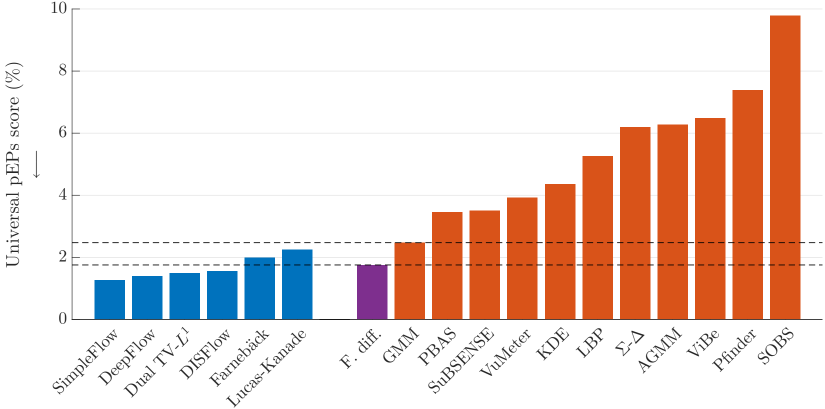
Optical flow $\mathbf{v}^t$



Normalized $n_2(.)$

Seg. map s^t

- **6 optical flow algorithms** (Bouguet 2001; Farnebäck 2003; Kroeger et al. 2016; Lucas et al. 1981; Tao et al. 2012; Weinzaepfel et al. 2013; Zach et al. 2007).
- **Universal performance on SBI** minimized with respect to $\mathcal{A}$.
- The ones **without memory** **vary around** the ones of the **frame difference**.
- Any $\mathcal{A}$ **without memory** is **better** than any $\mathcal{A}$ **with memory**.



Method	**Rank across cat.** $R_{<}^{AR^C}$ $\downarrow$
MSCL	**1**
LaBGen-OF	2
SPMD	3
BEWiS	4
LaBGen	5
LaBGen-P	6

30 methods in the rankings

Category	**Cat. rank** $R_{<}^{AR^c}$ $\downarrow$	
	LaBGen-OF	**LaBGen**
Basic	**4**	8
Intermittent Motion	**8**	10
Clutter	**1**	15
Jitter	**1**	7
Illumination Changes	12	**5**
Background Motion	**10**	13
Very Long	**3**	10
Very Short	**7**	9

−

- **Illumination** → Still no mechanisms
- **Intermittent** → Still no mechanisms
- **Background Motion**
- **Very Short**
- **FPS:** 1312 vs. 5 (CPU)

+

- Ranked **#2**
- Still **simpler** and **faster** than MSCL
- Better than LaBGen in **7/8 cat.**
- Ranked #1 in **Clutter**
- Ranked #1 in **Jitter**

Some visual improvements

LaBGen

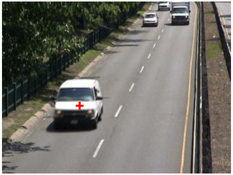
$\Downarrow$

pEPs = 0.05%!

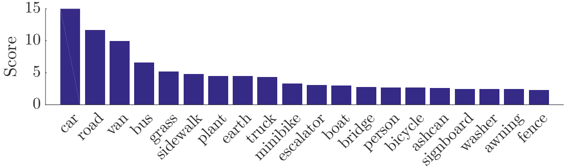
LaBGen-OF

Intra-frame motion detection leveraging semantic segmentation

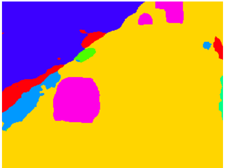
Semantic segmentation algorithms



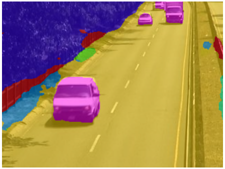
Input frame



Top-20 of the scores associated with the red cross



Semantic seg. map



Alpha blending

Motivation

Spatial features increase robustness against **intermittent motion**, **background motion**, and **very short** sequences.

Intra-frame motion detection algorithms

- Given a pixel $p \rightarrow \mathbf{s} = [s(1), s(2), \ldots, s(N)]^\top$ **vector of semantic scores**.
- $\mathrm{softmax}(\mathbf{s}) = \mathbf{u} = [u(1), u(2), \ldots, u(N)]^\top$.
- In proba. $\rightarrow o_i$ is the **real object class**.

$$\textbf{CV algorithm:} \quad P(\mathrm{FG} \mid \mathbf{s}) = \sum_{i=1}^{N} \underbrace{P(\mathrm{FG} \mid o_i)}_{\text{estimators}} \cdot \underbrace{P(o_i \mid \mathbf{s})}_{\approx u(i)}.$$

- **Scene-specific** estimators on a **sequence** $\boldsymbol{\gamma}$ with M-GT or...
- ...**universal** estimators on a **dataset** $\boldsymbol{\Gamma}$.
- Other method from the **most probable object class** $\hat{o} = o_j : \arg\max_j \mathbf{s}(j)$.
- **Assumption:** $P(\mathrm{FG} \mid \mathbf{s}) = P(\mathrm{FG} \mid \hat{o})$.

$$\textbf{MP algorithm:} \quad P(\mathrm{FG} \mid o_i).$$

Visual comparison

Frame

Frame difference

DeepFlow

CV+U

CV+S

MP+U

MP+S

From LaBGen-P to LaBGen-P-Semantic

- **Add a parameter** $\mathcal{V}$ to choose between CV and MP.
- We use **PSPNet** (Zhao et al. 2017) that is trained to recognize **150 objects**.

$$m_{x,y}^{t} = \begin{cases} P\left(\mathrm{FG} \,\middle|\, \mathbf{s}\right) & \text{if } \mathcal{V} = \mathrm{CV}, \\ P\left(\mathrm{FG} \,\middle|\, o_i\right) & \text{if } \mathcal{V} = \mathrm{MP}. \end{cases}$$

- Comparisons on SBI $\rightarrow$ we have **both M-GT and B-GT**.
- Maximizing **universal CQM** on SBI with respect to the different $\mathcal{V}$.

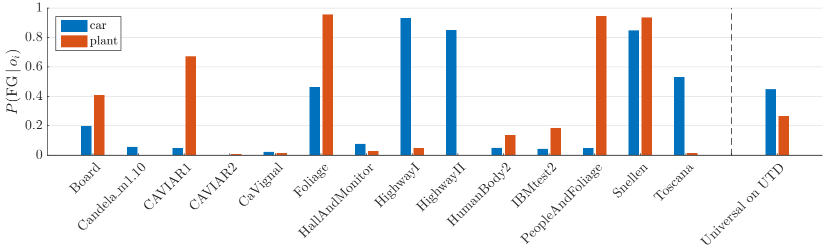
Algorithm $\mathcal{V}$	**Best parameter values for a given $\mathcal{V}$ and universal CQM scores $\uparrow$**			
	CV+U $(\mathcal{E},\mathcal{S}) = (43,94)$	**CV+S** $(\mathcal{E},\mathcal{S}) = (1,42)$	**MP+U** $(\mathcal{E},\mathcal{S}) = (1,48)$	**MP+S** $(\mathcal{E},\mathcal{S}) = (1,54)$
CV+U	34.1356	33.6932	33.8508	33.1893
CV+S	36.6552	**36.9663**	36.8952	36.5239
MP+U	32.9081	33.3753	33.5372	33.1257
MP+S	35.2960	36.3784	36.4490	**36.4883**

CV+S $\succ$ MP+S $\succ$ CV+U $\succ$ MP+U

Input sequence

Foliage: CV+U

Foliage: CV+S



- Could happen with **scene-specific** estimators on **complex scenes**.
- Probabilities $P(\mathrm{FG}\,|\,o_i)$ are **similar if an object moves or not**.

Some limitations: Information loss with MP

- The appropriate object class may have a **high score without being first**.
- Lost with MP, but **taken into account in CV**!

