DISSOLVE DETECTION IN ABSTRACT VIDEO CONTENTS

Bogdan IONESCU[1,2], Constantin VERTAN[1] and Patrick LAMBERT[2]

[1]LAPI - University POLITEHNICA of Bucharest, 061071 Bucharest, Romania
{bionescu, cvertan}@alpha.imag.pub.ro

[2]LISTIC – Polytech Annecy-Chambery, B.P. 806, 74016 Annecy, France
patrick.lambert@univ-savoie.fr

this work was partially supported under FSE - European Structural Funds
EXCEL POSDRU/89/1.5/S/62557 (2010-2013)

LAPI – The Image Processing and Analysis Laboratory

University POLITEHNICA of Bucharest

LISTIC Polytech – Laboratoire d'Informatique Systèmes, Traitement de l'Information et de la Connaissance

POLYTECH ANNECY-CHAMBÉRY – Relevons les défis de demain

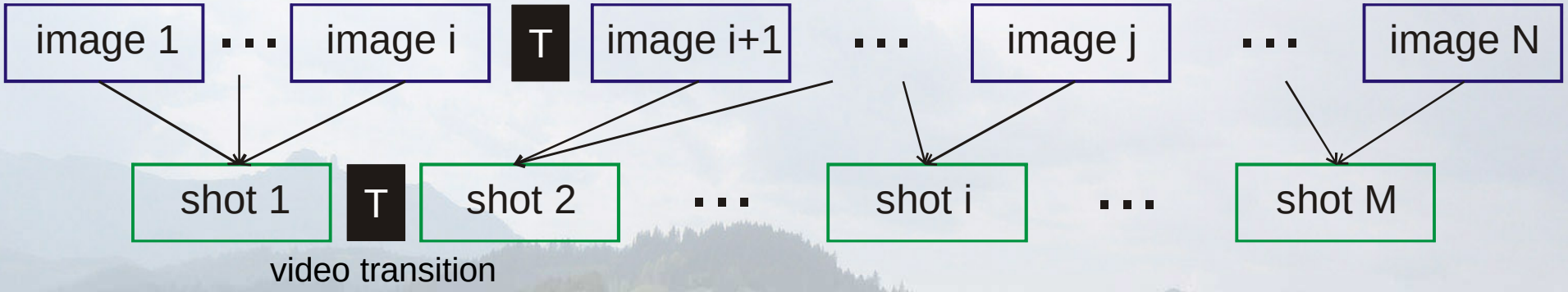
Presentation outline

- Problem statement
- State-of-the-art
- Proposed approach
- Experimental results
- Conclusions

Problem statement

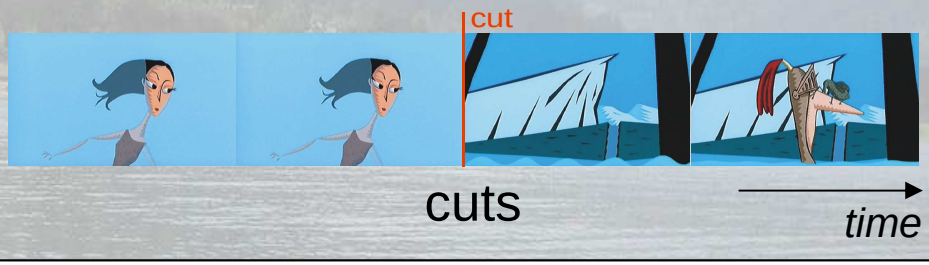
[movies from CITIA, Annecy]

> *temporal segmentation* = parsing the video into its basic temporal units, i.e. **video shots**;



⇒ *temporal segmentation* = detecting the video transitions;

A. sharp transitions or cuts

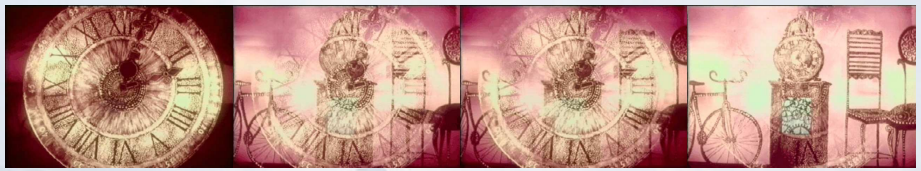


- the most frequent, 30 min ~ 300 cuts;
- existing techniques are highly accurate ~95% (see early TRECVID);

