Spacetime Forests with Complementary Features for Dynamic Scene Recognition

Christoph Feichtenhofer, Axel Pinz,
and Richard P. Wildes

Dynamic Scene Classification

- Assign a category label (e.g. beach, river, forest fire, highway, …) to a video




Challenges (1)

- Typical image classification challenges
- Changes in
  - Viewpoint
  - Illumination
  - Scale
  - Appearance
  - Background

Challenges (2)

- Small inter-class differences

river

waterfall

fountain

river

waterfall

Challenges (3)

- Large intra-class variations



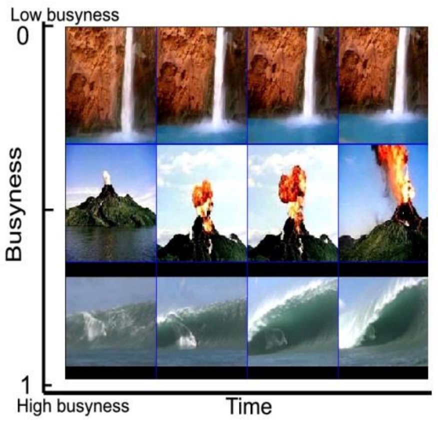


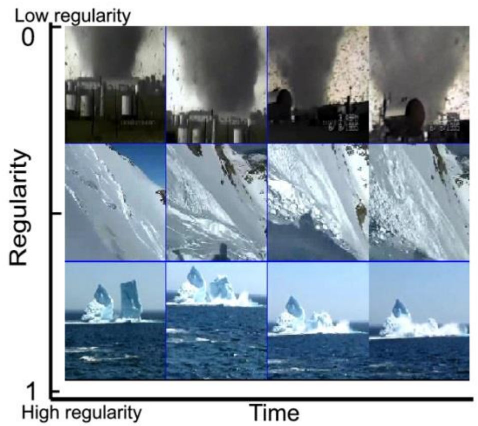
Challenges (4)