Frame

CV+S

MP+S

Method	Rank across cat. R^{AR^C} ↓
MSCL	**1**
LaBGen-OF	2
SPMD	3
BEWiS	4
CV+U	5
LaBGen	6
MP+U	7
LaBGen-P	8

Category	Cat. rank $R_{<}^{AR^c}$ ↓		
	CV+U	MP+U	LaBGen-P
Basic	**6**	10	7
Intermittent M.	**1**	2	5
Clutter	**5**	6	13
Jitter	8	10	**5**
Illumination C.	23	28	**10**
Background M.	2	**1**	13
Very Long	27	23	**19**
Very Short	**1**	2	11

30 methods in the rankings

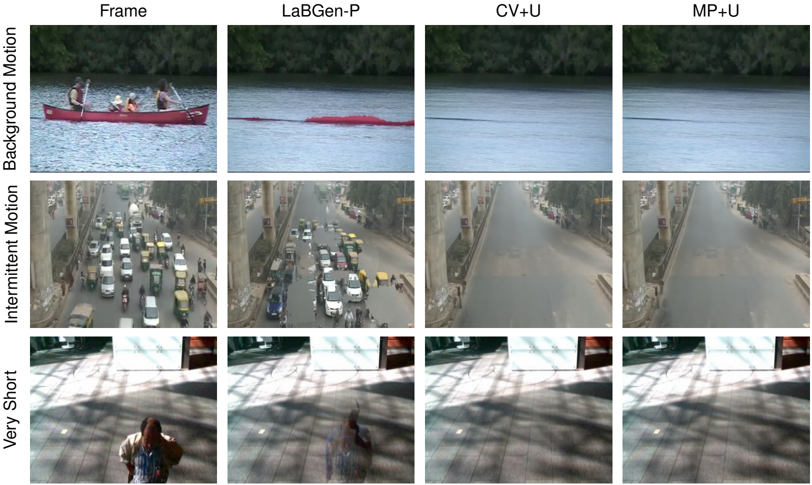
−

- **Illumination** → Still no mechanisms
- Still **below LaBGen-OF** → But metrics and rankings problems
- High **running time**

+

- Better than LaBGen-P in **5 cat.**
- Ranked #1 in **Intermittent Motion**
- Ranked #1 in **Background M.** → But intrinsic smoothing limitation
- Ranked #1 in **Very Short**

Some visual improvements



Conclusion

Contributions

1. New **median-based BE methods** leveraging motion detection.
2. Determination of the **best motion detection paradigms**.
3. New semantic-based **intra-frame motion detection algorithms**.
4. Not presented: Insights into **performance evaluation of online methods**.
5. Not presented: Detailed **discussion on performance evaluation** tools.
6. Not presented: Open-source **C++ implementations** (excluding semantic).

LaBGen

- LaBGen outperforms TMF $\rightarrow$ coupling **motion detection to TMF is relevant**.
- **Frame difference** is the most effective $\rightarrow$ **LaBGen 4/30** on SBMnet and faster.

LaBGen-P

- Made to avoid the **patch effect** as much as possible.
- **Selects pixels** regarding the motion information in the **spatial neighborhood**.
- Although ranked after LaBGen, a **subjective evaluation** proved that it is **better**.

LaBGen-OF

- A simple memory model showed that **temporal memory is undesired**.
- **Optical flow algorithms** $\rightarrow$ **LaBGen-OF**.
- On SBI, LaBGen-OF with **any OF** $\succ$ LaBGen with any $\mathcal{A}$ **with memory**.
- With DeepFlow $\rightarrow$ **LaBGen-OF 2/30** on SBMnet (and **1 in *Clutter* and *Jitter***).
- Let's go further $\rightarrow$ **intra-frame motion detection algorithms**.

LaBGen-P-Semantic

- Better ranked than LaBGen-P on SBMnet but **below LaBGen-OF**.
- However, **first in *Background Motion*, *Intermittent Motion*, and *Very Short***.
- Spatial features **insensitive to perturbations** induced by these challenges!
- Promising, but **some limitations** could be raised with a **temporal information**.

We believe that a *temporally hybrid* motion detection could be ideal!

Perspectives

- Develop the **temporally hybrid** paradigm.
- PhD focused on motion step $\rightarrow$ **improve the other steps**.
- Considering **pre/post-processing** (*e.g.*, registration, intensity adjustments).
- Fast and embedded **implementations**.
- A **combining** approach:

Category	**Best variant**	$R_{<}^{AR^c}$ $\downarrow$
Basic	LaBGen-OF	4
Intermittent Motion	LaBGen-P-Semantic (CV+U)	1
Clutter	LaBGen-OF	1
Jitter	LaBGen-OF	1
Illumination Changes	LaBGen	5
Background Motion	LaBGen-P-Semantic (MP+U)	1
Very Long	LaBGen-OF	3
Very Short	LaBGen-P-Semantic (CV+U)	1

Demonstration

شكراً جزيلاً

thank you
danke
謝謝
gracias
merci
obrigado
teşekkür ederim
go raibh maith agat
ngiyabonga
спасибо
bedankt
dziękuję
sukriya
kop khun krap
terima kasih
grazie
arigatô
dakujem
mochchakkeram
tapadh leat
dank je
hvala
mauruuru
sagolun
takk
taiku
ευχαριστώ
감사합니다
ありがとう
мерси
рахмат
Баярлалаа
faafetai lava
kiitos
dankie
dhanyavad
nanni
nandri
bayarlalaa
gràcie
köszönöm
enkosi
sobodi
dėkuji
mèsi
didi madloba
najis tuke
kam sah hamnida
rahmat
xièxie
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
mercé
trugarez
chnorakaloutioun
gratias ago
gràcies
sulpáy
akun dankon
aciú
vinaka
спасибі
blagodaram
mersi
kia ora
barka
welalin
tack
spas
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
murakoze
tenki
chokrane
mamnun
djiere dieuf
tau
дякую

Backup slides

Online constraints

- In the litterature (Bouwmans et al. 2017; Jodoin et al. 2017) $\rightarrow$ online methods are defined as **recursive methods**.

- **Too restrictive** according to us $\rightarrow$ memory constraint vs. methodology.

- Our constraints:

  1. **Compactness:** Must sufficiently summarize the input in its internals.
  2. **Real time:** Must be computationally efficient and run in real time.

Important note

Applying an **offline BE method** on the frames 1 to *i* to estimate the background of frame *i* **does not transform** this method into an online method!

Categories of methods

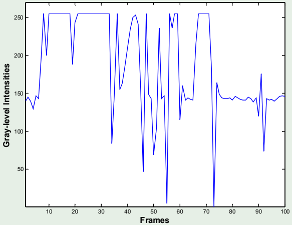
BE methods can be grouped into **categories** $\rightarrow$ represent the **methodology**.

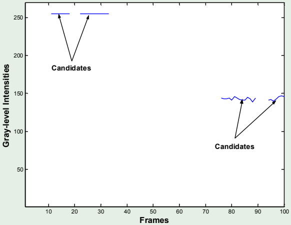
Temporal statistics (TS)

- Based on statistics (*e.g.*, mean, median) computed on **temporal information**.
- Statistics computed pixel-wise on the **whole sequence or random frames**.
- **Note:** Our methods are TS methods.

Subsequences of stable intensity (SSI)

- **Strong assumption:** The background has the **longest stable intensity**.
- Stable temporal subsequences are **located**, and the most reliable is **chosen**.
- **Unrealistic assumption** (*e.g.*, intermittent motion).





(Wang et al. 2006)

Optimal labeling (OL)

- Find a **labeling** $\mathscr{L} : \Phi \to \{1, 2, \ldots, T\}$
- Determined over a MRF with a **spatio-temporal energy** function.

Input frames

Labeling $\mathscr{L}$

Generated background

(Granados et al. 2008)

Neural networks (NN)

- **Learn** automatically the background **from the data**.
- The learning can be **supervised** or **unsupervised**.

Categories of methods (3)

Iterative model completion (IMC)

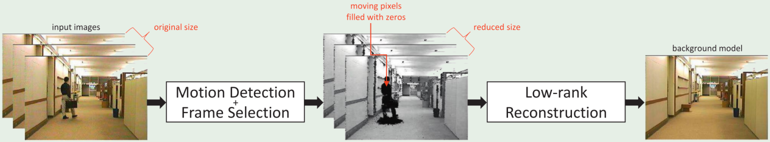
- Highly reliable **spatial areas** are kept in a **partial background** image.
- **Remaining areas** completed according to **spatial smoothness criteria**.

