Detection of salient motions and temporal alignment in videos

Katy Blanc

Laboratoire I3S, EDSTIC, UCA

02/03/2019, Sophia Antipolis, France

Overview

Context

- The video
- State of the art in video classification
- Issues
- Objectives

Singlets – Salient Motion Detection

- Sport Analysis
- From Singularities to Singlets
- Detection in Footbal matchs

Temporal Warping

- Temporal Warping in Computer Vision
- DTW and its extensions
- Alignment Framework
- Results

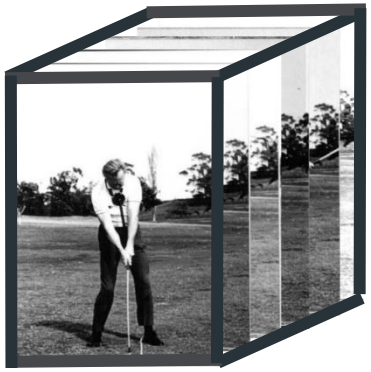
Conclusion

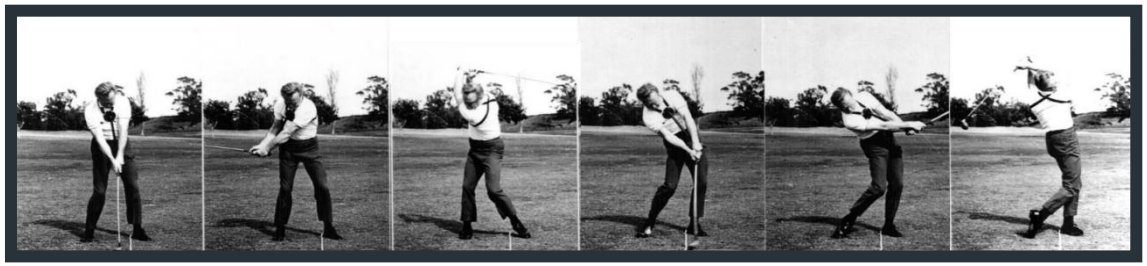
- Contributions
- Perspectives

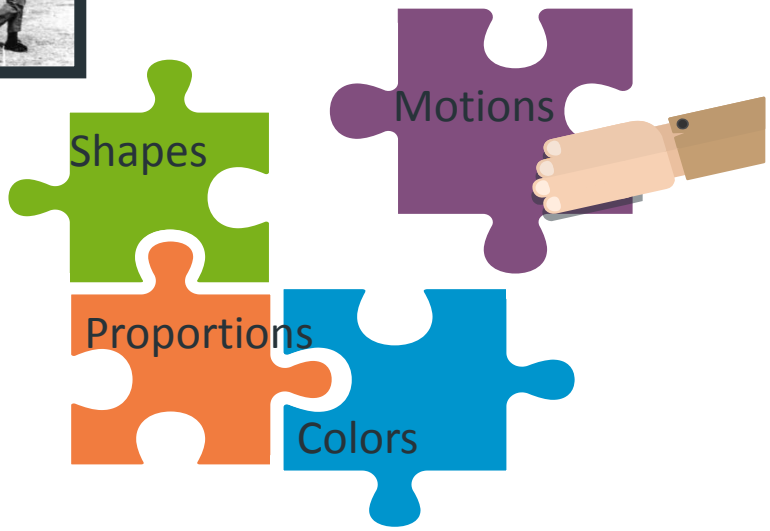
What is a video?

A video is considered as :

- Either a discrete suite of 2D images
- Or a cube of 3D pixel called voxel

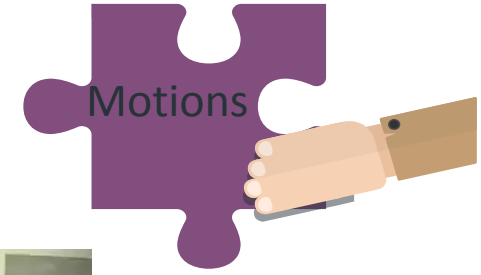






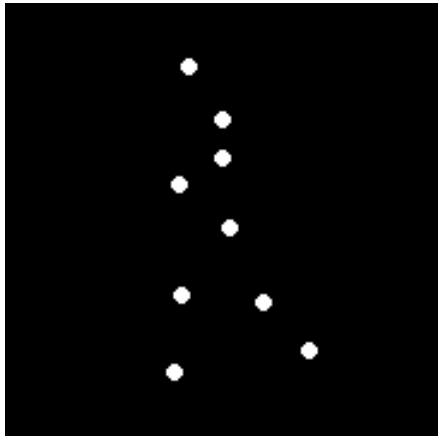
Motion: a primal information

Our attention



Coarse but informative interpretation

- Fast or Slow motion
- Cyclic Motion
- The timeline of the event

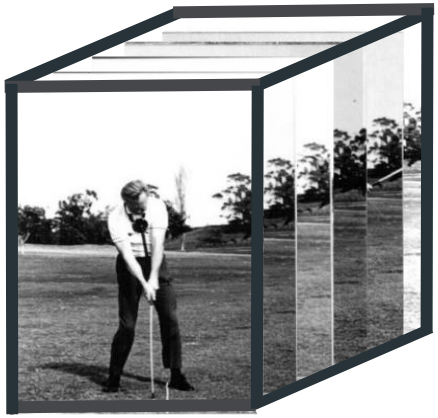


The temporal dimension

This dimension's characteristic:

- Redundancy

From 25 to 1000 images per seconde



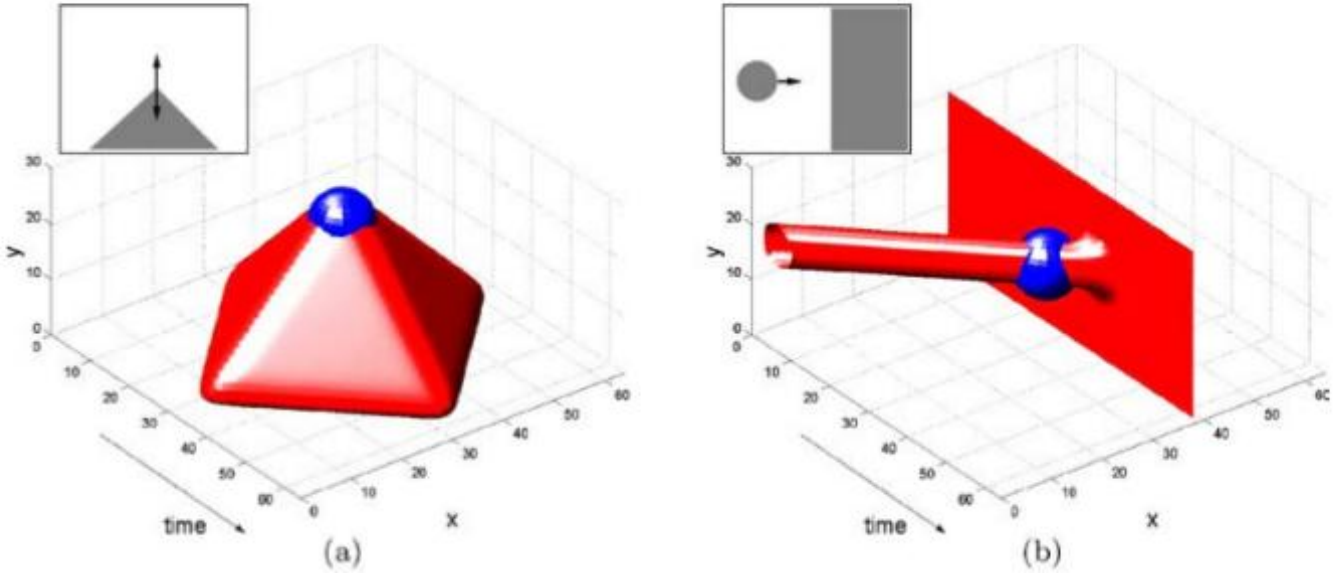
- Temporal Elasticity

Variation of execution speed

Robust Content to local warpings

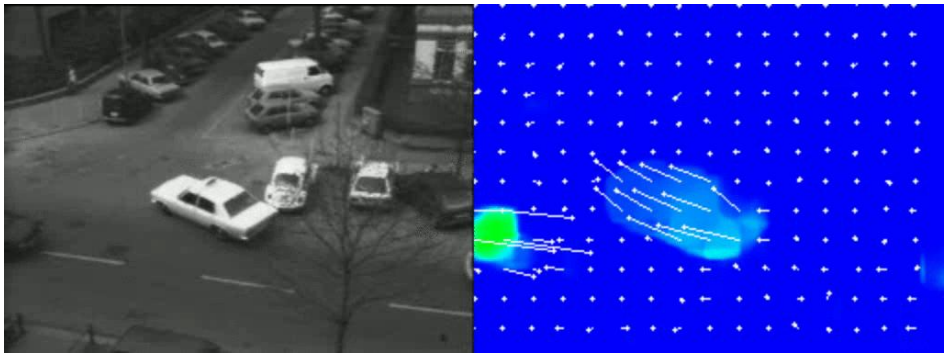
Handmade video descriptions

- Bag of Features (Sift)
- Spatio-Temporal Interest Point (STIP)[Laptev, 2005]

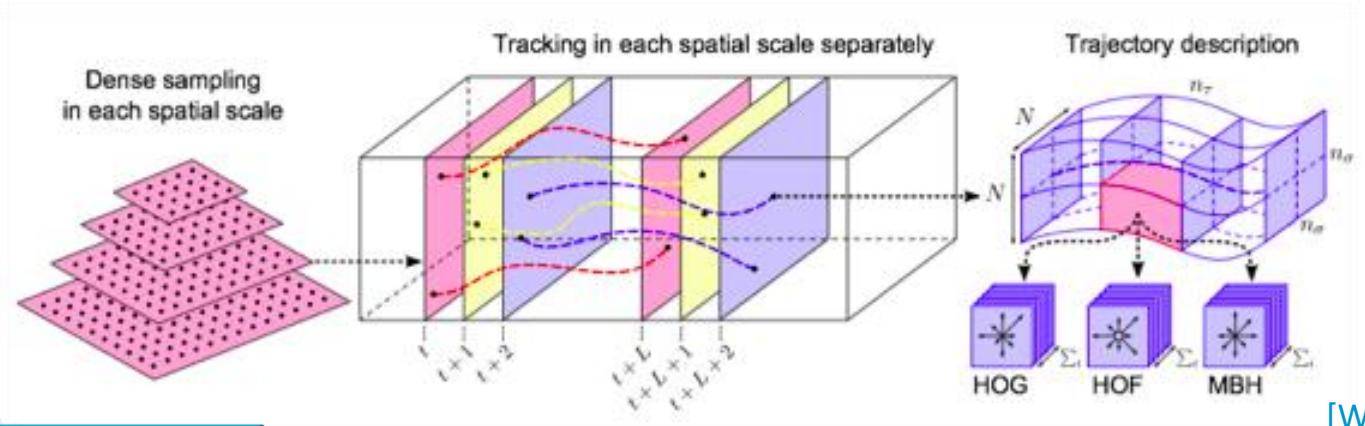


Handmade video descriptions

- Optical Flows



- Improved Dense Trajectories (IDT) [Wang, 2011]