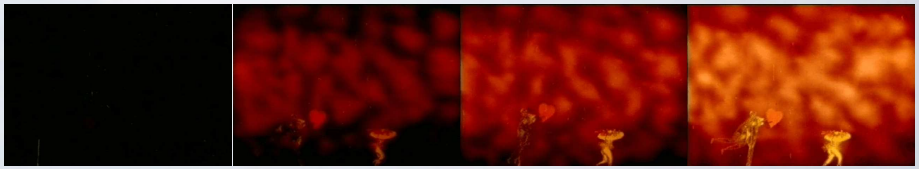
Problem statement

[movies from CITIA, Annecy]

B. gradual transitions = short-in-time visual effects; the most common are:



dissolves



fades

time

- less frequent (~one order measure less than cuts);
- much harder to detect, ~70-80%;

> the interest in temporal segmentation:

- for analysis purpose (temporal structure, localization, etc.);
- for content description purpose (summarization, action, etc.);

State-of-the-art

[Lienhart 01]

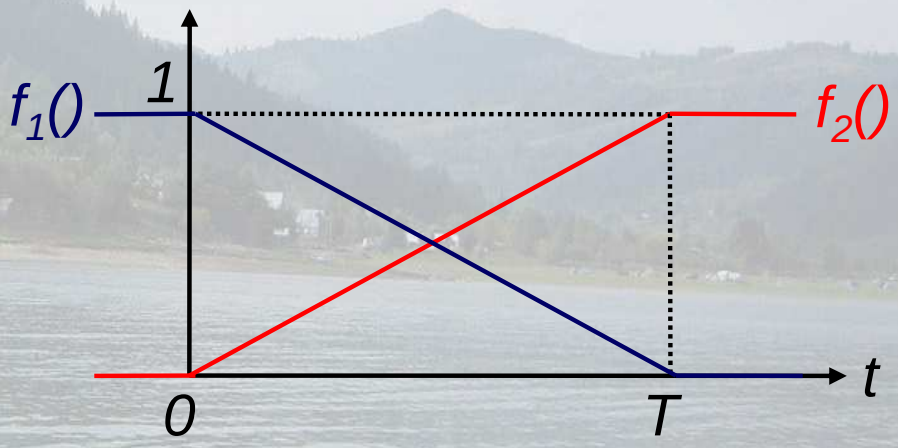
> *definition*: if $S_1(x,y,t)$ and $S_2(x,y,t)$ are two shot sequences (where x,y denote the spatial coordinates and t the time), then the dissolve sequence, $D(x,y,t)$, of duration T, can be defined thus:

$$D(x,y,t) = f_1(t)\cdot S_1(x,y,t) + f_2(t)\cdot S_2(x,y,t)$$

where $t=0,\ldots,T$ and $f_1()$, $f_2()$ are two monotonic functions.

$$f_1(t) = 1 - \frac{t}{T}, \quad f_2(t) = \frac{t}{T}$$

$$0 \le t \le T$$



$\Rightarrow$ superposition of a fade-out with a fade-in sequence;

State-of-the-art

> much more complex transitions (even compared to fades):

- hard to be temporally separated;
- similar time signatures with some camera/object movement;
- many variations (color effects, motion, etc.);
- similar color distributions and spatial structure for the start and end shots;

...

> existing approaches:

- *pixel intensity based*;
- *transform based*;
- *feature based*;
- *other*.

State-of-the-art

A. Pixel intensity based:

[H. Zhang, A. Kankanhalli, S.W. Smoliar'93] → image differences

Distance between consecutive frames accumulated should be greater than t_1, while for consecutive frames should stay below t_2 ($t_2<t_1$);

[B.T. Truong, C. Dorai, S. Venkatesh'00] → mathematical definition

Mean and variance of pixel intensities show a linear and quadratic behavior respectively;

$$Var\{D(x,y,t)\} = f_1^2(t)\cdot Var\{S_1(x,y)\} + f_2^2(t)\cdot Var\{S_2(x,y)\}$$

where S_1 and S_2 are time independent.

$$\Rightarrow Var\{D(x,y,t)\} = \frac{(T-t)^2}{T^2}\cdot Var\{S_1(x,y)\} + \frac{t^2}{T^2}\cdot Var\{S_2(x,y)\}$$