Partial back. Iterative completions of empty spatial areas

(Mseddi et al. 2019)

Missing data reconstruction (MDR)

- **Foreground elements** are first identified and **removed**.
- The **missing parts** are reconstructed through **matrix/tensor completion**.



(Sobral et al. 2017)

$$\text{AGE} = \frac{1}{W \times H} \cdot \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \left| B_{x,y} - G_{x,y} \right| .$$

Background image B

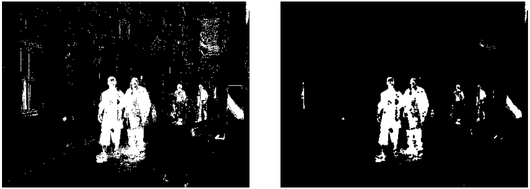
Ground truth G

Gray-level errors

Metric: Percentage of clustered error pixels (pCEPs)

$$\text{pCEPs}\ (\%) = \frac{100}{W \times H} \cdot \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} E'_{x,y}.$$

$$E'_{x,y} = \bigwedge_{z=-1}^{1} \left(E_{x+z,y} \wedge E_{x,y+z} \right) = E \ominus \begin{bmatrix} 0 & \color{red}{1} & 0 \\ \color{red}{1} & \color{red}{1} & \color{red}{1} \\ 0 & \color{red}{1} & 0 \end{bmatrix}.$$



Error map E Error map E'

Metric: Peak signal-to-noise ratio (PSNR)

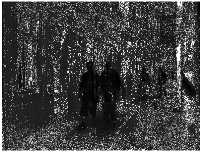
$$\text{PSNR (dB)} = 10 \cdot \log_{10}\left(\frac{L^2}{\text{MSE}}\right) = 20 \cdot \log_{10}\left(\frac{L}{\text{RMSE}}\right).$$

$$\text{MSE} = \frac{1}{W \times H} \cdot \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \left(B_{x,y} - G_{x,y}\right)^2 .$$

Background B

Ground truth G

Pixel RSEs



Pixel PSNRs

Metric: Structural similarity index (SSIM)

Assumptions

- HVS highly adapted to extract **structural information** (Wang et al. 2004a).
- Illumination and structural information are **independent**.
- Structural similarity should be a good **approximation of the perceived quality**.

$$\mathrm{SSIM} = l \cdot c \cdot s,$$

$$l = \frac{2 \cdot \mu_B \cdot \mu_G + C_1}{\mu_B^2 + \mu_G^2 + C_1}, \quad c = \frac{2 \cdot \sigma_B \cdot \sigma_G + C_2}{\sigma_B^2 + \sigma_G^2 + C_2}, \quad s = \frac{\Sigma_{B,G} + C_3}{\sigma_B \cdot \sigma_G + C_3}.$$

Symbols

- $\mu_B, \mu_G \rightarrow$ mean intensity in B and G
- $\sigma_B, \sigma_G \rightarrow$ standard deviation in B and G
- $\Sigma_{B,G} \rightarrow$ covariance of B and G
- $C_1, C_2, C_3 \rightarrow$ small constants to avoid divisions by zero

SSIM in practice

- Distortions are **not space-invariant**.
- **Local SSIMs** are computed using a 11×11 Gaussian **sliding window**.
- Let SSIM_i being the i-th local SSIM, then: $\text{SSIM} = \frac{1}{N}\sum_{i=1}^{N} \text{SSIM}_i$.

Background image B

Ground truth G

SSIM-map

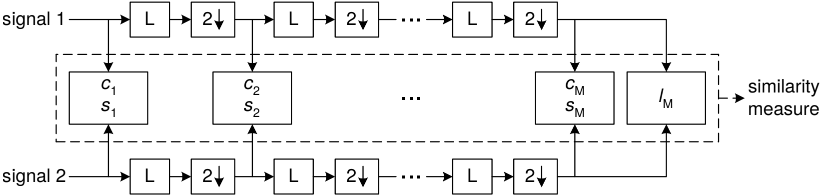
l-map

c-map

s-map

Assumption

SSIM computed at a single scale but the right one **depends on viewing conditions** such as resolution, distance, etc (Wang et al. 2003a).



Background image B

Ground truth G

Maps for levels 1, 2, 3, 4, and 5

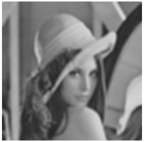
Problems related to these metrics

- **Statistical metrics** do not consider any characteristic modeled from the **HVS**.

Original

Contrast stretching

Compression



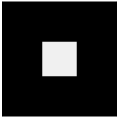
Blurring

MSE = 225 (Wang et al. 2003b)

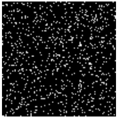
- They can be more **meaningful** if they are **interpreted together**.

AGE = 2.7
pEPs = 9
pCEPs = 7.8

AGE = 2.7
pEPs = 9
pCEPs = 0



AGE = 21.6
pEPs = 9
pCEPs = 7.8



AGE = 21.6
pEPs = 9
pCEPs = 0

- **SSIM** seems to be **related to MSE** (Dosselmann et al. 2011).
- **No consensus** on which metric is the best!
- At least in BE, **no methodology** enables to choose one metric over another.

“Homemade” datasets

SBI+SBMnet-GT

- **Objective:** Build the largest dataset with GT (*e.g.*, to learn parameters).
- Composed of **26 video sequences** (13 SBI + 13 SBMnet).
- Better but **remains small**...

Universal Training Dataset (UTD)

- **Objective:** Build the largest dataset with **motion ground truth** (M-GT).
- Composed of **60 video sequences**.

Comparing methods through rankings

- **Average cat. rank** AR^c $\rightarrow$ **Aggregates** m^c output by **all metrics** for a given c.
- **Average rank across cat.** $AR^{\mathcal{C}}$ $\rightarrow$ **Aggregates** AR^c for **all categories**.
- Ordering by AR^c or $AR^{\mathcal{C}}$ gives the **actual ranks**.

		Category-specific scores				**Rankings**		
Metric		$m_1^c \downarrow$		$m_2^c \downarrow$		$AR^{c_1} \downarrow$	$AR^{c_2} \downarrow$	$AR^{\mathcal{C}} \downarrow$
Category		c_1	c_2	c_1	c_2			
Method	α	**5.60**	**4.78**	**5.65**	**5.29**	**2**	**2.5**	**2.25**
	β	**6.67**	**5.45**	**6.90**	**1.80**	**3**	**2**	**2.5**
	δ	**1.40**	**1.50**	**1.84**	**2.47**	**1**	**1.5**	**1.25**

- *AR* or $AR^{\mathcal{C}}$?
- **Unreliable** (*e.g.*, a new method can modify a relative order).
- Some metrics are **strongly correlated** according to SBMnet data, examples:

$$\begin{array}{rcl} \text{AGE and pEPs} & \rightarrow & \rho = 0.97 \\ \text{pEPs and pCEPs} & \rightarrow & \rho = 0.96 \\ \text{PSNR and CQM} & \rightarrow & \rho = 1! \end{array}$$

- Strong correlations between **universal scores and rankings**:

ρ	AGE	pEPs	pCEPs	PSNR	CQM	MS-SSIM
AR	0.98	0.96	0.88	0.95	0.95	0.92
$AR^{\mathcal{C}}$	0.92	0.89	0.79	0.97	0.97	0.90

  - Optimize **AGE** to optimize **average rank**.
  - Optimize **PSNR** or CQM to optimize **average rank across categories**.