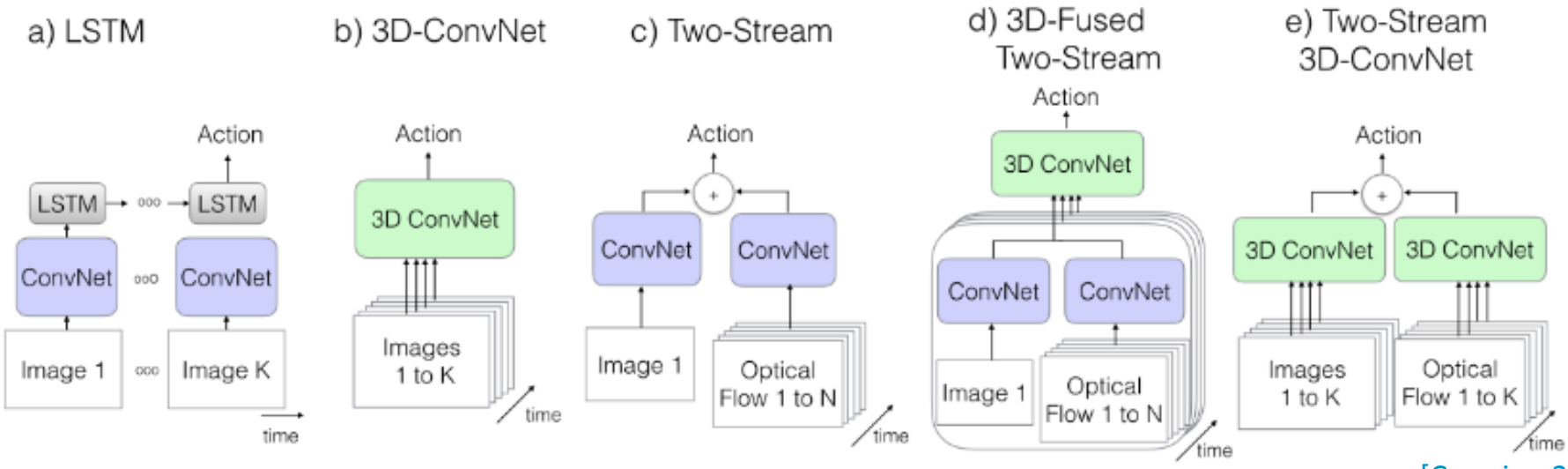


[Wang, 2011]

Deep Neural Network in Video

3 questions to characterise a video neural network :

- Are the convolutional kernels in 2D or in 3D ?
- Is the temporal aggregation a simple fusion or a recurrent fusion ?
- Is the input the original RGB pixels, the optical flows or both ?



[Carreira, 2017]

Issues

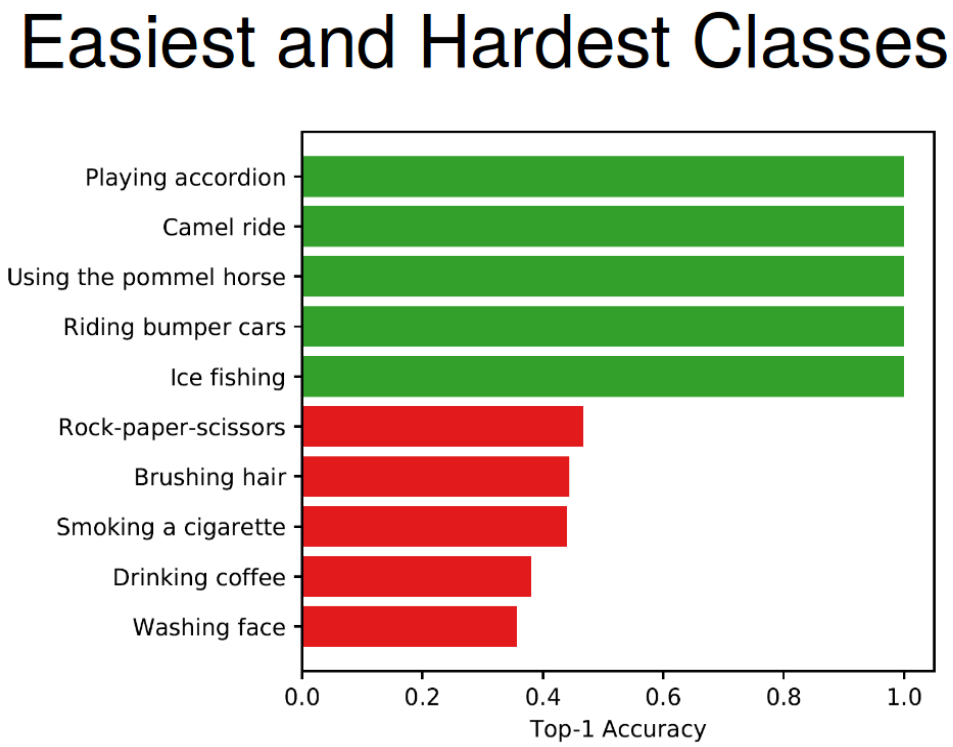
RNN are hard to train because of the temporal redundancy

The 3D convolutional networks and Temporal Linear Encoding [Diba, 2017] : higher quantity of parameters with the depth of the network

Best accuracies performed by 2-stream networks and combinaison of different architectures [Carreira, 2017]

Improvements:

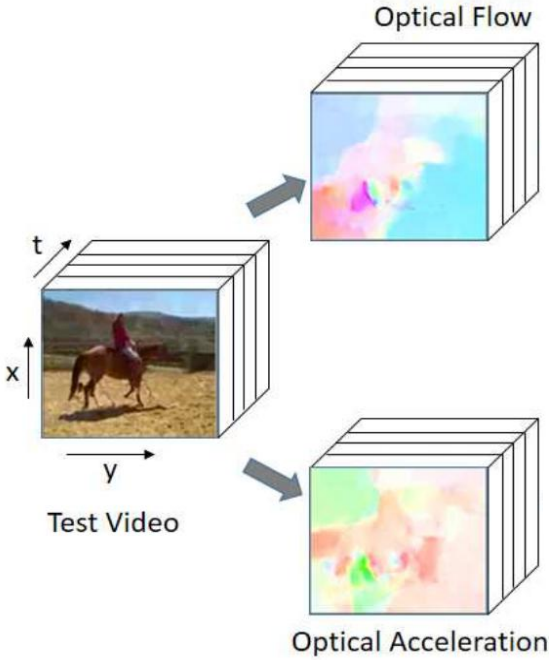
- Residual connections between content and motion [Feichtenhofer ,2016]
- Uniform Temporal Sampling (TSN) [Wang, 2018]



ACTIVITYNET

Other approaches

- Dynamical image network [Bilen, 2016]
- Attention mecanism [Li, 2018]
- Image ranking method [Fernando, 2015]
- Acceleration Optical Flows [Edison, 2017]



« *Brave new ideas for motion representation* »
(Workshop à ECCV'16, CVPR'17)

Issues

A Robust Video Motion Description

- Description and Classification : Relation between the static content and their motion; data classification
- Redundancy : relevant content selection
- Elasticity : robust to speed variations
- Generalisation : Adaptation to all type of motion

Overview

Two different applications using video and motion description:

- From Singularities to Singlets:
  local explicit description to detect salient motions

Singlets – Salient Motion Detection

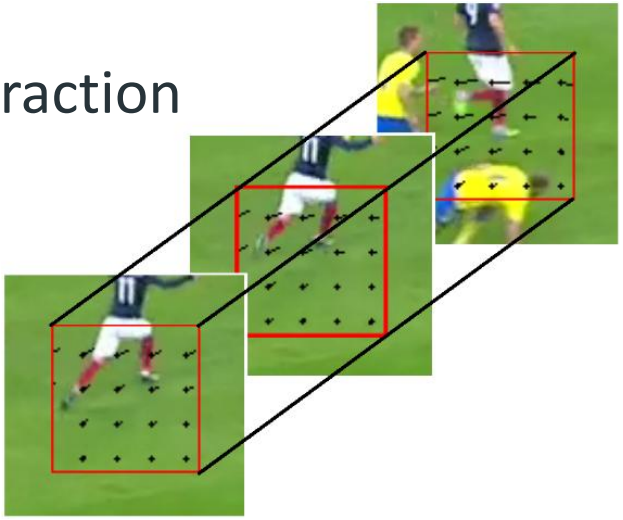
- **Temporal Elasticity Compensation** on a motion class:
  execution speed normalisation

Temporal Warping

Overview

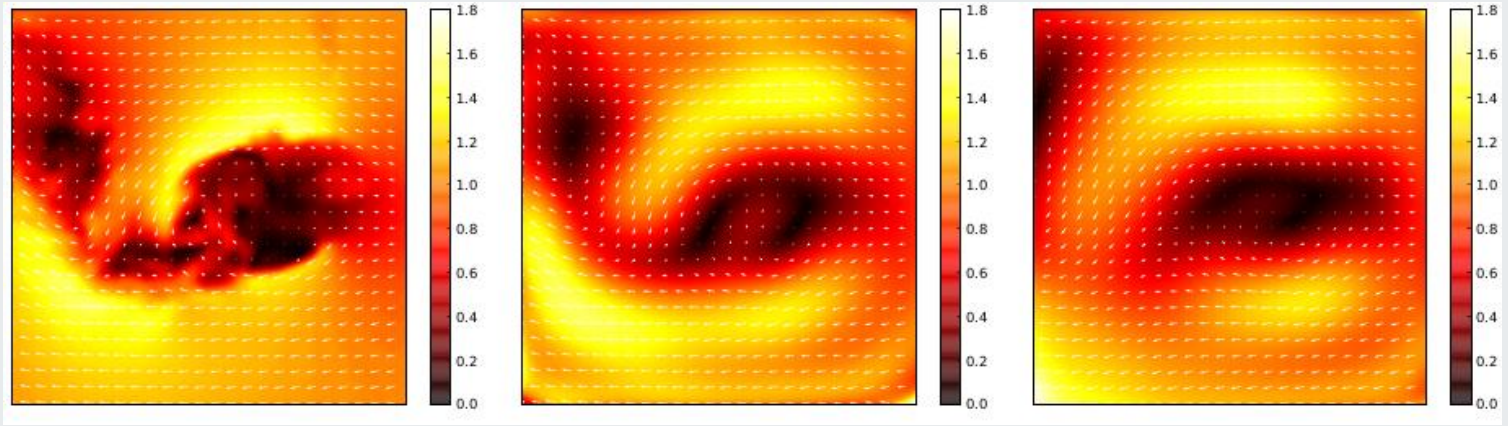
Singlets

- From Optical Flow to singularities chains : the Singlets
- Related work in Sport Abstraction
- Salient motion detection and video abstraction

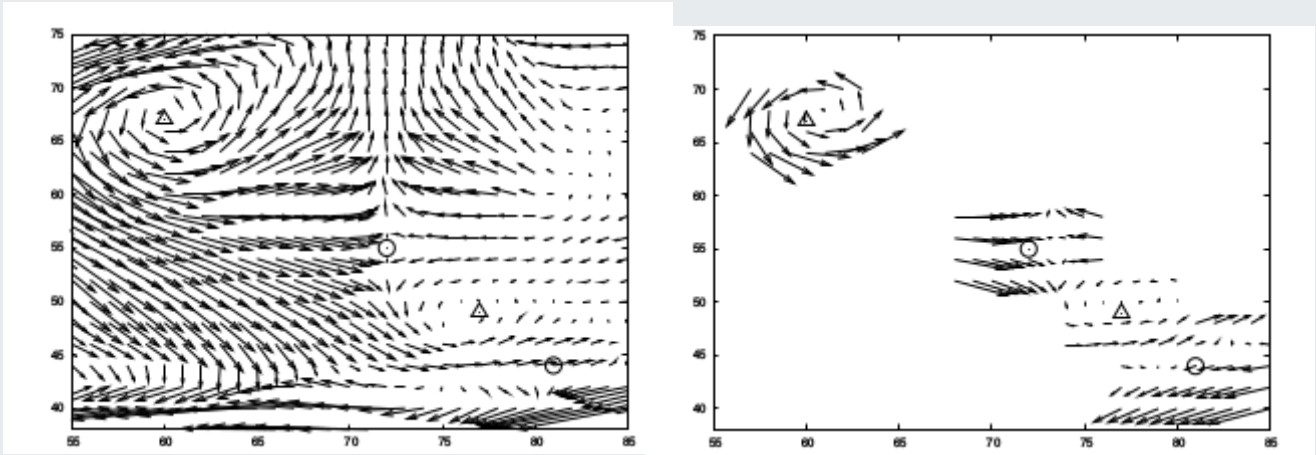


Inspiration: Fluid movements

Inspired of the works of Druon [Druon, 2006] and of Kihl [Kihl, 2008]