- Camera movement


- Scene cuts

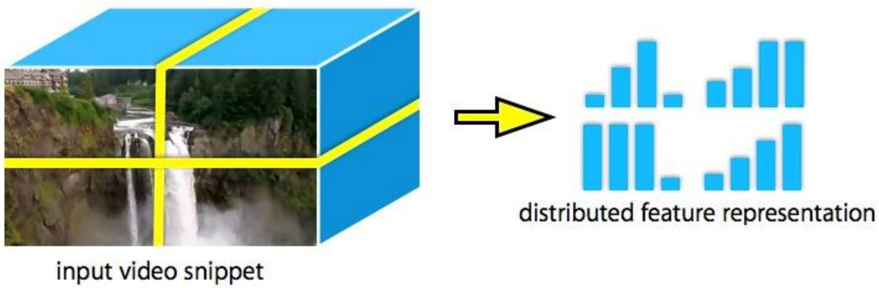

Related work





Combination of GIST and chaotic invariants

- Shroff et al. CVPR'10

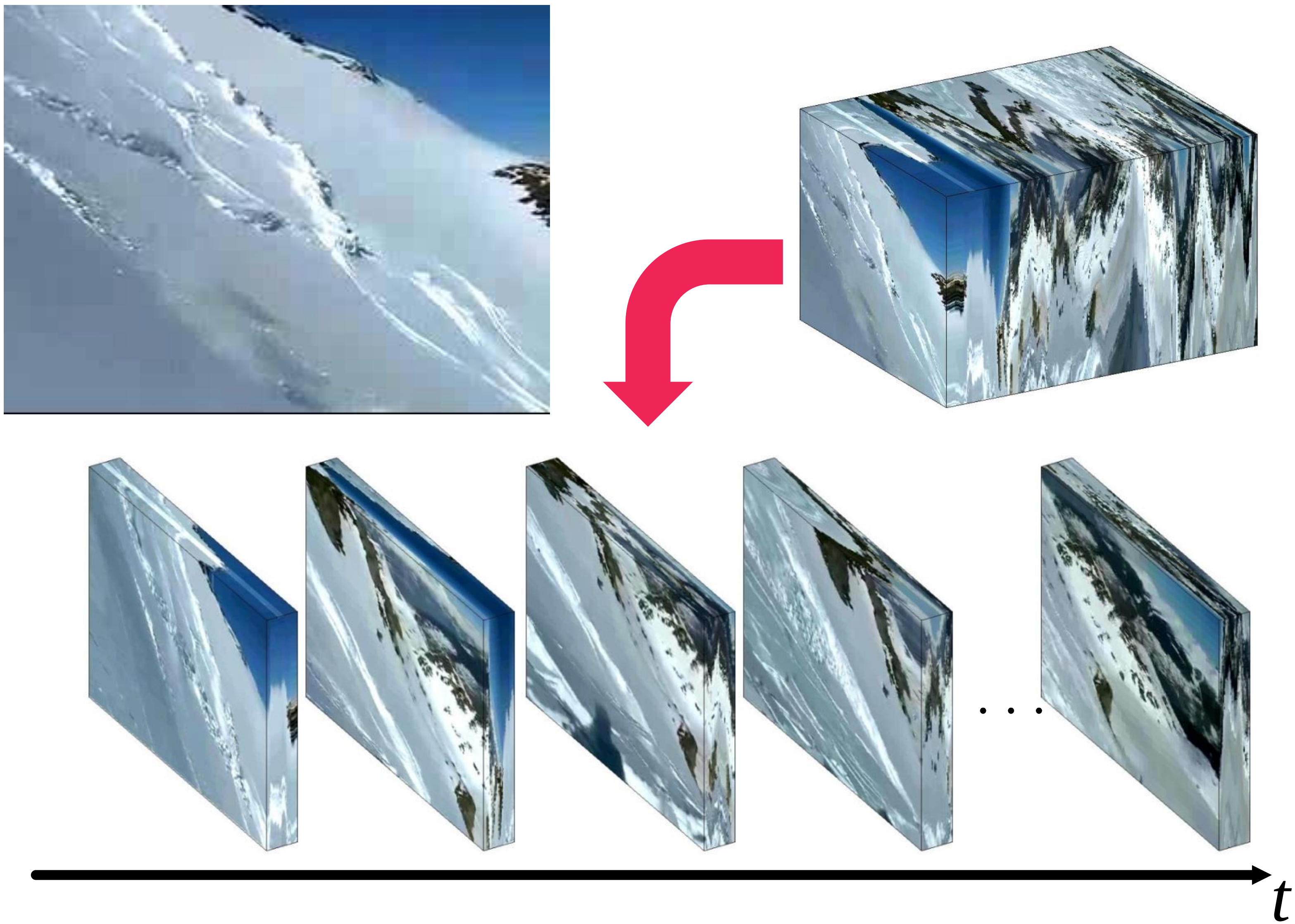


Spacetime Orientation Features

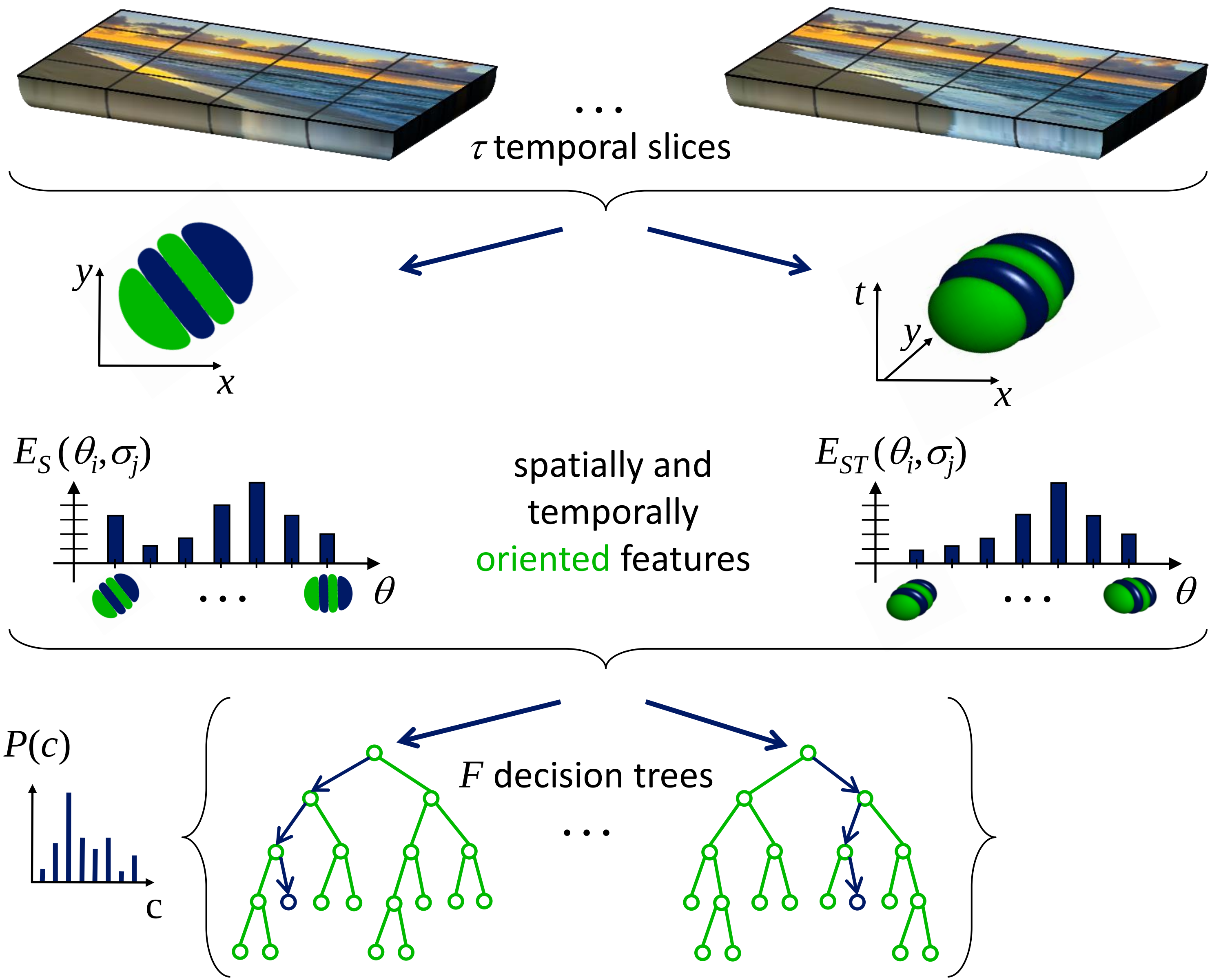
- Derpanis et al. CVPR'12

Existing methods compute a feature vector for the whole input sequence!

Proposed temporal slicing



Approach overview

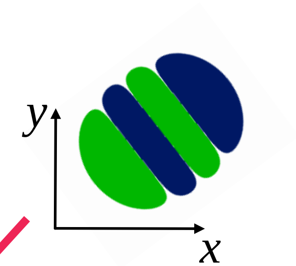


Complementary Spacetime Orientation (CSO) descriptor: Spatial information