Example of a ranking instability

		Metric m						$AR\downarrow$
		AGE ↓	pEPs ↓	pCEPs ↓	PSNR ↑	CQM ↑	MS-SSIM ↑	
α	$m^{\mathfrak{U}}$	6.7090	6.31	2.65	28.6396	29.4668	0.9266	1.50
	$R^{m^{\mathfrak{U}}}$	1	1	2	1	1	3	
β	$m^{\mathfrak{U}}$	7.0738	7.06	3.19	28.4660	29.3196	0.9278	2.67
	$R^{m^{\mathfrak{U}}}$	3	3	3	3	3	1	
δ	$m^{\mathfrak{U}}$	6.7778	6.71	2.27	27.9944	28.8810	0.9196	2.84
	$R^{m^{\mathfrak{U}}}$	2	2	1	4	4	4	
ε	$m^{\mathfrak{U}}$	7.3890	7.61	3.57	28.5050	29.3829	0.9267	3.00
	$R^{m^{\mathfrak{U}}}$	4	4	4	2	2	2	

		Metric m						$AR\downarrow$
		AGE ↓	pEPs ↓	pCEPs ↓	PSNR ↑	CQM ↑	MS-SSIM ↑	
α	$m^{\mathfrak{U}}$	6.7090	6.31	2.65	28.6396	29.4668	0.9266	2.00
	$R^{m^{\mathfrak{U}}}$	1	1	2	2	2	4	
β	$m^{\mathfrak{U}}$	7.0738	7.06	3.19	28.4660	29.3196	0.9278	3.67
	$R^{m^{\mathfrak{U}}}$	4	4	4	4	4	2	
δ	$m^{\mathfrak{U}}$	6.7778	6.71	2.27	27.9944	28.8810	0.9196	3.34
	$R^{m^{\mathfrak{U}}}$	2	2	1	5	5	5	
ε	$m^{\mathfrak{U}}$	7.3890	7.61	3.57	28.5050	29.3829	0.9267	4.00
	$R^{m^{\mathfrak{U}}}$	5	5	5	3	3	3	
η	$m^{\mathfrak{U}}$	6.8514	6.89	2.69	28.9450	29.7995	0.9387	2.00
	$R^{m^{\mathfrak{U}}}$	3	3	3	1	1	1	

Evaluating online BE methods

- Previous methodologies **dedicated to offline** methods.
- To date, there is **no methodology to assess online** BE methods!

Problem

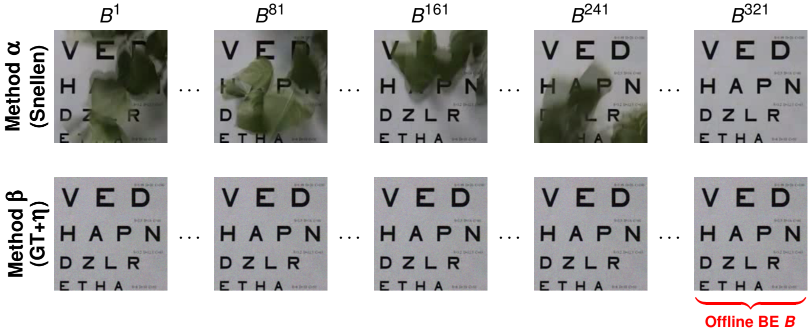
- Consider a **video sequence** $\gamma = F^1 F^2 \cdots F^T$.
- **Offline** BE methods evaluated on **the background** image $B^T = B$.
- What about the background images generated for **all the previous frames**?

Our two proposals

1. **Aggregating IQA** scores.
2. Using Full-Reference Video Quality Assessment (**VQA**) metrics.

Note: They remain compatible with aggregations and rankings!

Proof of concept



- **Offline:** More **noise** with β → $\boxed{\alpha \succ \beta}$.
- **Online:** Correct **noisy estimations** with β & **strong ghosts** with α → $\boxed{\beta \succ \alpha}$.

- **Averaging IQA scores** applied after each frame.
- Yields the **highest correlation** with **subjective scores** (Netflix: Li et al. 2018).
- Suppose that the K first frames are used for **training**:

$$\begin{aligned}
\text{pEPs}_{\text{online}} &= \frac{1}{T-K} \cdot \sum_{t=K+1}^{T} \text{pEPs}\left(F^t, G^t\right), \\
\text{CQM}_{\text{online}}\ (\text{dB}) &= 10 \cdot \log_{10}\left(\frac{1}{T-K} \cdot \sum_{t=K+1}^{T} 10^{\frac{\text{CQM}\left(F^t, G^t\right)}{10}}\right).
\end{aligned}$$

		Scene-specific scores	
		pEPs $\downarrow$	**CQM** $\uparrow$
Method α **(Snellen)**	**Offline**	0.0000	48.3636
	Online	58.5234	24.2629
Method β **(GT+**η**)**	**Offline**	0.0000	36.5866
	Online	0.0001	36.5855

Spoiler alert

It also works with VQA metrics!

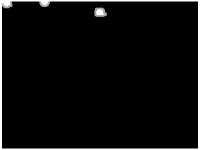
- **VSSIM: Weighted average of SSIMs** determined on Y, Cb, and Cr components (Wang et al. 2004b).
- **VQM:** Mainly designed for **broadcasting**. Linear **combination of "parameters"** measuring the perceptual effect of an impairment (Pinson et al. 2004).
- **MOVIE:** Spatio-spectrally localized multiscale evaluation based on a **model of the visual cortex** (Seshadrinathan et al. 2010).
- **VMAF: Combine** the strenghts of different metrics **through SVM** (Li et al. 2016).

Metric	**Best**	**Values**	**Method α (Snellen)**	**Method β (GT+η)**
VSSIM	$\uparrow$	$[-1, 1]$	0.6244	0.8907
VQM	$\downarrow$	$[0, > 1]$	1.1306	0.0289
MOVIE	$\downarrow$	$\mathbb{R}_+$	0.0306	$2.7769 \cdot 10^{-5}$
VMAF	$\uparrow$	$[0, 100]$	16.4222	94.7228

Augmentation step efficacy

Frame 1

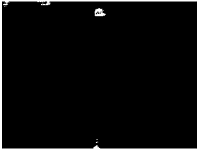
Frame 319



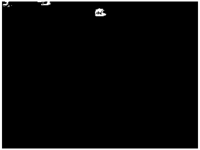
M-GT

$\mathcal{P}^1$ (PWC $= 2.29\%$)

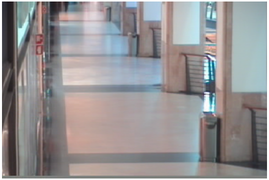
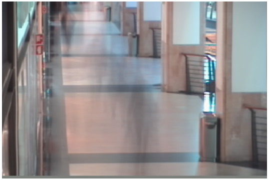
$\mathcal{P}^2$ (PWC $= 1.03\%$)



$\mathcal{P}^3$ (PWC $= 0.41\%$)



$\mathcal{P}^5$ (PWC $= 0.15\%$)

	CAVIAR1	Hall&Monitor	People&Foliage
LaBGen Pfinder	0.19% 	0.23% 	0.73% 
Model Pfinder	8.78% 	2.88% 	63.40% 

t^{α} :	1	2	3	4	5	6	7	8	9	10	11	12	13
t :	1	2	3	4	3	2	1	2	3	4	3	2	1
$\mathcal{P}^{t^{\alpha}}$:	1	1	1	1	2	2	2	3	3	3	4	4	4
$\rightleftharpoons$:	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
FΔ :	✘	|2-1|	|3-2|	|4-3|	|4-3|	|3-2|	|2-1|	|2-1|	|3-2|	|4-3|	|4-3|	|3-2|	|2-1|

Given any submultiset Ω_i, increasing $\mathcal{P}$ by keeping the same $\mathcal{S}$ leads to:

1. If $\mathcal{S}$ patches with q_i^- in a pass $\rightarrow \Omega_i$ has the $\mathcal{S}$ last patches with q_i^-.
2. If $\mathcal{S} > T \rightarrow \gamma$ entirely duplicated into Ω_i until the pass that totally populates Ω_i.
3. Let f_i^- be the multiset of patches with q_i^- in a forward and/or backward pass:
   - If $\left|f_i^-\right| < \mathcal{S} \rightarrow f_i^-$ will be duplicated in Ω_i over time.
   - Depending on $\left|f_i^-\right|$, other patches with $q_i > q_i^-$ are ejected during the first passes.
   - When Ω_i is entirely populated with patches $\in f_i^-$, it is composed of $\left\lfloor \mathcal{S}/|f_i^-| \right\rfloor$ duplicates of $f_i^- +$ a duplicate of the $\mathcal{S}$ mod $\left|f_i^-\right|$ most recently observed patches in f_i^-.
   - If $\left(\mathcal{S} \bmod \left|f_i^-\right|\right) \neq 0$, the content of Ω_i depends on the direction of the last pass.