[Druon, 2006]



[Kihl, 2008]

Singularity in a optical flow

- The optical flow F : $\Omega = [-1;1]^2 \subseteq \mathbb{R}^2 \to \mathbb{R}^2$
  $(x_1, x_2) \to \big(U(x_1, x_2), V(x_1, x_2)\big)$

- Projection to the polynomial space of degree D

$$P(x_1, x_2) = \sum_{i,j}^{i+j \leq D} a_{i,j} x^i y^j$$

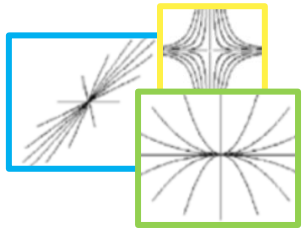
  - Scalar product : $< F_1, F_2 > = \iint_\Omega F_1(x_1, x_2) F_2(x_1, x_2)\, dx_1 dx_2$

  - Legendre basis : $a_n = \frac{2n+1}{n+1}$; $b_n = 0$; $c_n = \frac{n}{n+1}$

$$\begin{cases} P_{-1,j} = 0 \\ P_{i,-1} = 0 \\ P_{0,0} = 1 \\ P_{i+1,j} = (a_i x_1 + b_i) P_{i,j} - c_i P_{i-1,j} \\ P_{i,j+1} = (a_j x_2 + b_j) P_{i,j} - c_j P_{i,j-1} \end{cases}$$

Singularity in a optical flow

Optical flow F :
$$\Omega = [-1;1]^2 \subseteq \mathbb{R}^2 \to \mathbb{R}^2$$
$$(x_1, x_2) \to \big(U(x_1, x_2), V(x_1, x_2)\big)$$



Projection to the polynomial space of degree D

$$\begin{cases} U = u_{0,0}P_{0,0} + u_{0,1}P_{0,1} + u_{1,0}P_{1,0} \\ V = v_{0,0}P_{0,0} + v_{0,1}P_{0,1} + v_{1,0}P_{1,0} \end{cases}$$

Affine expression of the projection in the canonical basis

$$\begin{pmatrix} U \\ V \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + b \qquad \text{with } A \in M_{2,2} \text{ and } b \in \mathbb{R}^2$$

If $\det(A) \neq 0$ and if $A^{-1}b \subseteq [-1;1]^2$

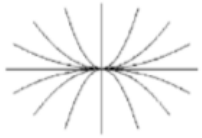
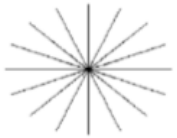
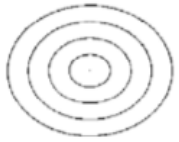
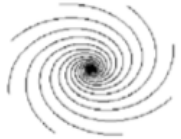
Then a singularity is detected in the position $\begin{pmatrix} x \\ y \end{pmatrix} = A^{-1}b$

An optical flow singularity is a root of its polynomial approximation

Singularity categories

If a singularity is detected in the position $\begin{pmatrix} x \\ y \end{pmatrix} = A^{-1}b$,

We determine its type among 6 singularity shapes

node	saddle	star node	improper node	center	spiral
					
$\lambda_1\lambda_2 > 0$ $\Delta(A) > 0$ and $tr(A) > 0$	$\lambda_1\lambda_2 < 0$ $\Delta(A) > 0$ and $tr(A) < 0$	$\lambda_1 = \lambda_2$ $\Delta(A) = 0$ and $a_{12} - a_{21} = 0$	$\lambda_1 = \lambda_2$ $\Delta(A) = 0$ and $a_{12} - a_{21} \neq 0$	$\lambda_{1,2} = \alpha \pm \imath\beta$ $\Delta(A) < 0$ and $tr(A) = 0$	$\lambda_{1,2} = \alpha \pm \imath\beta$ $\Delta(A) < 0$ and $tr(A) \neq 0$

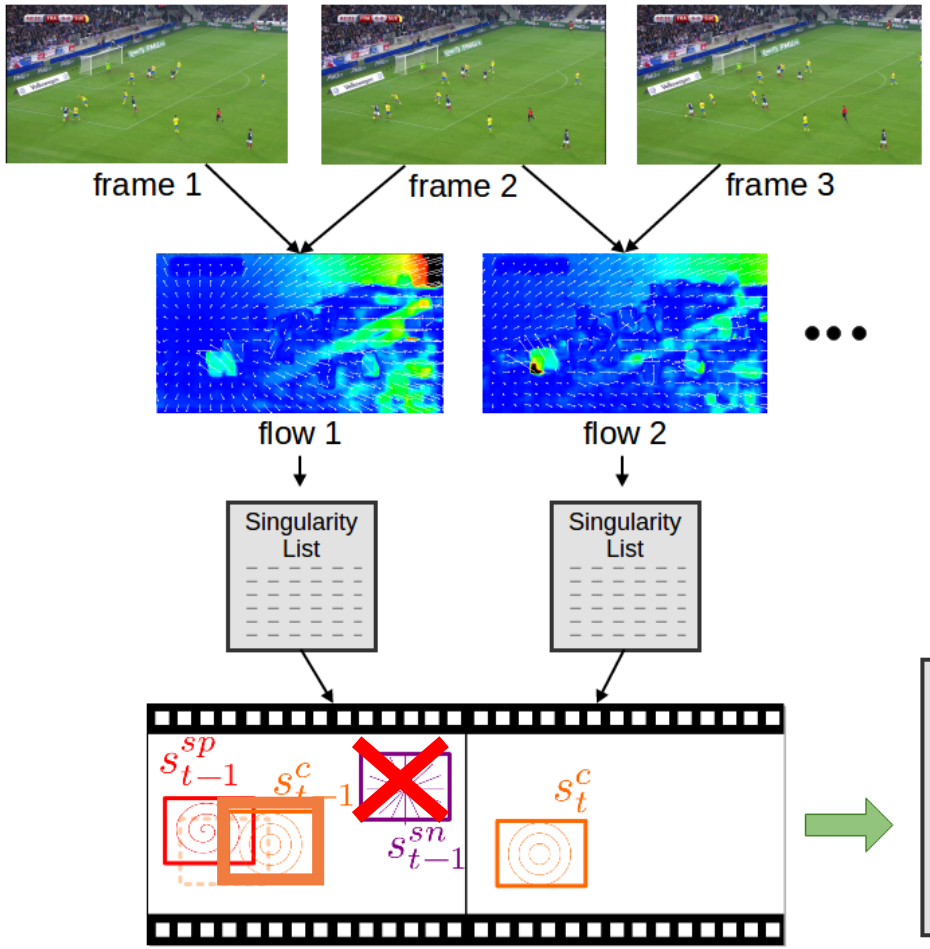
[Druon, 2009]

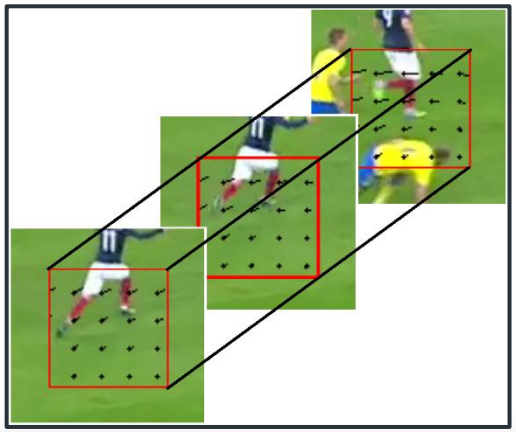
With [redacted] and λ_1, λ_2 the eigen values of A

-> Sliding window methodology

Singlets: Singularities chains

From singularities of optical flows to their trajectories : Singlets





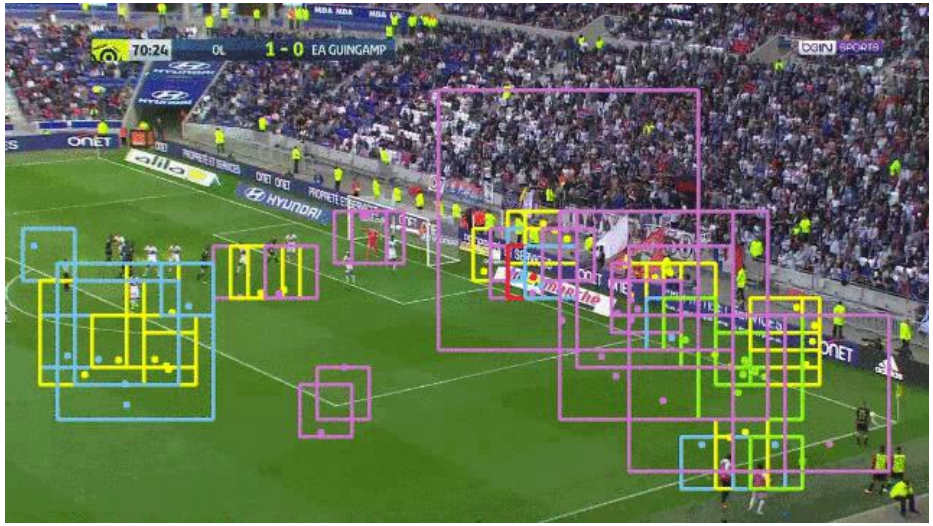
$$V(sin_{\cdot} d(\begin{pmatrix}\mathbf{A}\\x\\y\end{pmatrix},\begin{pmatrix}\mathbf{A}'\\x'\\y'\end{pmatrix}) = ||A-A'||_F+\lambda||\begin{pmatrix}x\\y\end{pmatrix}-\begin{pmatrix}x'\\y'\end{pmatrix}||_2 \quad \alpha\}$$

Singlets List

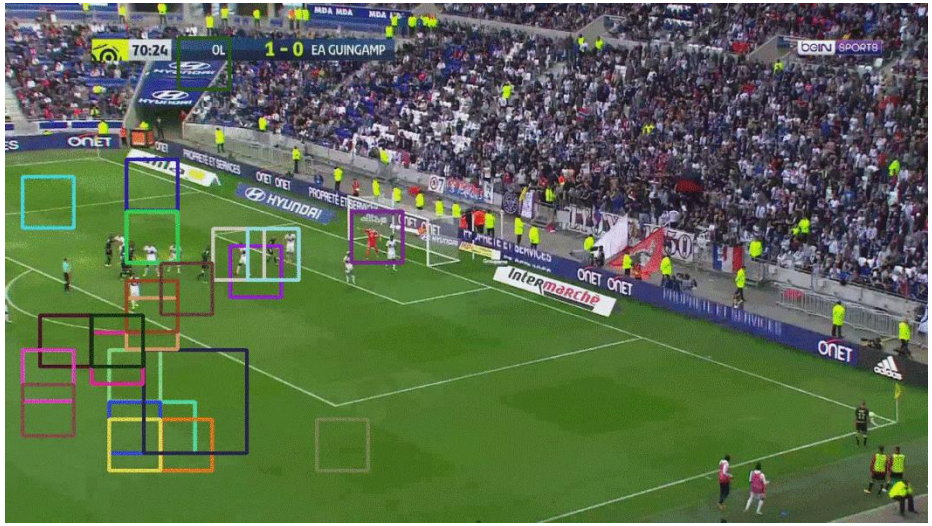
frame, length, sing1, sing2, ...