State-of-the-art

A. Pixel intensity based (following):

$$\Rightarrow Var\{D(x,y,t)\} = c \cdot (t-a)^2 - b \qquad \Rightarrow \frac{\partial}{\partial t} Var\{D(x,y,t)\}$$

where *a, b,* and *c* are constants, thus:

$$\frac{\partial^2}{\partial t^2} Var\{D(x,y,t)\}$$

$$a = \frac{T \cdot Var\{S_1(x,y)\}}{Var\{S_1(x,y)\} + Var\{S_2(x,y)\}}$$

$$b = \frac{Var\{S_1(x,y)\} \cdot Var\{S_2(x,y)\}}{Var\{S_1(x,y)\} + Var\{S_2(x,y)\}}$$

$$c = \frac{Var\{S_1(x,y)\} + Var\{S_2(x,y)\}}{T^2}$$

State-of-the-art

A. Pixel intensity based (following):

[C.W. Su, H.-Y.M. Liao, H.R. Tyan, K.C. Fan, L.H. Chen'05] → optical effect

Superposition of two fade sequences: detects the amount of pixels whose intensity is either monotonously increasing or decreasing in a certain observation window;

- good reliability and invariance to image fluctuations/noise;

B. Transform based:

[O. Bao, M. Lian, L. Guan'07] → compressed domain

Performing the detection on a transformed domain, e.g. DCT;

- real-time performance capabilities; increasing accuracy may require decompressing to a certain level of detail;

State-of-the-art

C. Feature based:

[R. Zabih, J. Miller, K. Mai'99] → contour/edge information

During dissolve, object contours from the start image disappear progressively, while contours from the final image appear gradually;

$$ECR_{out} = \frac{X_k^{out}}{N_k}, \quad ECR_{in} = \frac{X_{k+1}^{in}}{N_{k+1}}$$

where N_k is the number of contour points in image k and X_k^{out} is the number contour points which disappeared from image k.

[Y. Wang, Y. Yang, T. Ren, G. Wu'09] → feature points, e.g. SURF

- vulnerable to motion and visual effects;

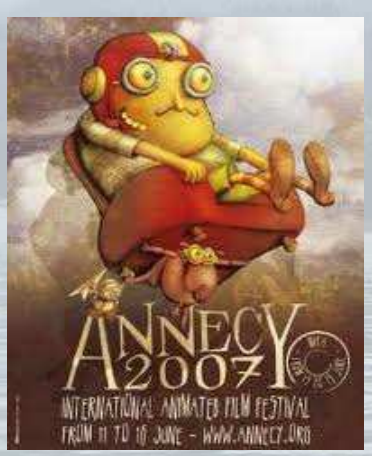
Proposed approach

> regardless the information used, most of the existing dissolve detection methods are naturally designed and validated on *natural movies*;

> we address a particular domain:

→ **artistic animated movies** ≠ cartoons

> a major entertainment industry:



Annecy-France,
Ottawa-Canada,
Portugal,
…

↳ e.g. more than 31,000 movie titles, 22,924 companies and 60,879 professionals;

Proposed approach

[movies from CITIA, Annecy]