Comparison of LaBGen to other algorithms on SBI

		BE methods										
		TMF (baseline)	BE-AAPSA	WS2006	RSL2011	Photomontage	LRGeomCG	TMac	BEWiS	LaBGen (default)	LaBGen (scene-s.)	**Mean sequence pEPs**
Scene-specific pEPs scores ↓	Board	23.765	**0.290**	6.095	5.308	2.412	35.954	36.241	2.217	2.729	0.320	11.533
	Candela_m1.10	3.332	0.012	1.905	0.375	3.584	0.658	1.015	0.793	1.681	**0.000**	1.336
	CAVIAR1	0.350	**0.009**	0.126	0.160	0.133	6.627	6.702	0.459	0.633	0.066	1.527
	CAVIAR2	**0.000**	**0.000**	0.039	0.131	**0.000**	0.327	0.329	**0.000**	**0.000**	**0.000**	0.083
	CaVignal	10.485	4.810	1.500	**0.015**	11.221	6.412	6.507	**0.015**	**0.015**	**0.015**	4.100
	Foliage	47.705	59.980	2.851	12.309	**0.000**	20.892	21.983	0.017	**0.000**	**0.000**	16.574
	Hall&Monitor	0.979	0.320	0.556	1.649	0.361	0.224	0.242	1.435	0.130	**0.018**	0.591
	HighwayI	0.163	2.760	0.685	0.234	0.408	0.202	0.199	0.466	0.436	**0.061**	0.561
	HighwayII	0.332	0.280	0.488	0.513	0.589	0.356	0.376	0.414	0.303	**0.014**	0.367
	HumanBody2	0.327	0.080	0.639	0.310	13.005	4.650	4.729	1.501	0.263	**0.052**	2.556
	IBMtest2	0.033	0.001	1.953	2.701	0.069	1.454	1.483	1.501	0.087	**0.000**	0.928
	People&Foliage	36.009	31.000	3.572	9.402	0.004	57.781	57.189	13.018	0.003	**0.001**	20.798
	Snellen	62.235	76.080	23.167	14.429	33.497	50.434	51.997	5.276	6.337	**0.005**	32.346
	Toscana	10.985	**0.103**	5.894	27.379	0.452	11.857	12.016	6.888	6.944	0.485	8.300
Universal pEPs score ↓		14.050	12.552	3.534	5.351	4.695	14.131	14.358	2.429	1.397	**0.074**	
Universal rank ↓		8	7	4	6	5	9	10	3	2	**1**	

- Ranked **#2** in **universal** → #1 for 3/14 sequences.
- Ranked **#1** in **scene-specific** → #1 for 11/14 sequences (worst rank: #3).
- Significantly **above average** (even when it is high).
- **Outperforms other methods** (putting apart the very short *Toscana*).

1. **Motion step / BGS:** Applied to each pixel $\rightarrow \mathcal{O}\left(\mathcal{P}TWH\mathcal{A}\right)$.
2. **Estimation step:** Requires iterating each pixel $\rightarrow \mathcal{O}\left(\mathcal{P}TWH\right)$.
3. **Selection step:** $\mathcal{S}$ comparisons per spatial area in the worst case and 1 in the best $\rightarrow \mathcal{O}\left(\mathcal{P}T\mathcal{S}\mathcal{E}^2\right)$ or $\mathcal{O}\left(\mathcal{P}T\mathcal{E}^2\right)$.
4. **Generation step / Pixel-wise median:** Using *Introselect* (Musser 1997) linear computation $\rightarrow \mathcal{O}\left(\mathcal{S}\right)$ per pixel position (offline) or $\mathcal{O}\left(\mathcal{S}\right)$ per pixel (online with $\mathcal{P}=1$) $\rightarrow \mathcal{O}\left(WH\mathcal{S}\right)$ (offline) or $\mathcal{O}\left(TWH\mathcal{S}\right)$ (online).

LaBGen		Best case	Worst case
Time	Offline	$\mathcal{O}\left(\mathcal{P}T\cdot\left(WH\mathcal{A}+\mathcal{E}^2\right)+WH\mathcal{S}\right)$	$\mathcal{O}\left(\mathcal{P}T\cdot\left(WH\mathcal{A}+\mathcal{E}^2\mathcal{S}\right)+WH\mathcal{S}\right)$
	Online	$\mathcal{O}\left(T\cdot\left(WH\mathcal{A}+\mathcal{E}^2+WH\mathcal{S}\right)\right)$	$\mathcal{O}\left(T\cdot\left(WH\mathcal{A}+\mathcal{E}^2\mathcal{S}+WH\mathcal{S}\right)\right)$
Space		$\mathcal{O}\left(\mathcal{S}\left(\mathcal{E}^2+WH\right)\right)$	

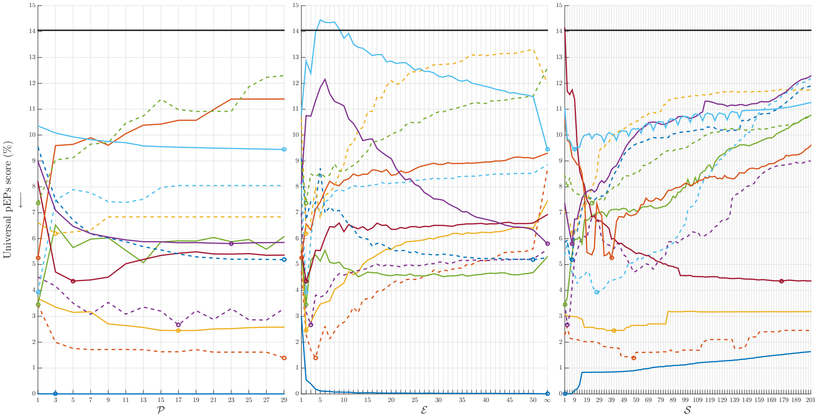
- Mean pixel **throughput** $\approx 239\cdot 10^6$ px/s.
- Requires to store $\mathcal{S}\mathcal{E}^2$ **quantities of motion and patches** ($=\mathcal{S}WH$ patches) $\rightarrow \mathcal{O}\left(\mathcal{S}\left(\mathcal{E}^2+WH\right)\right) \rightarrow \mathcal{S}\cdot\left(32\cdot\mathcal{E}^2+24\cdot WH\right)$ bits.
- With **default parameter values**: $800\times 600 \sim$ 86Mo & $4K \sim$ 1.5Go.

LaBGen running times

Sequence	Dimensions $W \times H$	Frames T	Time (ms) Default	Default $\mathcal{P}=1$	Default $\mathcal{E}=1$	Default $\mathcal{S}=1$	Default $\mathcal{E}=\infty$	Default $\mathcal{S}=201$
Board	200×164	228	958	124	506	485	992,452	3,420
Candela_m1.10	352×288	350	6,439	570	2,322	2,109	5,383,488	28,087
CAVIAR1	384×256	610	19,721	938	6,988	3,731	9,836,691	71,923
CAVIAR2	384×256	460	15,396	820	4,719	2,862	7,404,927	59,682
CaVignal	200×136	258	1,806	89	1,003	548	918,378	11,143
Foliage	200×144	394	1,008	126	707	706	1,077,841	2,719
Hall&Monitor	352×240	296	1,987	386	1,582	1,346	3,791,332	6,180
HighwayI	320×240	440	3,013	376	2,010	1,844	4,202,820	9,262
HighwayII	320×240	500	2,432	399	2,339	2,028	5,566,130	5,453
HumanBody2	320×240	740	7,836	530	3,219	3,258	8,590,112	30,218
IBMtest2	320×240	90	624	222	613	366	999,054	2,550
People&Foliage	320×240	341	2,181	317	1,660	1,425	3,282,700	6,702
Snellen	144×144	321	718	83	449	462	641,250	2,023
Toscana	800×600	6	1,220	66	1,142	135	264,243	3,235

Important note

$\mathcal{P}$ has a bigger impact than $\mathcal{E}$ that has a bigger impact that $\mathcal{S}$.