Singlets: Singularities chains

Singularities

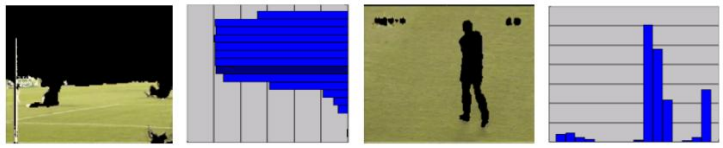


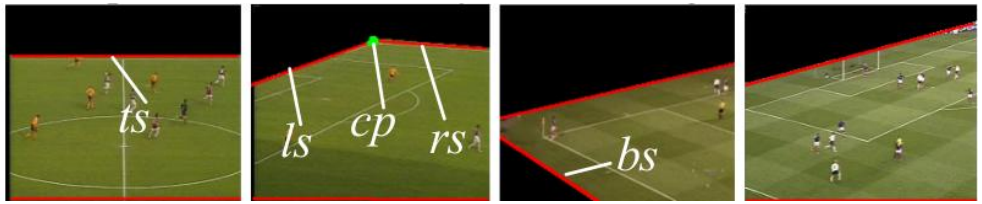
Singlets



Related works: Football Abstraction Video

- Ground lines position, extract segmentation [Ye, 2005]





- Logo and score apparitions, goal mouth position in the extract [Zwabba, 2012]

- Face and skin detection, whistle detection [Raventos, 2014]

Our database

Lack of standard benchmark for the evaluation …

Germany vs Portugal

4-0

Nigeria vs Argentina

2-3

Switzerland vs France

2-5

France vs Honduras

3-0

Matchs	FIFA ground-truth
Germany vs Portugal	30
Nigeria vs Argentine	51
France vs Honduras	54
Switzerland vs France	40

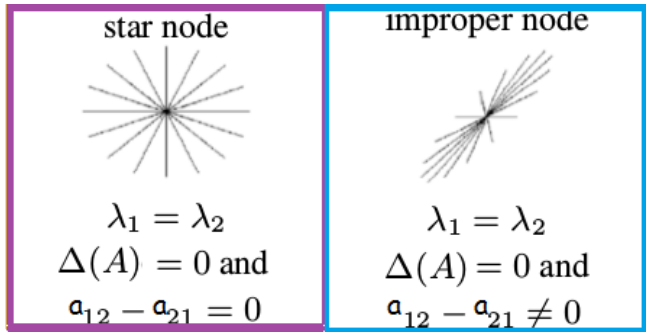
© http://www.fifa.com/worldcup/matches/

Zoom detection

- Zoom = Star singularity
- Zoom + translation = Improper Node

Conditions:

- A singularity is detected
- $\Delta(A) = 0$
- Temporal consistency : minimum 1 sec

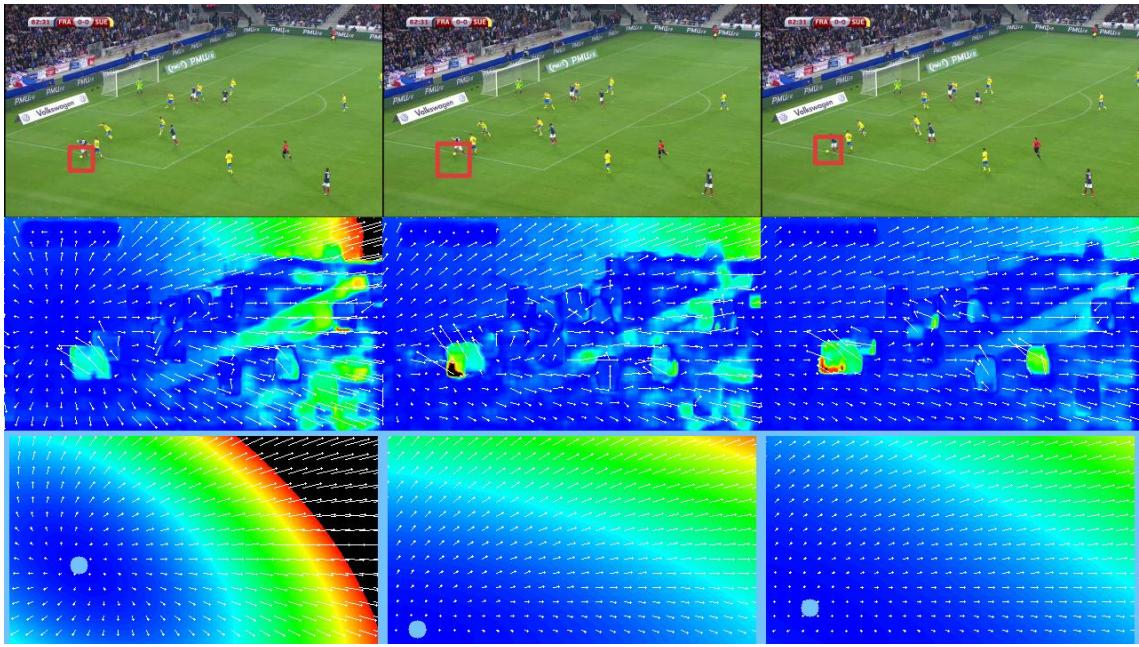


✓ Zoom's type ?

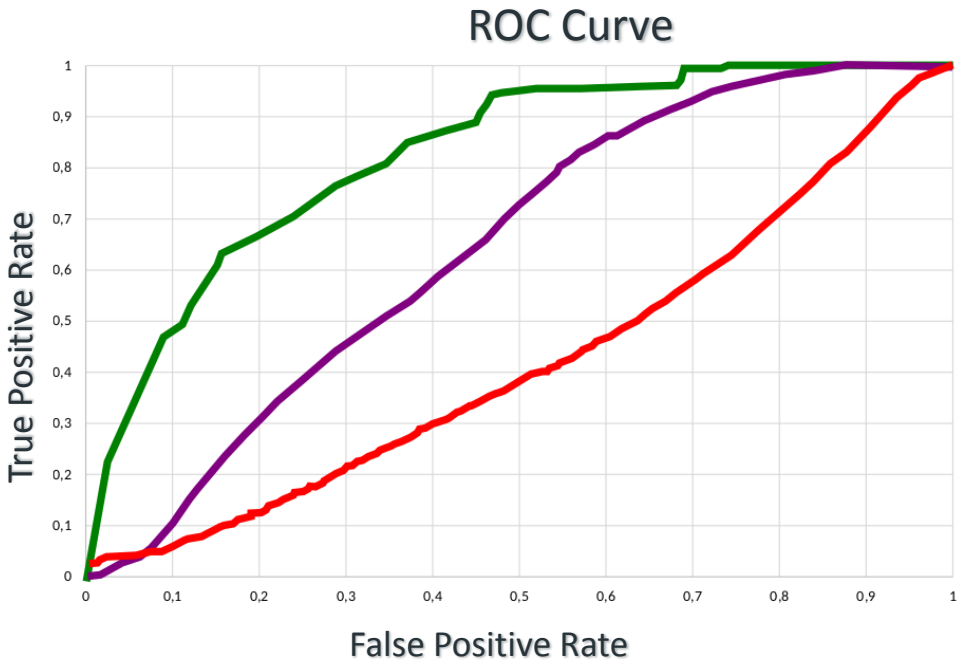
Positive eigen values -> zoom out

Negative eigen values -> zoom in

✓ Zoom center position



Zoom detection



Our method

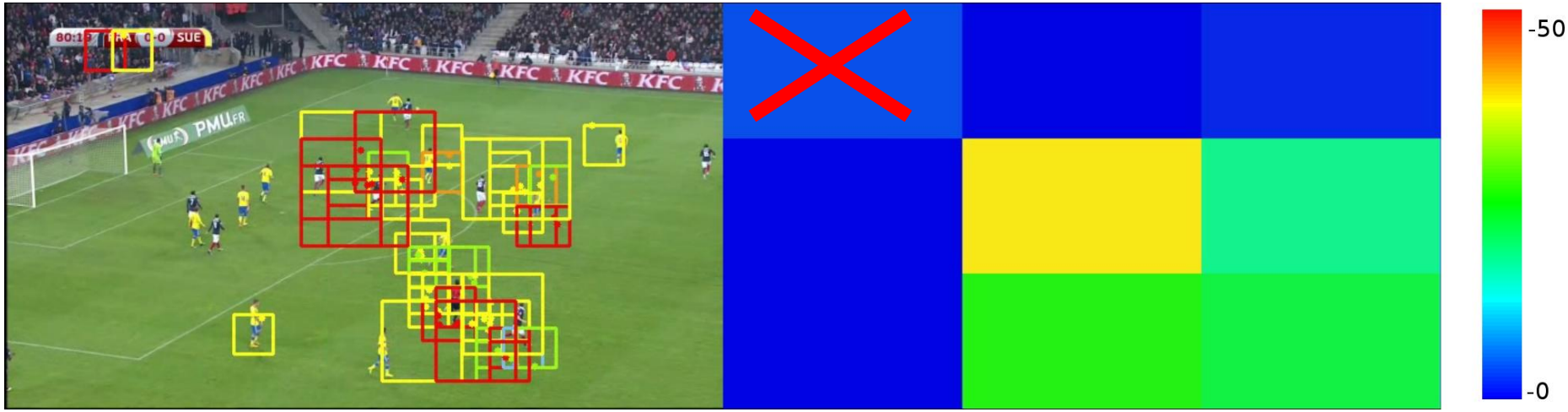
Global Motion Estimation(GME) [Qian, 2012]

Method of [Duan, 2005]

Method	Precision	Recall	Accuracy
GME	3.68 %	68.4 %	19.79 %
Duan	8.92 %	50.62 %	75.06 %
ours	19.45 %	63.47 %	86.82 %

Global Excitement

- 2D spatial histogram of 3x3 bins
- Histogram sum on 10 frames
- Threshold to select most agitated moment



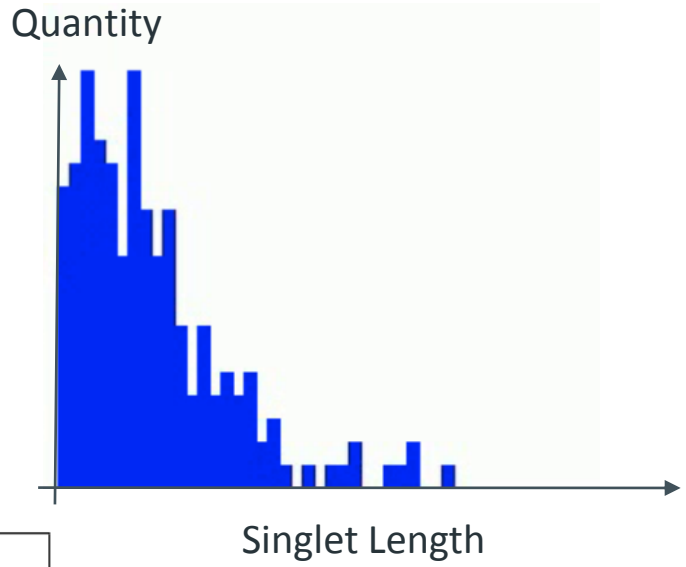
Slow Motion Detection

How to differentiate a fast motion that has been artificially slowed down and a real slow motion?