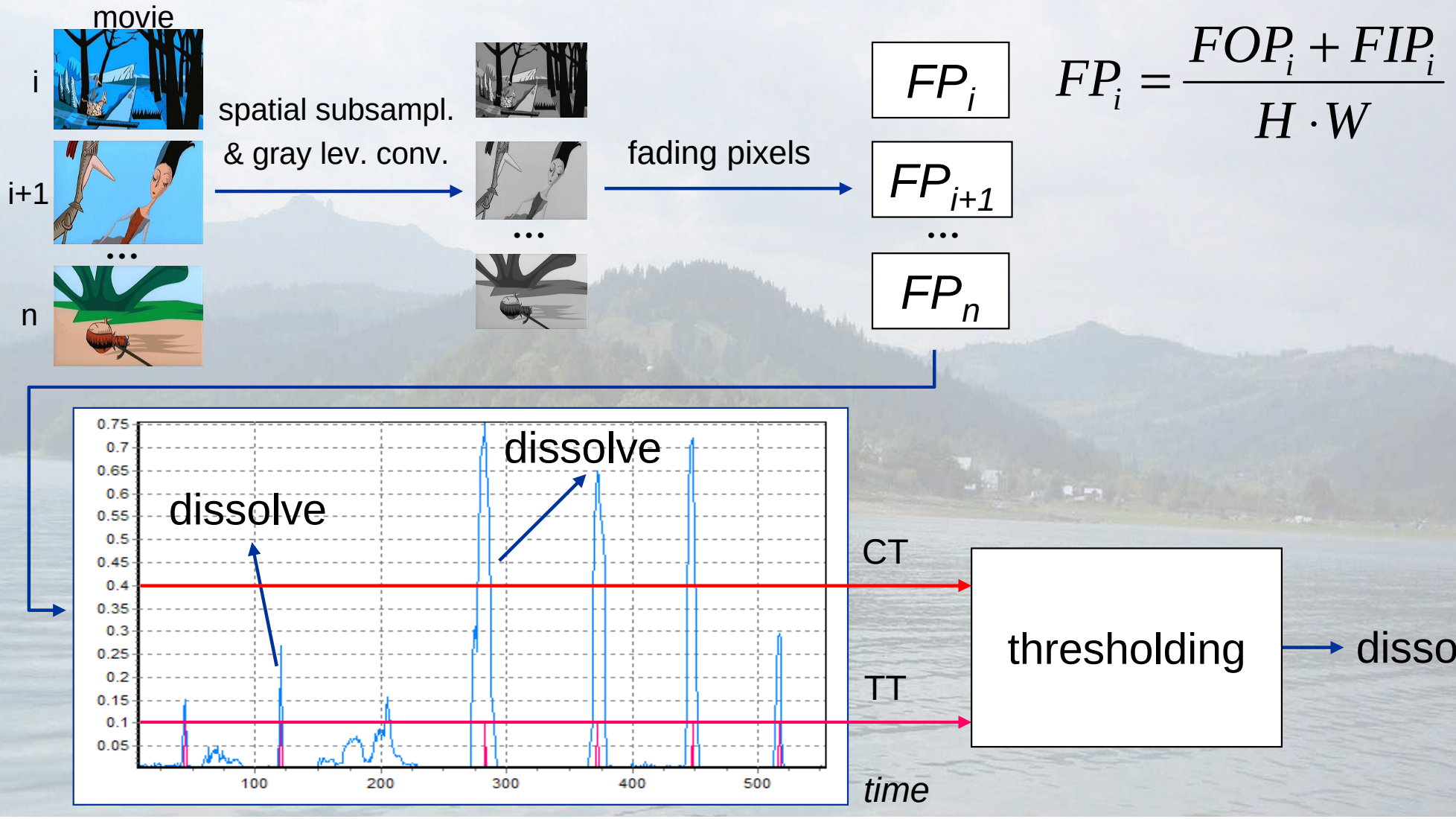
> artistic animated movies are different in many respects:

⇒ dissolve transitions usually show atypical patterns;

Proposed approach

> we propose a straightforward fast and yet efficient approach:

[inspired from C.W. Su, H.-Y.M. Liao, H.R. Tyan, K.C. Fan, L.H. Chen'05]



Proposed approach

> method details:

1. *FOP* and *FIP* are computed on a restraint time window (*N=3* frames) to overcome visual discontinuity;

2. Dissolve localization using twin-thresholding:

2.A. if $FP_i > t_{CT}$ (Certain Threshold, e.g. 0.4) + local max

$\Rightarrow$ dissolve declared in $[i-1/2t_{max}; i+1/2t_{max}]$

2.B. if $FP_i > t_{TT}$ (Tolerant Threshold, e.g. 0.1) + $FP_i < t_{CT}$

$\Rightarrow$ potentially a dissolve middle frame,

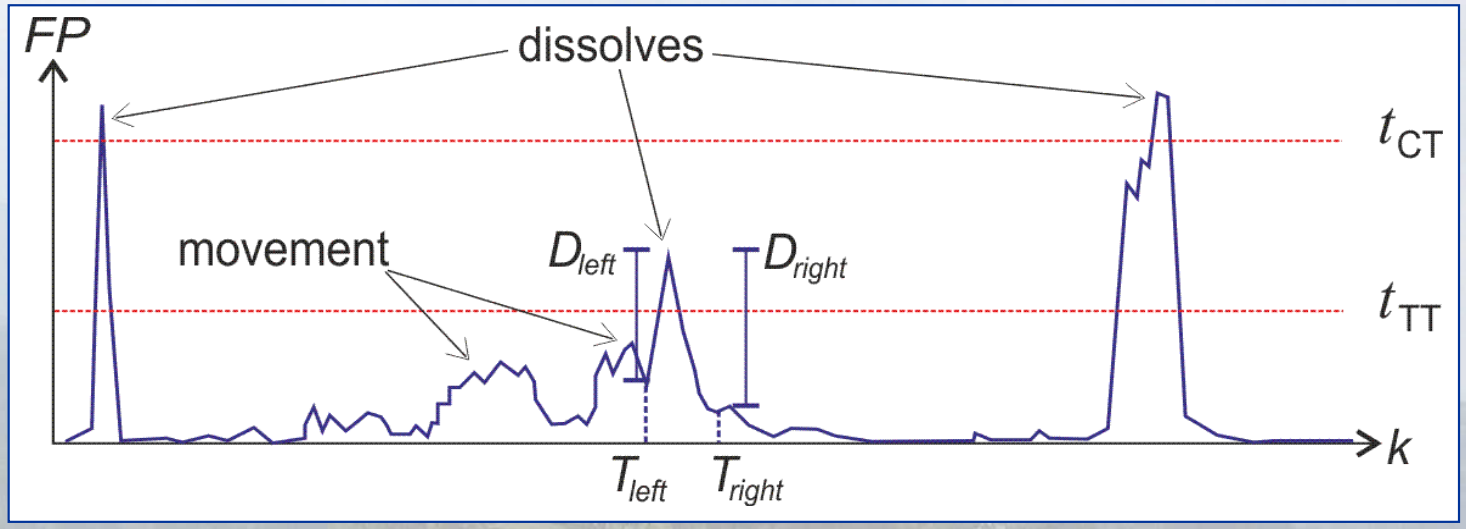
for a dissolve, *FP* function should decrease enough on both sides;