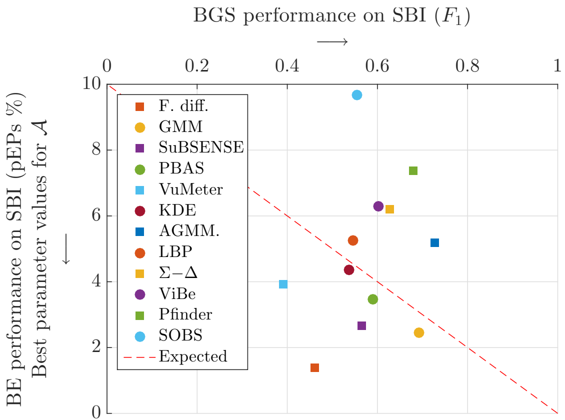


Oracle (—), F. diff. (---), GMM (—), SuBSENSE (---), PBAS (—), VuMeter (---), KDE (—), AGMM (---), LBP (—), Σ-Δ (---), ViBe (—), Pfinder (---), SOBS (—)

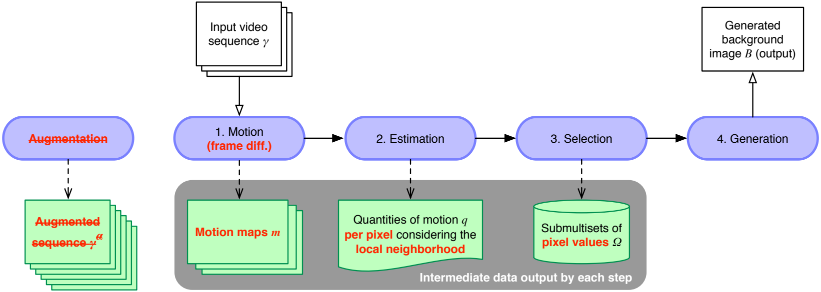
Important note

In general $\mathcal{S}$ has a bigger impact than $\mathcal{E}$ that has a bigger impact that $\mathcal{P}$.

(Non-)correlation between BE and BGS performance



- **Expectation:** BE perf. increases when BGS perf. increases.
- **Best BGS** algorithm (AGMM) → **7-th best BE**...
- **2-nd worst BGS** algorithm (F. diff.) → **Best BE!**
- Other metrics does not help...



Input video sequence γ
Generated background image B (output)
~~Augmentation~~
1. Motion (frame diff.)
2. Estimation
3. Selection
4. Generation
~~Augmented sequence γ^α~~
Motion maps m
Quantities of motion q per pixel considering the local neighborhood
Submultisets of pixel values Ω
Intermediate data output by each step

Illumination Changes

Intermittent Motion

Background Motion

Very Short

LaBGen-P computational performance

1. **Motion step / Frame difference:** Applied to each pixel $\rightarrow \mathcal{O}(TWH)$.
2. **Estimation step / Summed-area tables:** Linear frame initialization and constant pixel estimation $\rightarrow \mathcal{O}(TWH + TWH) = \mathcal{O}(TWH)$.
3. **Selection step:** $\mathcal{S}$ comparisons per pixel in the worst case and 1 in the best $\rightarrow \mathcal{O}(T\mathcal{S}WH)$ or $\mathcal{O}(TWH)$.
4. **Generation step / Pixel-wise median:** Using *Introselect* (Musser 1997) linear computation $\rightarrow \mathcal{O}(\mathcal{S})$ per pixel position (offline) or $\mathcal{O}(\mathcal{S})$ per pixel (online) $\rightarrow \mathcal{O}(WH\mathcal{S})$ (offline) or $\mathcal{O}(TWH\mathcal{S})$ (online)

LaBGen-P		**Best case**	**Worst case**
Time	**Offline**	$\mathcal{O}(WH \cdot (T + \mathcal{S}))$	$\mathcal{O}(WHT\mathcal{S})$
Time	**Online**	$\mathcal{O}(WHT\mathcal{S})$	$\mathcal{O}(WHT\mathcal{S})$
Space		$\mathcal{O}(WH\mathcal{S})$	$\mathcal{O}(WH\mathcal{S})$

- Mean pixel **throughput** $\approx 38 \cdot 10^6$ px/s.
- Requires at least $56 \cdot WH\mathcal{S}$ **free bits for processing RGB sequences** with a bit depth of 8 bits (assuming that the quantities of motion are encoded on 32 bits).
- Unlike LaBGen, the memory footprint **cannot be reduced by tuning** $\mathcal{E}$.

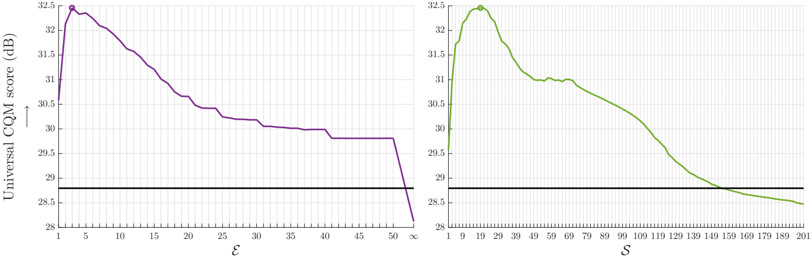
LaBGen-P running times

Sequence	Dimensions $W \times H$	Frames T	Time (ms) Default	Default $\mathcal{E} = 1$	Default $\mathcal{S} = 1$	Default $\mathcal{E} = \infty$	Default $\mathcal{S} = 201$
Board	200×164	228	183	175	85	150	$1,594$
Candela_m1.10	352×288	350	$1,007$	982	472	$1,003$	$8,166$
CAVIAR1	384×256	610	$1,562$	$1,523$	788	$1,510$	$12,423$
CAVIAR2	384×256	460	$1,199$	$1,170$	595	$1,167$	$9,884$
CaVignal	200×136	258	170	153	87	174	$1,483$
Foliage	200×144	394	247	241	126	183	$2,383$
Hall&Monitor	352×240	296	670	632	284	562	$5,474$
HighwayI	320×240	440	843	818	380	677	$7,049$
HighwayII	320×240	500	983	977	433	760	$8,674$
HumanBody2	320×240	740	$1,368$	$1,366$	644	$1,246$	$11,937$
IBMtest2	320×240	90	228	210	80	185	$1,006$
People&Foliage	320×240	341	667	653	294	575	$5,913$
Snellen	144×144	321	152	147	75	116	$1,420$
Toscana	800×600	6	111	110	53	100	178

Important note

$\mathcal{S}$ has a bigger impact than $\mathcal{E}$.

Influence of the parameters on the LaBGen-P universal performance

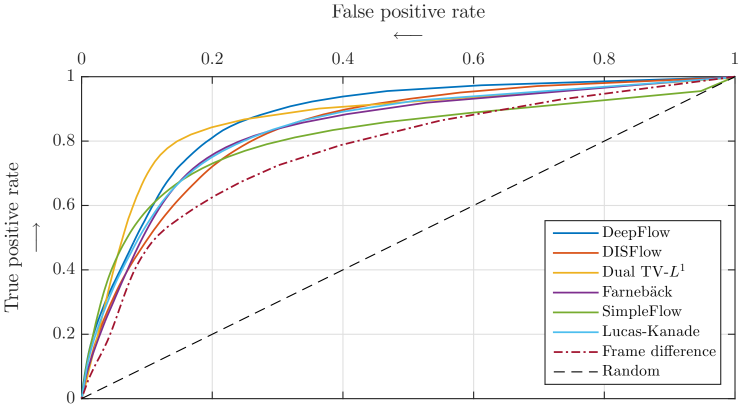


Important note

Equivalent impact but $\mathcal{S}$ leaves the performance slightly more stable.

Optical flow vs. frame difference

- **Better temporally memoryless algorithms** on SBI!
- Good reason to **try them!**



Problematic visual results (LaBGen-OF)

Illumination Changes

Intermittent Motion

Background Motion

Very Short

Best parameter values with respect to each OF algorithm

- Maximizing **universal CQM** on **SBI+SBMnet-GT**.
- Maximization done **with respect to each optical flow** algorithm ($\mathcal{A}$).
- The **threshold value is considered** in the maximization.