Slow Motion Detection

SVM Classification

from the histograms of singlet length

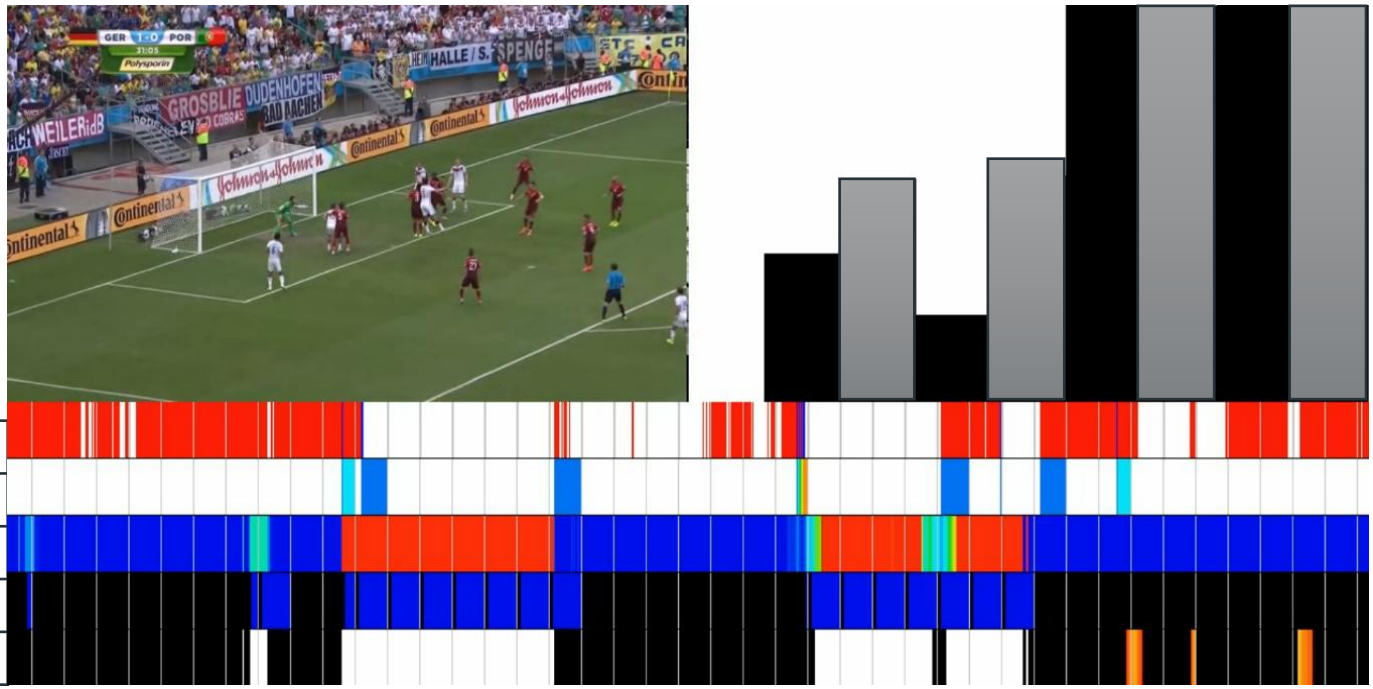


Database	Precision	Recall	Accuracy
Training	97.06 %	80.49 %	89.41 %
Test	76.32 %	87.88 %	79.36 %
Test on handball	100 %	20 %	60 %

Saliant moment in a match

Within 30 secondes

- At least two changes in zoom directions
- An activity peak higher than 1500 in a large view
- A slow motion in a close view



Saliant moment in a match

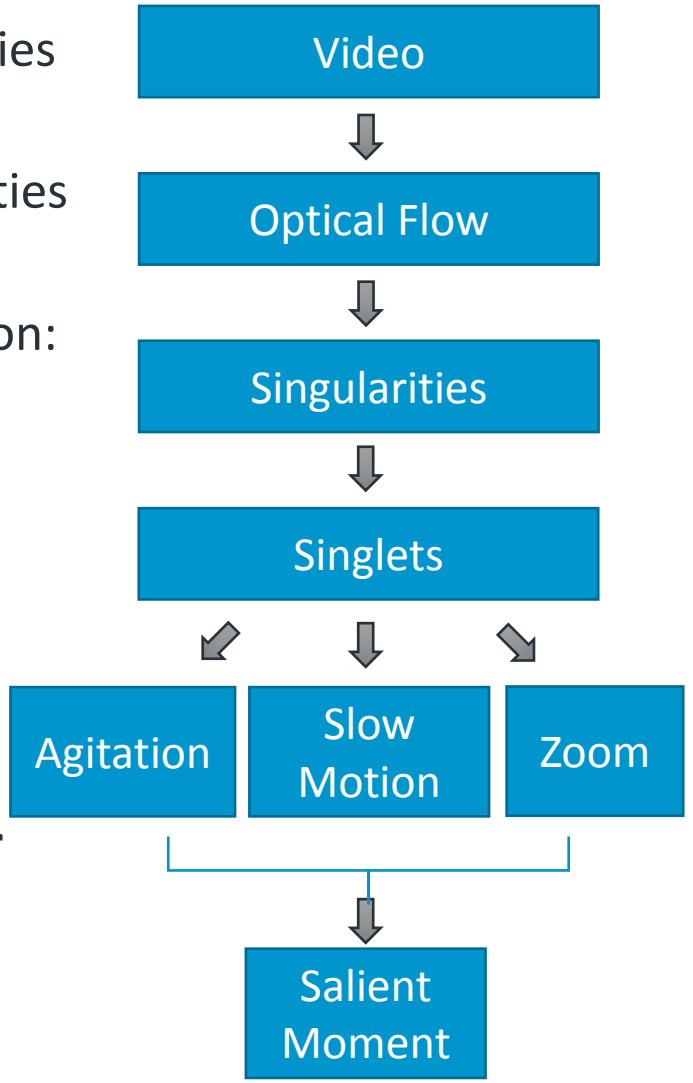
Within 30 secondes

- At least two changes in zoom directions
- An activity peak higher than 1500 in a large view
- A slow motion in a close view

Match	FIFA database	Extended database
Germany vs Portugal	80 %	88.9 %
Nigeria vs Argentine	53 %	77.2 %
France vs Honduras	53.7 %	90.7 %
Switzerland vs France	62.5 %	96.6 %
Mean accuracy	62.3 %	88.2 %

Singlets

- Local explicit motion description using singularities detection
- Tracking and accumulation of chains of singularities to describe a video shot
- Use of this description in a sport video abstraction: zooms, slowmotions, excitement and salient moments



K. Blanc, D. Lingrand and F. Precioso

Singlets: Multi-resolution Motion Singularities for Soccer Video Abstraction

IEEE Conference on Computer Vision and Pattern Recognition Workshop (CvSport), 2017.

Overview

Temporal Warping

- Temporal Warping in recognition
- DTW and its extensions
- Temporal alignment framework
- Results

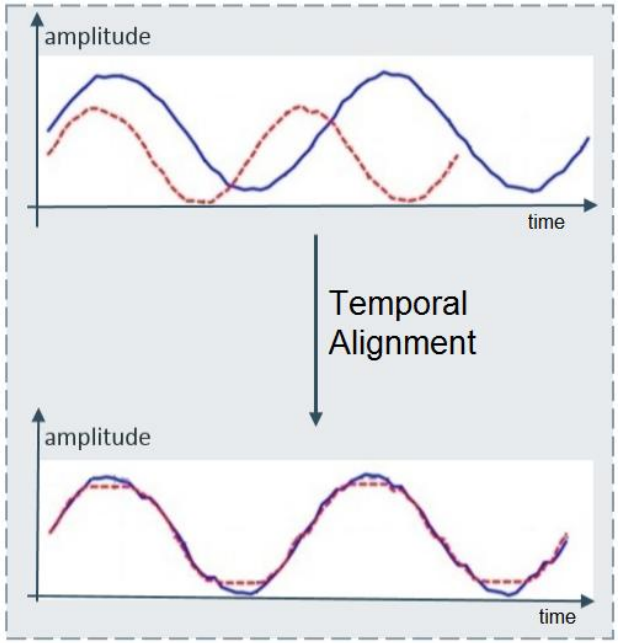
Dynamical Time Warping

Dynamical Time Warping (DTW) is used to temporally align sequences which are assumed correlated

-> Warp sequences to maximize the correlation

Related works:

- IMW: Split the action execution and the user style [Hsu, 2005]
- Match two viewpoints [Junejo, 2011]
- Align actions described by 3D point cloud [Vu, 2012]
- Extension de DTW: CTW [Zhou, 2009], DCTW [Trigeorgis, 2016], GCTW [Zhou, 2016]
- TCTW [Wang, 2017], DDATW [Trigeorgis, 2018]



The Alignment

Non-aligned videos

Aligned videos with gctw

DTW and its extensions

- DTW is used to align two sequences $X = [x_1, \cdots, x_{n_x}] \in \mathbb{R}^{d \times n_x}$ and $Y = [y_1, \cdots, y_{n_y}] \in \mathbb{R}^{d \times n_y}$ such that

$$\min_{\{p_x, p_y\} \in \Psi} J_{DTW} = \sum_{t=1}^{l} \| x_{p_t^x} - y_{p_t^y} \|^2$$

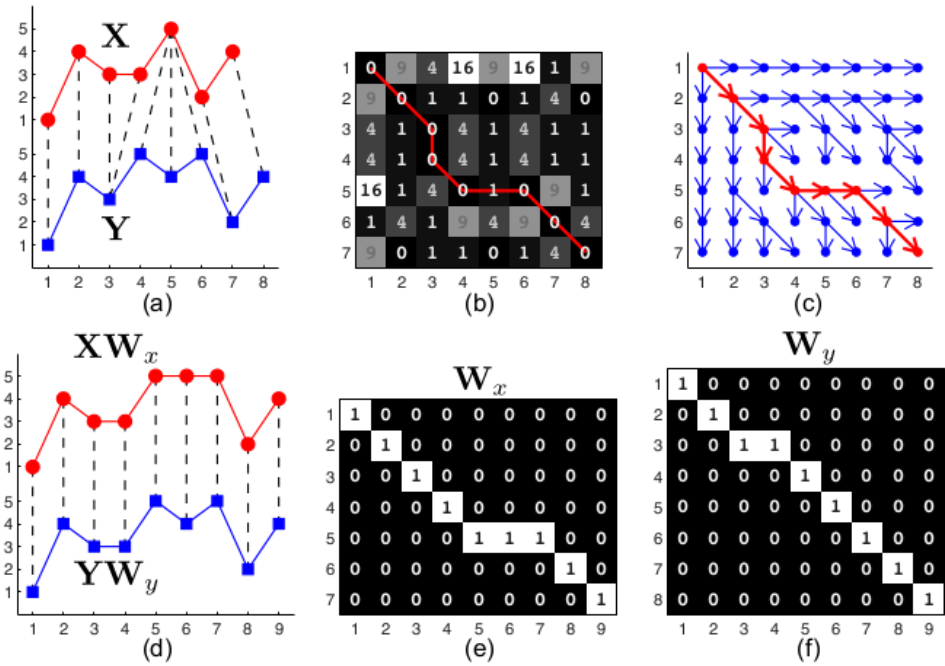
With the constraints :

✓ Boundary : $[p_1^x, p_1^y] = [1, 1]$ and $[p_l^x, p_l^y] = [n_x, n_y]$

✓ Monotony : $t_1 \leq t_2 \Rightarrow p_{t_1}^x \leq p_{t_2}^x$ and $p_{t_1}^y \leq p_{t_2}^y$

✓ Continuity :

$$[p_t^x, p_t^y] - [p_{t-1}^x, p_{t-1}^y] \in \{[0, 1], [1, 0], [1, 1]\}$$



DTW and its extensions

$$\min_{\{p_x,p_y\}\in\Psi} J_{DTW} = \sum_{t=1}^{l} \|x_{p_t^x} - y_{p_t^y}\|^2$$