where i is the time index, t_{max} represents the average dissolve duration; threshold values were empirically determined;

Proposed approach

> method details (following):

2. Dissolve localization using twin-thresholding (following):



- we determine T_{right} and T_{left} where *FP* starts increasing again;
- we determine D_{right} and D_{left} to quantify the decrease;
- if $D_{right} > 0.5\ FP_i$ and $D_{left} > 0.5\ FP_i$

  $\Rightarrow$ dissolve declared in $[i-1/2t_{max}; i+1/2t_{max}]$;

Experimental results

[movies from CITIA, Annecy]

> test database: 61 minutes and 452 dissolves;

legend:
↑ = highly difficult (abstract, very complex visual contents, motion discontinuity);
↔ = average difficulty (average amount of visual effects, motion content less discontinuous)

> asses performance:

$$precision = \frac{GD}{GD + FD}, \quad recall = \frac{GD}{N_{dissolves}}$$

where *GD* = good detections, *FD* = false detections and $N_{dissolves}$ = 452;

Experimental results

[movies from CITIA, Annecy]

> detection results:

• overall:

precision = 94%, **recall = 79.6%**, 360 good detections, only 23 false detections;

• per movie:

movie	count	GD	FD	P	R
Ex-Enfant↔	75	65	8	89%	86.7%
Le Moine et le Poisson↔	61	47	2	95.9%	77.1%
M. Pascal↑	98	76	2	97.4%	77.6%
Une Bonne Journée↔	19	19	0	**100%**	**100%**
Paradise↑	60	44	7	86.3%	73.3%
Cœur de Secours↑	67	47	2	95.9%	70.2%
The Sand Castle↔	72	62	2	96.9%	86.1%