Rank	**Parameter values**					**Universal CQM** $\uparrow$
	$\mathcal{A}$	$\mathcal{P}$	$\mathcal{E}$	$\mathcal{S}$	τ	
1	**DeepFlow**	**3**	**8**	**119**	**0.04**	**33.7400**
2	Lucas-Kanade	1	6	63	0.03	33.5150
3	DISFlow	1	3	57	0.02	32.9819
4	Farnebäck	3	3	83	0.05	32.8954
5	SimpleFlow	3	6	49	0.06	32.5169
6	Dual TV-L^1	5	10	75	0.06	32.4793

Details of the CV algorithm

$$
\begin{aligned}
P\left(M=\mathrm{FG} \,|\, \mathbf{S}=\mathbf{s}\right) &= \sum_{i=1}^{N} P\left(M=\mathrm{FG}, O=o_i \,|\, \mathbf{S}=\mathbf{s}\right) \\
\text{(product rule)} &= \sum_{i=1}^{N} \underbrace{P\left(M=\mathrm{FG} \,|\, \mathbf{S}=\mathbf{s}, O=o_i\right)}_{\approx P\left(M=\mathrm{FG} \,|\, O=o_i\right)} \cdot P\left(O=o_i \,|\, \mathbf{S}=\mathbf{s}\right) \\
\text{(hypothesis)} &= \sum_{i=1}^{N} P\left(M=\mathrm{FG} \,|\, O=o_i\right) \cdot P\left(O=o_i \,|\, \mathbf{S}=\mathbf{s}\right) \\
\text{(Bayes)} &= \sum_{i=1}^{N} \frac{P\left(M=\mathrm{FG}, O=o_i\right)}{P\left(O=o_i\right)} \cdot \underbrace{P\left(O=o_i \,|\, \mathbf{S}=\mathbf{s}\right)}_{\approx u(i)} \\
\text{(s.-s. or u. estimators)} &\approx \sum_{i=1}^{N} \frac{\sum_{p\in\gamma} g \cdot u(i)}{\sum_{p\in\gamma} u(i)} \cdot u(i) \quad \text{OR} \quad \sum_{i=1}^{N} \frac{\sum_{\gamma'\in\Gamma} \frac{1}{|\gamma'|} \cdot \sum_{p\in\gamma'} g \cdot u(i)}{\sum_{\gamma'\in\Gamma} \frac{1}{|\gamma'|} \cdot \sum_{p\in\gamma'} u(i)} \cdot u(i)
\end{aligned}
$$

Differences between LaBGen variants

Characteristics	LaBGen	LaBGen-OF	LaBGen-P	LaBGen-P-Semantic
Augmentation	$\mathcal{P}$ passes are performed		✘	
Motion detection	Any BGS algorithm $\mathcal{A}$	Any OF algorithm $\mathcal{A}$	Frame difference	CV or MP with PSPNet
Motion temporality	Aware or memoryless	Memoryless		Intra-frame
Motion information	Crisp segmentation maps s^t		Fuzzy motion maps m^t	
Quantity of motion	One q_i^t per patch f_i^t		One $q_{x,y}^t$ per pixel $p_{x,y}^t$ w. r. t. $\Psi_{x,y}$	
Spatial aggregation	Sum of crisp motion classifications in Ψ_i		Sum of fuzzy motion scores in $\Psi_{x,y}$	
Submultisets content	Ω_i contains maximum $\mathcal{S}$ patches		$\Omega_{x,y}$ contains maximum $\mathcal{S}$ pixels	
Selection criterion	At least one element in the submultiset has a larger quantity of motion			
Cardinality control	Remove the oldest element with the largest quantity of motion			
Background generation	Pixel-wise median filter			
Operating mode	Offline or online with no augmentation		Offline or online	
Number of parameters	4 ($\mathcal{A}$, $\mathcal{P}$, $\mathcal{E}$, $\mathcal{S}$)		2 ($\mathcal{E}$, $\mathcal{S}$)	3 ($\mathcal{V}$, $\mathcal{E}$, $\mathcal{S}$)

- All steps applied consecutively on each frame.
- Quantities of motion simplification:

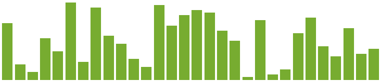
$$q_i^t < q_i^{t'} \iff \sum_{(x,y)\in\Psi_i} \frac{s_{x,y}^t}{|\Psi_i|} < \sum_{(x,y)\in\Psi_i} \frac{s_{x,y}^{t'}}{|\Psi_i|} \iff \sum_{(x,y)\in\Psi_i} s_{x,y}^t < \sum_{(x,y)\in\Psi_i} s_{x,y}^{t'}.$$

- Submultisets as ordered lists:

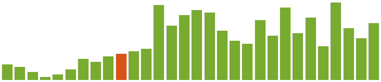
$$\begin{array}{r|c|c|c|c|c|c:c}
 & (f_i^2,2) & (f_i^1,2) & (f_i^4,4) & (f_i^3,4) & (f_i^6,9) & (f_i^5,9) & \\
\text{insert } (f_i^7,5): & (f_i^2,2) & (f_i^1,2) & (f_i^4,4) & (f_i^3,4) & (f_i^7,5) & (f_i^6,9) & \cancel{(f_i^5,9)} \\
 & & & & & \uparrow & & \\
\text{insert } (f_i^8,4): & (f_i^2,2) & (f_i^1,2) & (f_i^8,4) & (f_i^4,4) & (f_i^3,4) & (f_i^7,5) & \cancel{(f_i^6,9)} \\
 & & & \uparrow & & & & \\
 & & & & & & \mathcal{S} &
\end{array}$$

Implementation tricks (2)

- Median through selection algorithm:
  - $\mathcal{S}$ **is odd:** select $\left(A, \left\lceil \frac{\mathcal{S}}{2} \right\rceil\right)$.
  - $\mathcal{S}$ **is even:** $\dfrac{\text{select}\left(A, \frac{\mathcal{S}}{2}\right) + \min A\left[\frac{\mathcal{S}}{2} + 1..\mathcal{S}\right]}{2}$.

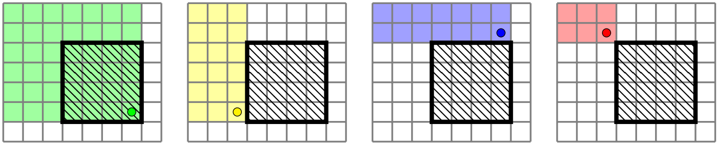


Before selection



Selection of the 10-th element in orange

- Summed area-tables:
  - **Initialization:** $S^t_{x,y} = m^t_{x,y} + S^t_{x-1,y} + S^t_{x,y-1} - S^t_{x-1,y-1}$.
  - **Quantity of motion:** $q^t_{x,y} = S^t_{xr,yb} - S^t_{xl-1,yb} - S^t_{xr,yt-1} + S^t_{xl-1,yt-1}$.



References

References

[Bar+11] O. Barnich and M. Van Droogenbroeck. "ViBe: A universal background subtraction algorithm for video sequences." In: *IEEE Transactions on Image Processing* 20.6 (2011), pp. 1709–1724. DOI: 10.1109/TIP.2010.2101613.

[Bou+17] T. Bouwmans, L. Maddalena, and A. Petrosino. "Scene Background Initialization: a Taxonomy." In: *Pattern Recognition Letters* 96 (2017), pp. 3–11. DOI: 10.1016/j.patrec.2016.12.024.

[Bou01] J.-Y. Bouguet. "Pyramidal Implementation of the Lucas Kanade Feature Tracker - Description of the Algorithm." In: *Intel Corporation* 5.1-10 (2001), p. 4.

[Dos+11] R. Dosselmann and X. D. Yang. "A comprehensive assessment of the structural similarity index." In: *Signal, Image and Video Processing* 5.1 (2011), pp. 81–91. DOI: 10.1007/s11760-009-0144-1.

[Elg+00] A. Elgammal, D. Harwood, and L. Davis. "Non-parametric Model for Background Subtraction." In: *European Conference on Computer Vision (ECCV)*. Vol. 1843. Lecture Notes in Computer Science. Springer, 2000, pp. 751–767. DOI: 10.1007/3-540-45053-X_48.

References

[Far03] G. Farnebäck. "Two-Frame Motion Estimation Based on Polynomial Expansion." In: *Scandinavian Conference on Image Analysis (SCIA)*. Vol. 2749. Lecture Notes in Computer Science. Springer, 2003, pp. 363–370. DOI: 10.1007/3-540-45103-X_50.