- Canonical Time Warping (CTW) [Zhou, 2009]:
  -> adding a **linear transform**

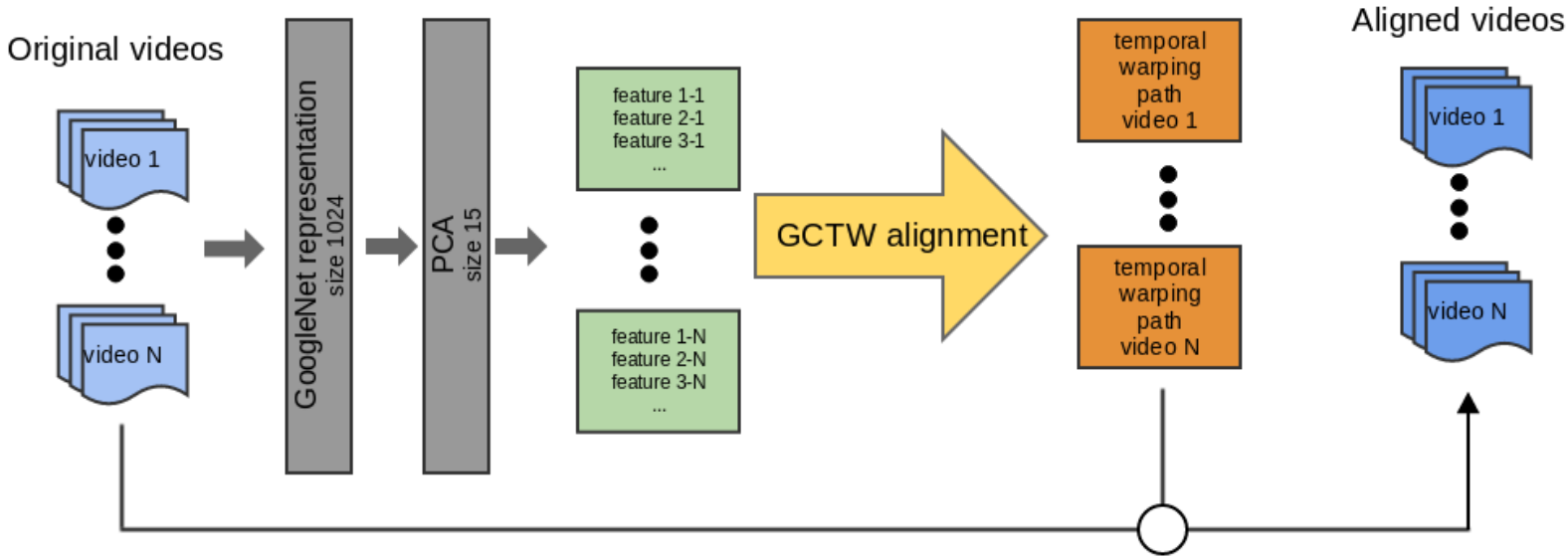
$$\min_{\{V_x,V_y\}\in\Phi,\{p_x,p_y\}\in\Psi} J_{CTW} = \|V_x^T X W_x - V_y^T Y W_y\|_F^2 + \phi(V_x) + \phi(V_y)$$

- Generalized Canonical Time Warping(GCTW) [Zhou, 2016]:
  -> adding **several** sequences and temporal transform are parametrized

$$\min_{\{V_i\}_i\in\Phi,\{p_i\}_i\in\Psi} J_{GCTW} = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{1}{2}\|V_i^T X_i W_i - V_j^T X_j W_j\|_2^F + \sum_{i=1}^{m}(\phi(V_i)+\psi(p_i))$$

Temporal Alignment Framework

To normalize the speed execution within a class



The databases

Video Database on **sign language**: ASL and IsoGD

The Temporal Alignment

Original videos

Aligned videos

Classification protocols

Prediction made by a 3D convolutional neural network (C3D)

Classification protocols:

- **Baseline** : Training and Prediction on the non-aligned database
- **Protocol 1**: Training on aligned videos and Testing on non-aligned videos
- **Protocol 2**: Training and Prediction on the aligned database
- **Protocol 3**: Training on aligned video and Testing on video aligned with each class

Classification protocol	ASL		IsoGD	
	top-1 acc	top-5 acc	top-1 acc	top-5 acc
Baseline	76.7	96.2	45	89.05
Protocol 1	14.5	37.9	41.65	71.03
Protocol 2	91.7	98.7	81.27	97.01
Protocol 3	50.4	92	81.34	97.11

Conclusion: Temporal Alignment

The recognition on the aligned videos is higher than on the normalised database

-> The alignment of the execution speed improves the C3D classification

The protocol 3 improves the recognition rate on IsoGD

K. Blanc, D. Lingrand, A. Paladini, L. Coviello, D. Mitrev, E. Söhler, L. Guzman and F. Precioso

Analysis of temporal alignment for Video Classification

IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2019.

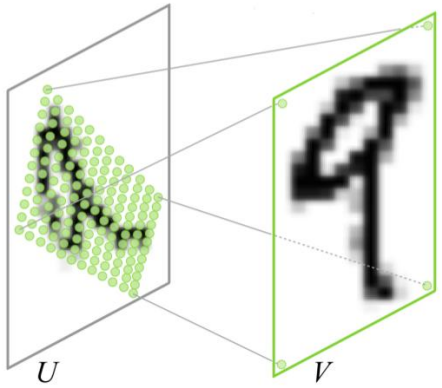
Perspective: Spatial Transformer Network

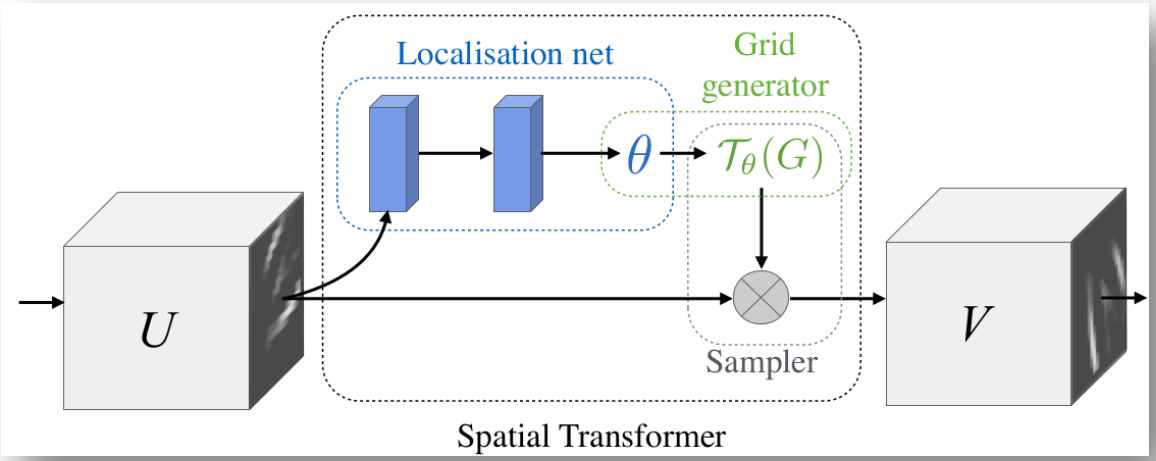
[Jaderberg, 2015]

$$J_{GCTW} = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|V_i^T X_i W_i - V_j^T X_j W_j\|_2^F$$

Network in 2 parts:

- Localisation
- Prediction of the affine transform parameters θ
- Representation and Classification





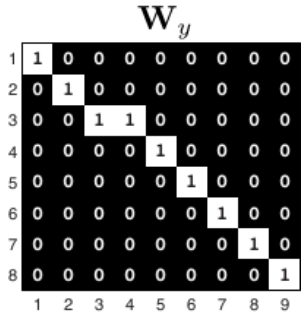
[Jaderberg, 2015]

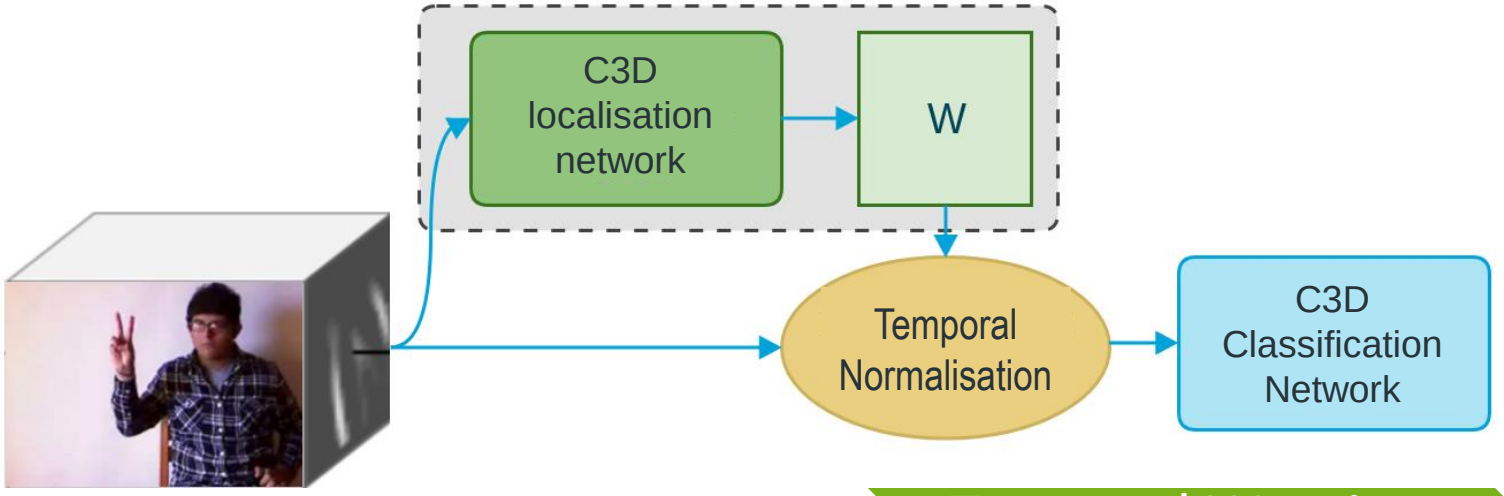
Perspective: Temporal Transformer Network

Adaption of the network for a temporal transform:

- The input U is a video of length n
- The localisation network is inpired from C3D
- The output of the 1^{st} network is the temporal transform matrix $W \in \mathbb{R}^{N \times n}$
- The alignment function is $U_a = W^T.U$
- The prediction part is inpired from C3D

$$J_{GCTW} = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|V_i^T X_i W_i - V_j^T X_j W_j\|_2^F$$





Conclusion

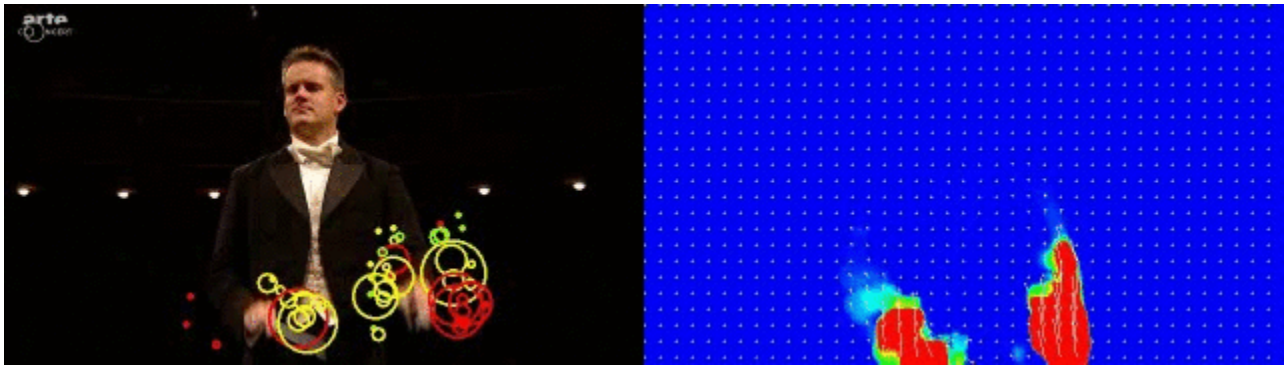
- The videos, the motions are relevant content and can be used to reduce the redundancy
- Need of a robust video description
- 2 applications using motion description:

  - Local and explicit description: *Les Singlets*
    - ➢ *Motion Detection in Football matchs*

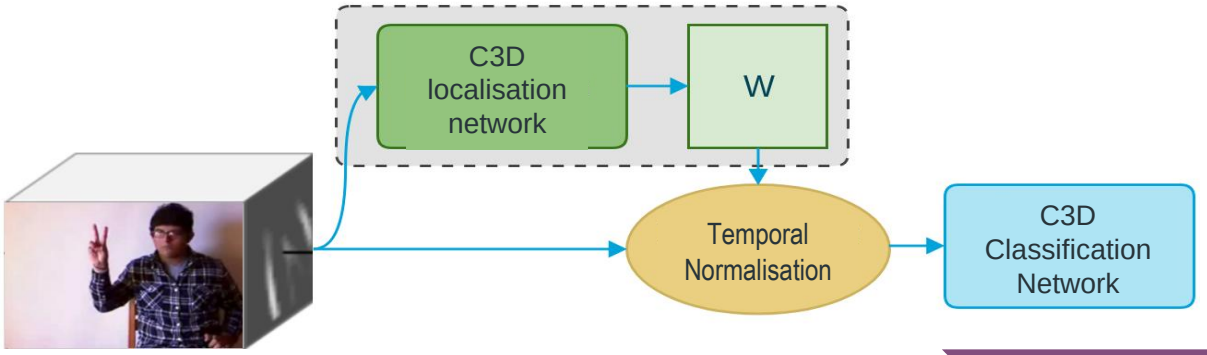
  - Dynamical Time Warping: Alignment of execution speed within each class and prediction protocols
    - ➢ Sign Language classification

Perspectives

- Singularities for small movements



- Temporal normalisation as a first part of a neural network



Katy Blanc

Detection of salient motions
and temporal alignment in videos

The 2nd April 2019

References

[Laptev, 2005] I. Laptev. **On space-time interest points**. *International journal of computer vision, 64(2-3) :107–123, 2005.*

[Wang, 2011] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. **Action recognition by dense trajectories**. *In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3169–3176. IEEE, 2011*.

[Carreira, 2017] J. Carreira and A. Zisserman. **Quo vadis, action recognition ? a new model and the kinetics dataset**. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733. IEEE, 2017.*

[Diba, 2017] A. Diba, V. Sharma, and L. Van Gool. **Deep temporal linear encoding networks.** In Computer Vision and Pattern Recognition, 2017.

[Feichtenhofer ,2016] C. Feichtenhofer, A. Pinz, and R. Wildes. **Spatiotemporal residual networks for video action recognition**. *In Advances in neural information processing systems, pages 3468–3476, 2016.*

[Wang, 2018] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. **Temporal segment networks for action recognition in videos.** *IEEE transactions on pattern analysis and machine intelligence, 2018*

[Li, 2018] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek. **VideoLSTM convolves, attends and flows for action recognition**. *Computer Vision and Image Understanding, 166 :41–50, 2018.*

[Bilen, 2016] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. **Dynamic image networks for action recognition**. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3034–3042, 2016.*

[Fernando, 2015] B. Fernando, S. Gavves, O. Mogrovejo, J. Antonio, A. Ghodrati, and T. Tuytelaars. **Modeling video evolution for action recognition**. *In Proceedings CVPR 2015, pages 5378–5387, 2015.*

[Edison, 2017] A. Edison and C. Jiji. **Optical acceleration for motion description in videos.** *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 39–47, 2017*

References

[Ye, 2005] Q. Ye, Q. Huang, W. Gao, and S. Jiang. **Exciting event detection in broadcast soccer video with mid-level description and incremental learning.** *In Proceedings of the 13th annual ACM international conference on Multimedia, pages 455–458. ACM, 2005.*

[Zawbaa, 2012] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. Kim**. Event detection based approach for soccer video summarization using machine learning.** *International Journal of Multimedia and Ubiquitous Engineering, 7(2) :63–80, 2012.*

[Raventos, 2014] A. Raventos, R. Quijada, L. Torres, and F. Tarres. **Automatic summarization of soccer highlights using audio-visual descriptors.** *arXiv preprint arXiv :1411.6496, 2014.*

[Druon, 2006] Druon, Martin, Benoit Tremblais, and Bertrand Augereau. **Vector fields modelization using basis of polynomials: Application to the analysis of simple face movements.** *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Vol. 2. IEEE, 2006.*

[Kihl, 2008] O. Kihl, B. Tremblais, and B. Augereau. **Multivariate orthogonal polynomials to extract singular points.** *In IEEE International Conference on Image Processing 2008. ICIP 2008., pages –, San Diego, CA, United States, Oct. 2008.*

[Druon, 2009] Druon, Martin**. Modélisation du mouvement par polynômes orthogonaux: application à l'étude d'écoulements fluides**. Diss. Université de Poitiers, 2009.

[Safdarnejad, 2015] Safdarnejad, Seyed Morteza, Xiaoming Liu, and Lalita Udpa. **Robust Global Motion Compensation in Presence of Predominant Foreground**. *BMVC. 2015.*

[Qian, 2012] X. Qian. **Global Motion Estimation and Its Applications**. *INTECH Open Access Publisher, 2012*

[Duan, 2005] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin. **A unified framework for semantic shot classification in sports video.** *IEEE Transactions on Multimedia, 7(6) :1066–1083, 2005.*

[De Lathauwer, 2000] L. De Lathauwer, B. De Moor, and J. Vandewalle. **A multilinear singular value decomposition.** *SIAM journal on Matrix Analysis and Applications, 21(4) :1253–1278, 2000.*

[Lebedev, 2014] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky. **Speeding-up convolutional neural networks using fine-tuned cp-decomposition**. *arXiv preprint arXiv :1412.6553, 2014*

References

[Novikov, 2015] Novikov Alexander, Podoprikhin Dmitrii, Osokin Anton, **Tensorizing neural networks.** *In : Advances in Neural Information Processing Systems. p. 442-450. 2015.*

[Garipov, 2016] T. Garipov, D. Podoprikhin, A. Novikov, and D. Vetrov. **Ultimate tensorization : compressing convolutional and fc layers alike.** *arXiv preprint arXiv :1611.03214, 2016.*

[Yang, 2017] Y. Yang, D. Krompass, and V. Tresp. **Tensor-train recurrent neural networks for video classification**. *arXiv preprint arXiv :1707.01786, 2017.*

[Kossaifi, 2017] J. Kossaifi, A. Khanna, Z. Lipton, T. Furlanello, and A. Anandkumar. Tensor contraction layers for parsimonious deep nets. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 1940–1946. IEEE, 2017.

[Ben-Younes, 2017] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome. Mutan : Multimodal tucker fusion for visual question answering. In Proc. IEEE Int. Conf. Comp. Vis, volume 3, 2017

References

[Zhao, 2013] Q. Zhao, C. F. Caiafa, D. P. Mandic, Z. C. Chao, Y. Nagasaka, N. Fujii, L. Zhang, and A. Cichocki. **Higher order partial least squares (hopls) : a generalized multilinear regression method.** *IEEE transactions on pattern analysis and machine intelligence, 35(7) :1660–1673, 2013.*

[Zhou, 2016] G. Zhou, A. Cichocki, Y. Zhang, and D. P. Mandic. **Group component analysis for multiblock data : Common and individual feature extraction.** *IEEE transactions on neural networks and learning systems, 27(11) :2426–2439, 2016.*

[Hsu, 2005] E. Hsu, K. Pulli, and J. Popović. **Style translation for human motion.** *In ACM Transactions on Graphics (TOG), volume 24, pages 1082–1089. ACM, 2005*

[Junejo, 2011] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez**. View-independent action recognition from temporal self-similarities**. *IEEE transactions on pattern analysis and machine intelligence, 33(1) :172–185, 2011.*

[Vu, 2012] H. T. Vu, C. Carey, and S. Mahadevan. **Manifold warping : Manifold alignment over time.** *In AAAI, volume 1, page 8, 2012*

[Zhou, 2009] F. Zhou and F. Torre. **Canonical time warping for alignment of human behavior.** *In Advances in neural information processing systems, pages 2286–2294, 2009.*

[Trigeorgis, 2016] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller. **Deep canonical time warping.** *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5110–5118, 2016.*

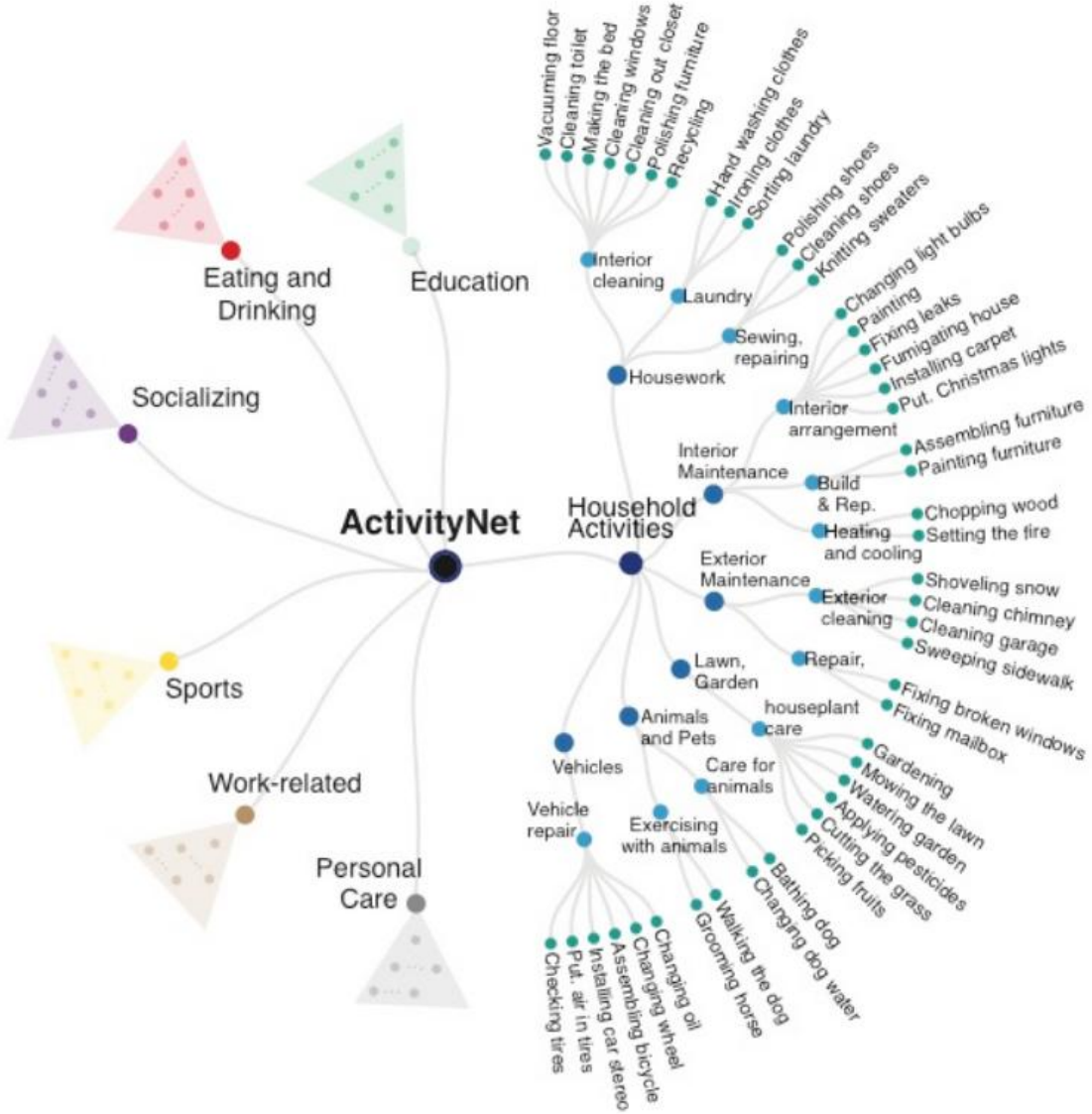
[Zhou, 2016] F. Zhou and F. De la Torre. **Generalized canonical time warping**. *IEEE transactions on pattern analysis and machine intelligence, 38(2) :279–294, 2016.*