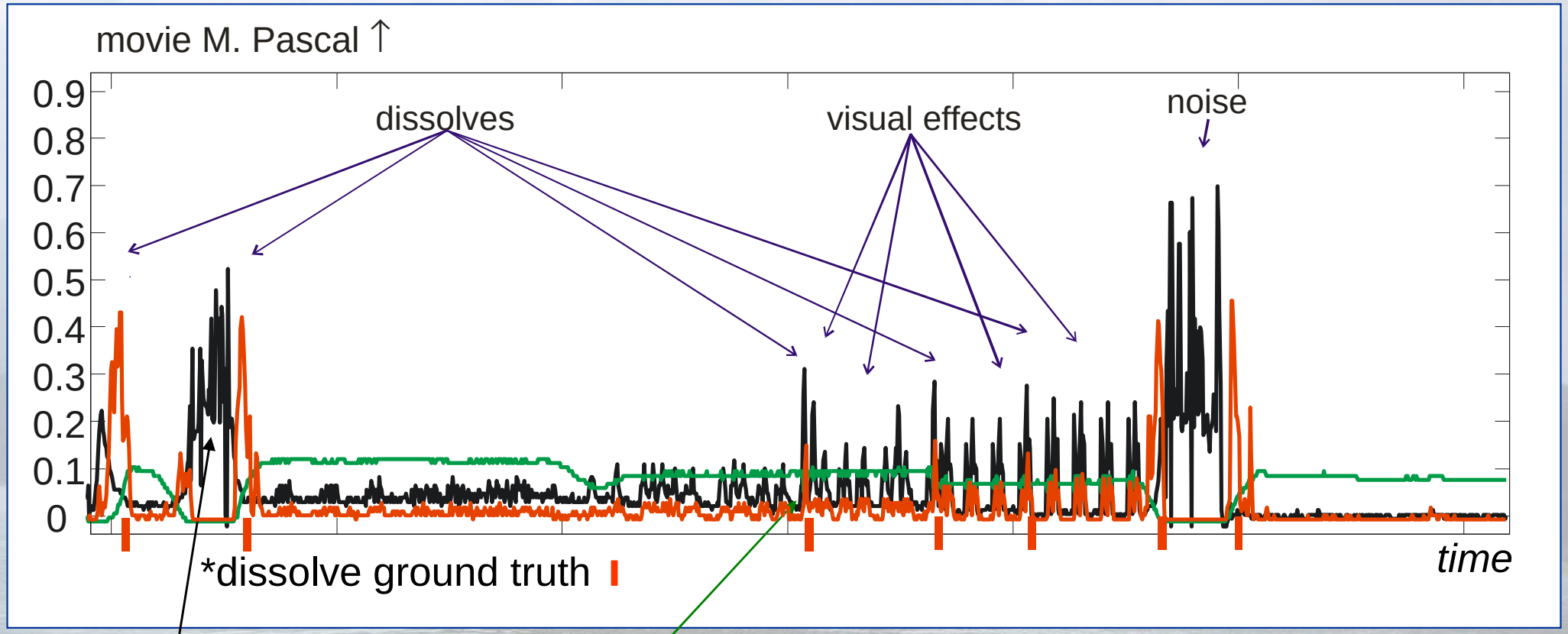
$P \in [86.3;100]\%$

$R \in [70.2;100]\%$

Experimental results

[movies from CITIA, Annecy]

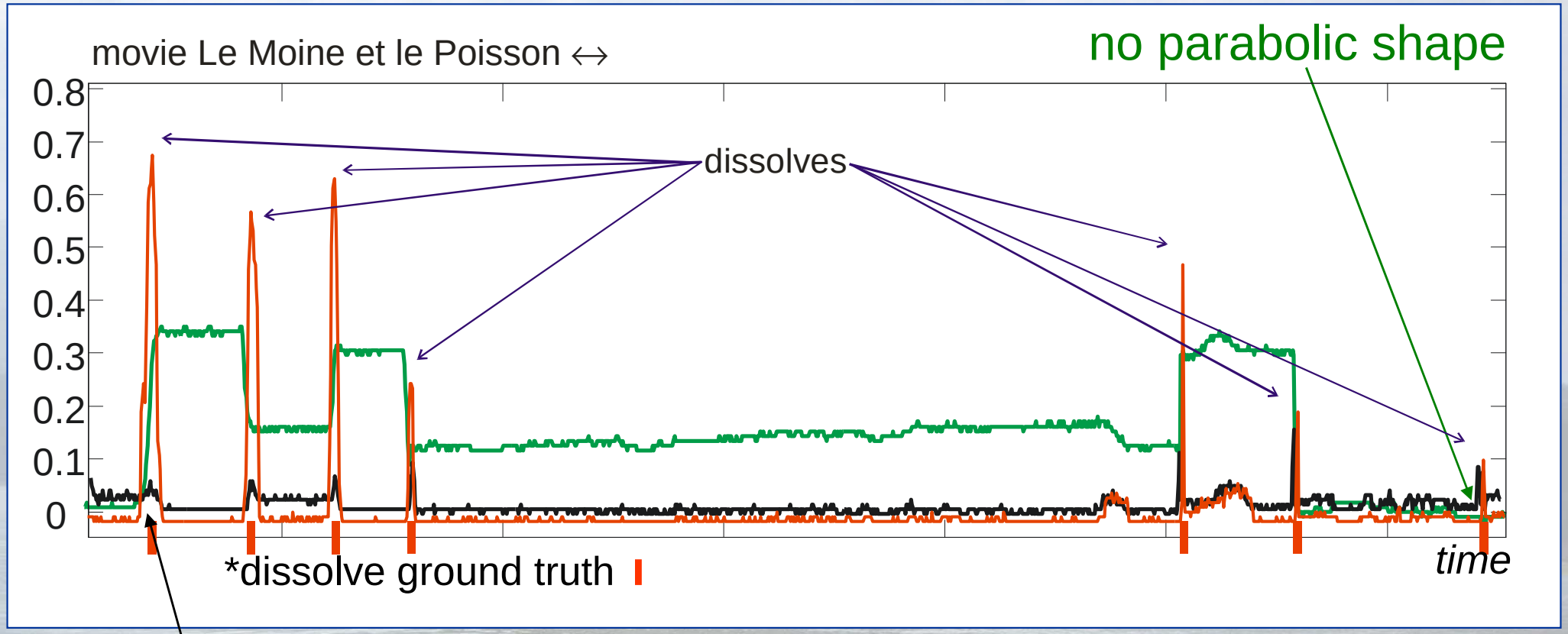
> comparison with other approaches:



Experimental results

[movies from CITIA, Annecy]

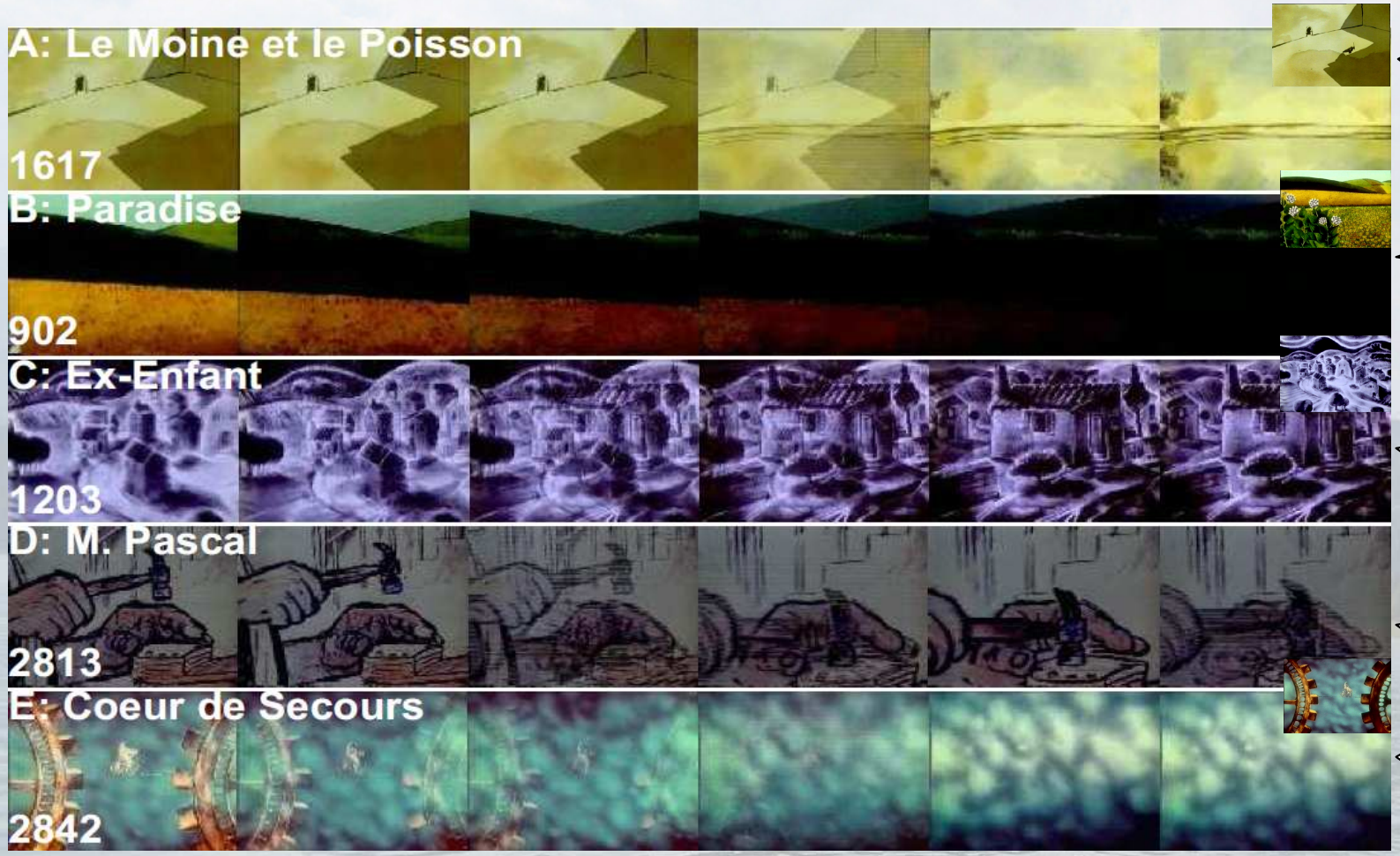
> comparison with other approaches:



Experimental results

[movies from CITIA, Annecy]

> e.g. of dissolves successfully detected:



+very short dissolves ~3 frames

+atypical

+similar color and structure

+highly discontinuous

+global motion and intensity fluctuations

Conclusions

> we proposed an *intensity-based approach to dissolve detection* which was adapted to the specificity of animated movies;

> it achieves precision and recall ratios of up to 94% and 79.6% respectively (while some other classic approaches proved to be inefficient to this task);

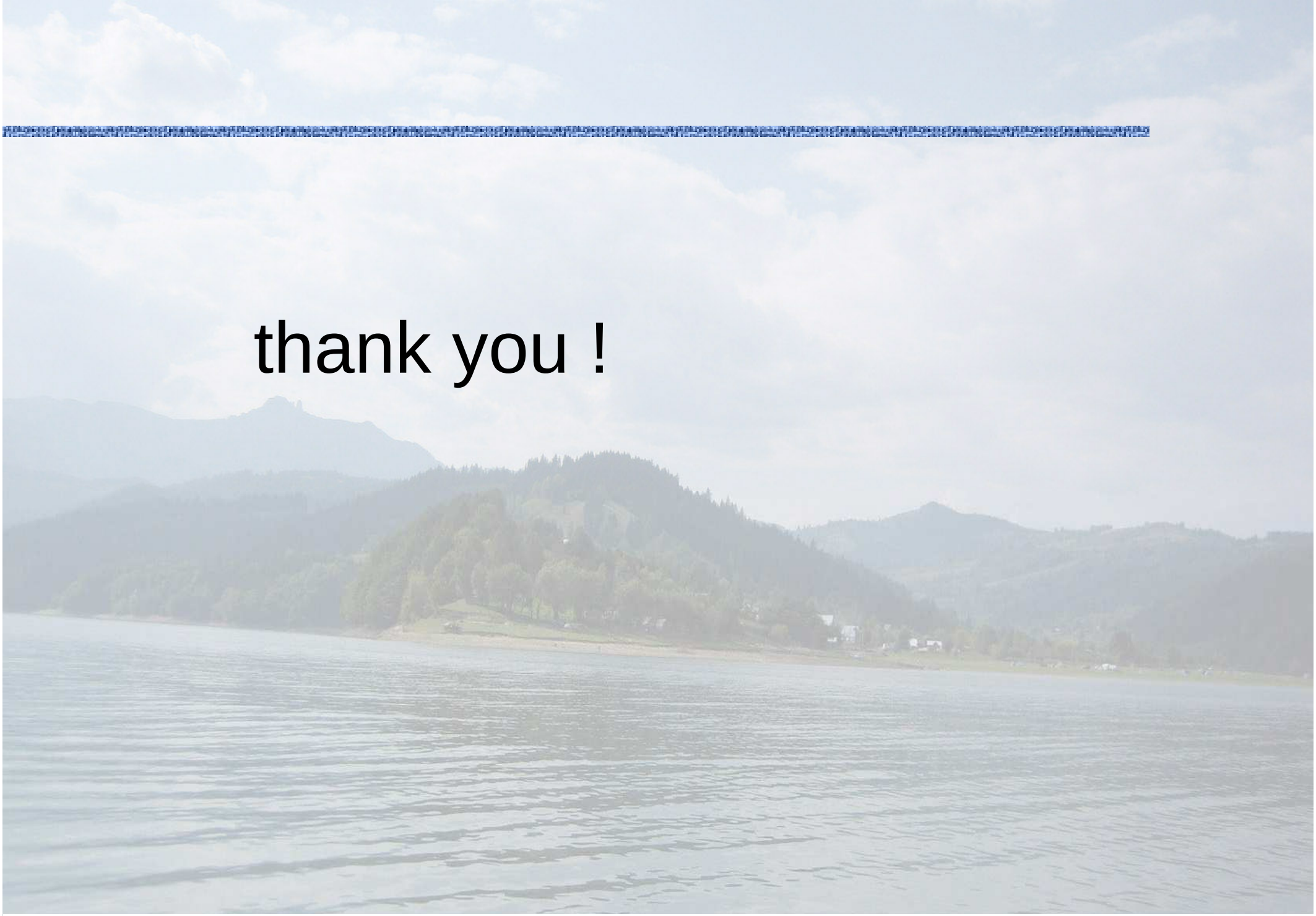
> main limitations:

→ choice of several thresholds;

→ unable to detect the dissolve boundaries;

→ beware of fades!

Acknowledgements: we would like to thank CITIA - The City of Moving Images (International Animated Film Festival of Annecy) and Folimage French Animation Company for their support.


thank you !