[Goy+06] Y. Goyat et al. "Vehicle trajectories evaluation by static video sensors." In: *International Conference on Intelligent Transportation Systems (ITSC)*. Toronto, Canada, 2006, pp. 864–869. DOI: 10.1109/ITSC.2006.1706852.

[Gra+08] M. Granados, H.-P. Seidel, and H. Lensch. "Background Estimation from Non-time Sequence Images." In: *Graphics Interface*. Windsor, Ontario, Canada, 2008, pp. 33–40.

[Hei+06] M. Heikkilä and M. Pietikäinen. "A Texture-Based Method for Modeling the Background and Detecting Moving Objects." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.4 (2006), pp. 657–662. DOI: 10.1109/TPAMI.2006.68.

[Hof+12] M. Hofmann, P. Tiefenbacher, and G. Rigoll. "Background Segmentation with Feedback: The Pixel-Based Adaptive Segmenter." In: *IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Providence, Rhode Island, USA, 2012, pp. 38–43. DOI: 10.1109/CVPRW.2012.6238925.

[Jod+16] P.-M. Jodoin, L. Maddalena, and A. Petrosino. *SceneBackgroundModeling.NET (SBMnet) Dataset and Web Platform*. http://www.scenebackgroundmodeling.net. 2016.

[Jod+17] P.-M. Jodoin et al. "Extensive Benchmark and Survey of Modeling Methods for Scene Background Initialization." In: *IEEE Transactions on Image Processing* 26.11 (2017), pp. 5244–5256. DOI: 10.1109/TIP.2017.2728181.

[Kro+16] T. Kroeger et al. "Fast Optical Flow Using Dense Inverse Search." In: *European Conference on Computer Vision (ECCV)*. Vol. 9908. Lecture Notes in Computer Science. Amsterdam, The Netherlands: Springer, 2016, pp. 471–488. DOI: 10.1007/978-3-319-46493-0_29.

[Li+16] Z. Li et al. *Toward A Practical Perceptual Video Quality Metric*. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652. The Netflix Tech Blog. 2016.

[Li+18] Z. Li et al. *VMAF: The Journey Continues*. https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12. The Netflix Tech Blog. 2018.

[Luc+81] B. D. Lucas and T. Kanade. "An Iterative Image Registration Technique with an Application to Stereo Vision." In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 2. Vancouver, Canada, 1981, pp. 674–679.

[Mad+08] L. Maddalena and A. Petrosino. "A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications." In: *IEEE Transactions on Image Processing* 17.7 (2008), pp. 1168–1177. DOI: 10.1109/TIP.2008.924285.

[Mad+15a] L. Maddalena and A. Petrosino. *Scene Background Initialization (SBI) Dataset*. http://sbmi2015.na.icar.cnr.it/SBIdataset.html. 2015.

References

[Mad+15b] L. Maddalena and A. Petrosino. “Towards Benchmarking Scene Background Initialization.” In: *International Conference on Image Analysis and Processing Workshops (ICIAP Workshops)*. Vol. 9281. Lecture Notes in Computer Science. 2015, pp. 469–476. DOI: 10.1007/978-3-319-23222-5_57.

[Man+04] A. Manzanera and J. Richefeu. “A robust and computationally efficient motion detection algorithm based on Sigma-Delta background estimation.” In: *Indian Conference on Computer Vision, Graphics and Image Processing*. Kolkata, India, 2004, pp. 46–51.

[Mse+19] W. S. Mseddi, M. Jmal, and R. Attia. “Real-time scene background initialization based on spatio-temporal neighborhood exploration.” In: *Multimedia Tools and Applications* 78.6 (2019), pp. 7289–7319. DOI: 10.1007/s11042-018-6399-1.

[Mus97] D. R. Musser. “Introspective Sorting and Selection Algorithms.” In: *Software: Practice and Experience* 27.8 (1997), pp. 983–993. DOI: 10.1002/(SICI)1097-024X(199708)27:8<983::AID-SPE117>3.0.CO;2-\#.

References

[Pin+04] M. H. Pinson and S. Wolf. "A New Standardized Method for Objectively Measuring Video Quality." In: *IEEE Transactions on Broadcasting* 50.3 (2004), pp. 312–322. DOI: 10.1109/TBC.2004.834028.

[Ses+10] K. Seshadrinathan and A. C. Bovik. "Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos." In: *IEEE Transactions on Image Processing* 19.2 (2010), pp. 335–350. DOI: 10.1109/TIP.2009.2034992.

[Sob+17] A. Sobral and E.-H. Zahzah. "Matrix and tensor completion algorithms for background model initialization: A comparative evaluation." In: *Pattern Recognition Letters* 96 (2017), pp. 22–33. DOI: 10.1016/j.patrec.2016.12.019.

[Sta+99] C. Stauffer and E. Grimson. "Adaptive background mixture models for real-time tracking." In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. Fort Collins, Colorado, USA, 1999, pp. 246–252. DOI: 10.1109/CVPR.1999.784637.

[Tao+12] M. W. Tao et al. "SimpleFlow: A Non-iterative, Sublinear Optical Flow Algorithm." In: *Computer Graphics Forum (Eurographics 2012)* 31.2 (2012), pp. 345–353. DOI: 10.1111/j.1467-8659.2012.03013.x.

[Wan+03a] Z. Wang, E. P. Simoncelli, and A. C. Bovik. “Multiscale structural similarity for image quality assessment.” In: *Asilomar Conference on Signals, Systems and Computers*. Vol. 2. Pacific Grove, California, USA, 2003, pp. 1398–1402. DOI: 10.1109/ACSSC.2003.1292216.

[Wan+03b] Z. Wang, H. R. Sheikh, and A. C. Bovik. “Objective Video Quality Assessment.” In: *The Handbook of Video Databases: Design and Applications*. CRC Press, 2003. Chap. 41, pp. 1041–1078.

[Wan+04a] Z. Wang et al. “Image quality assessment: from error visibility to structural similarity.” In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[Wan+04b] Z. Wang, L. Lu, and A. C. Bovik. “Video quality assessment based on structural distortion measurement.” In: *Signal Processing: Image Communication* 19.2 (2004), pp. 121–132. DOI: 10.1016/S0923-5965(03)00076-6.

[Wan+06] H. Wang and D. Suter. “A Novel Robust Statistical Method for Background Initialization and Visual Surveillance.” In: *Asian Conference on Computer Vision (ACCV)*. Vol. 3851. Lecture Notes in Computer Science. Berlin, Heidelberg, 2006, pp. 328–337. DOI: 10.1007/11612032_34.

References

[Wei+13] P. Weinzaepfel et al. "DeepFlow: Large displacement optical flow with deep matching." In: *IEEE International Conference on Computer Vision (ICCV)*. Sydney, Australia, 2013, pp. 1385–1392. DOI: 10.1109/ICCV.2013.175.

[Wre+97] C. Wren et al. "Pfinder: Real-Time Tracking of the Human Body." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (1997), pp. 780–785. DOI: 10.1109/34.598236.

[Yal+13] Y. Yalman and I. Ertürk. "A new color image quality measure based on YUV transformation and PSNR for human vision system." In: *Turkish Journal of Electrical Engineering & Computer Sciences* 21.2 (2013), pp. 603–613. DOI: 10.3906/ELK-1111-11.

[Zac+07] C. Zach, T. Pock, and H. Bischof. "A Duality Based Approach for Realtime TV-L^1 Optical Flow." In: *Joint Pattern Recognition Symposium (JPRS)*. Vol. 4713. Lecture Notes in Computer Science. Springer, 2007, pp. 214–223. DOI: 10.1007/978-3-540-74936-3_22.

[Zha+17] H. Zhao et al. "Pyramid Scene Parsing Network." In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, USA, 2017, pp. 6230–6239. DOI: 10.1109/CVPR.2017.660.

References

[Ziv04] Z. Zivkovic. "Improved adaptive Gaussian mixture model for background subtraction." In: *IEEE International Conference on Pattern Recognition (ICPR)*. Vol. 2. Cambridge, UK, 2004, pp. 28–31. DOI: 10.1109/ICPR.2004.1333992.

[St-+15] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. "SuBSENSE: A Universal Change Detection Method with Local Adaptive Sensitivity." In: *IEEE Transactions on Image Processing* 24.1 (2015), pp. 359–373. DOI: 10.1109/TIP.2014.2378053.