[Wang, 2017] H. Wang, W. Yang, C. Yuan, H. Ling, and W. Hu. **Human activity prediction using temporally-weighted generalized time warping.** *eurocomputing, 225 :139–147, 2017*

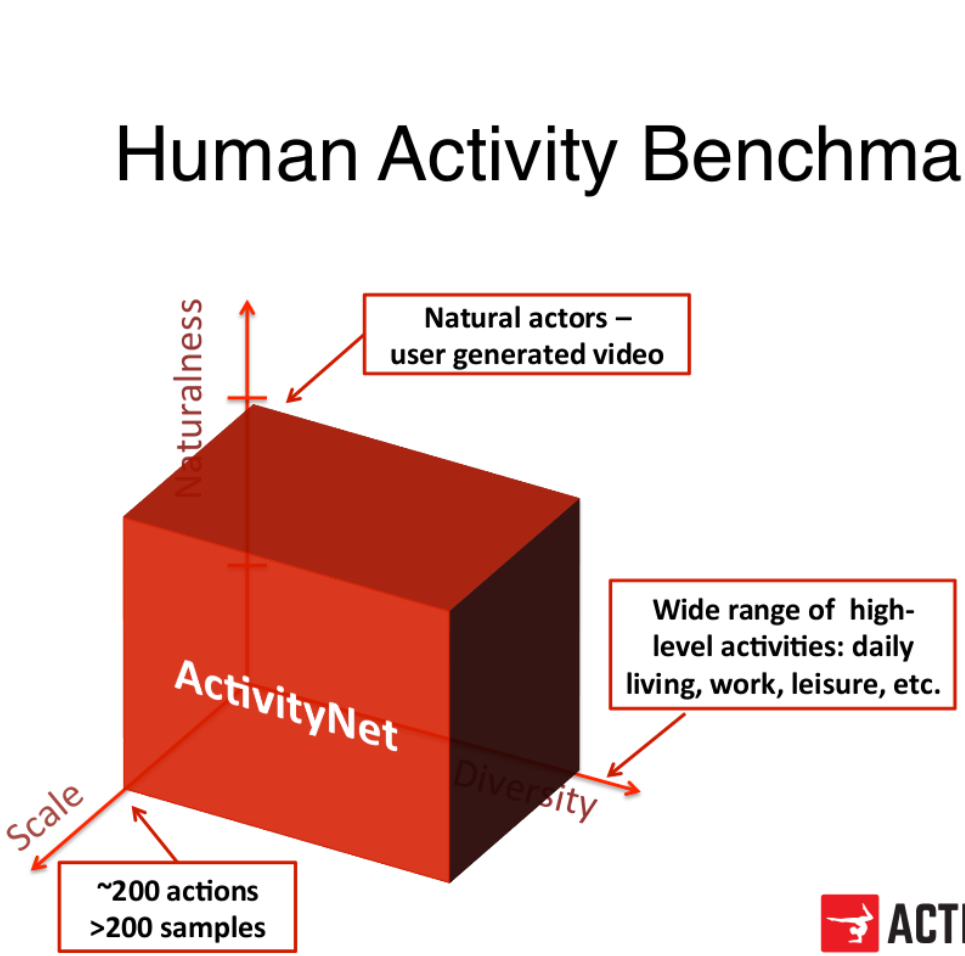
[Trigeorgis, 2018] G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. **Deep canonical time warping for simultaneous alignment and representation learning of sequences.** *IEEE Trans. PAMI, pages 1128–1138, 2018.*

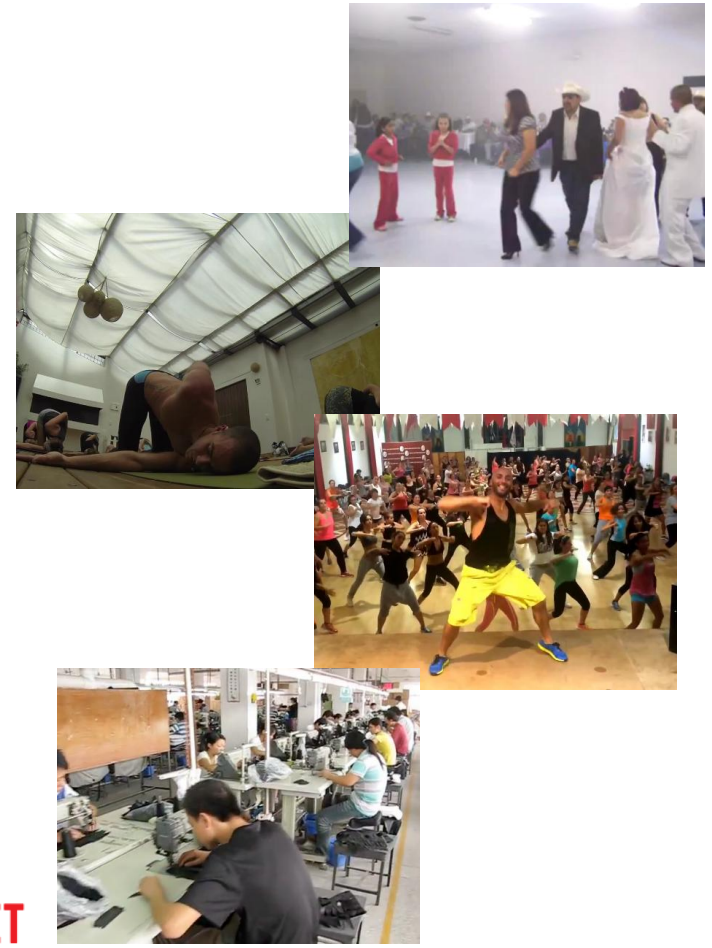
[Jaderberg, 2015] M. Jaderberg, K. Simonyan, A. Zisserman, et al. **Spatial transformer networks.** *In NIPS, pages 2017–2025, 2015.*

ActivityNet Dataset : 272 classes, 20K vidéos



ActivityNet Dataset : 272 classes, 20K vidéos





Football related works

Ye et al, Exciting Event Detection in Broadcast Soccer Video with Mid-level Description and Incremental Learning

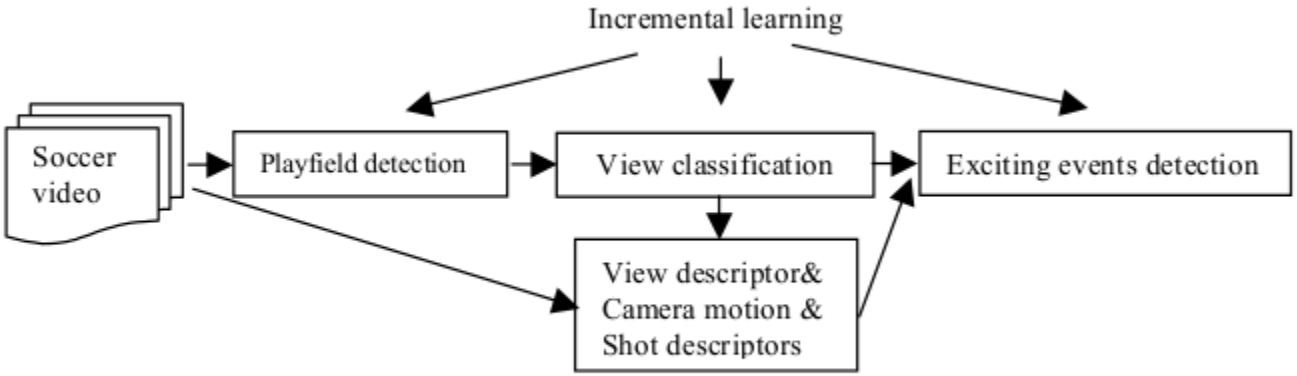


Figure 1. Framework for exciting event detection.

Zwabaa et al, Event detection based approach for soccer video summarization using machine learning

Logos and scores -> recognition by NN ; Hough method to detect goal mouth;
Classification of viewpoints and the shot length to get the play time and the break
Audio detection of crowd excitation

Optical Flow Estimation

Gunnar Farneback method:

$$f(x) = x^T A x + b^T x + c$$

Pixel Neighbourhood approximation

Approximation Translation :

$$f_2(x) = f_1(x - v) = (x - d)^T A_1 (x - d) + b_1^T (x - d) + c_1 = x^T A_2 x + b_2^T x + c_2$$

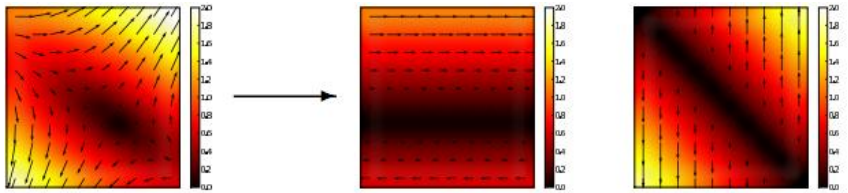
Matching :

$$A_2 = A_1$$
$$b_2 = b_1 - 2A_1 v$$
$$c_2 = v^T A_1 v - b_1^T v + c_1$$

Speed vector estimation

$$v = \frac{-1}{2} A_1^{-1} (b_2 - b_1)$$

$$\mathcal{F}: \quad \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^2$$
$$(x_1, x_2) \mapsto (\mathcal{V}^1(x_1, x_2), \mathcal{V}^2(x_1, x_2))$$



$\mathcal{F}$
Champ original

$\mathcal{V}^1$
Mouvements horizontaux

$\mathcal{V}^2$
Mouvements verticaux

Discussion: Protocol 3

The protocol 3: Alignment of a video to a class

→ Que se passe-t-il ?

X the video, c_1 its class, c_2 another class, W_{c_i} the alignment path for the class c_i and $F_{C3D}^{c_i}$ the score function of the class c_i by the network C3D trained on the aligned base

On ASL, the new execution speed specific to the class c_2 obtains a higher score than the new execution speed specific to the class c_1

$$F^{c_1}(XW_{c_1}) < F^{c_2}(XW_{c_2})$$

On IsoGD:

$$F^{c_2}(XW_{c_2}) < F^{c_1}(XW_{c_1})$$

The temporal normalisation depends on the input video and the target alignment class

Generative method (diminution of the intra-class variation)

-> Add a discriminant criterion to increase the inter-class variation after the alignment