$(x, y)^{\mathrm{T}}$

orientation

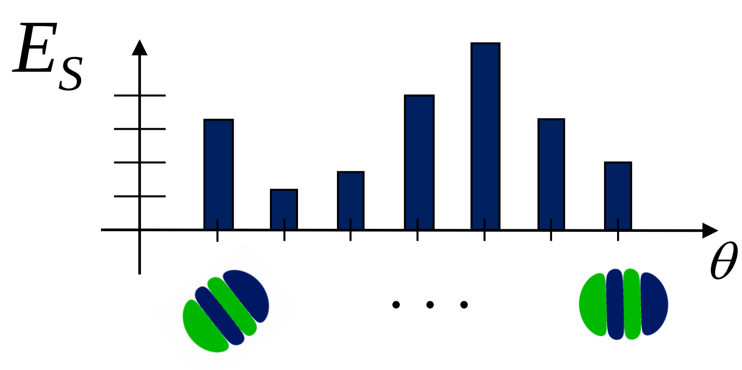
scale



image

$$E_S(\mathbf{x}; \theta_i, \sigma_j) = \sum_{\Omega} \left| G^{(3)}_{2D}(\theta_i, \sigma_j) * \mathcal{I}(\mathbf{x}) \right|^2$$

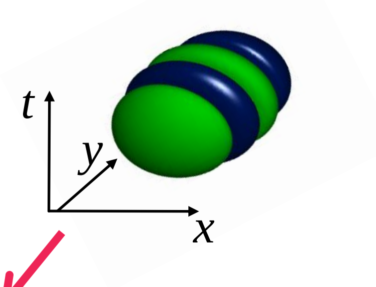
aggregation region



Histogram of spatially oriented energies

Complementary Spacetime Orientation (CSO) descriptor: Temporal information

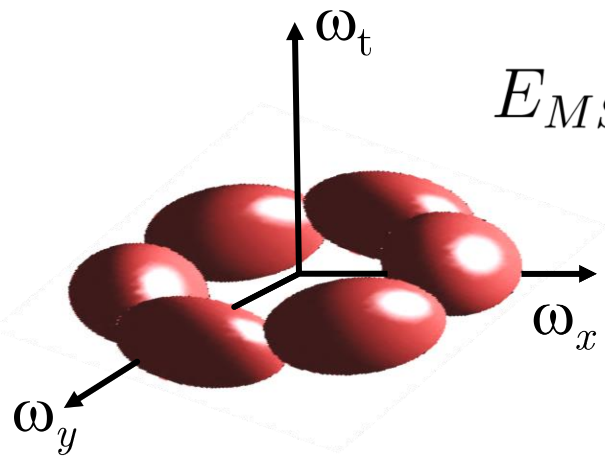
$(x,y,t)^T$ 3D orientation scale



spacetime volume

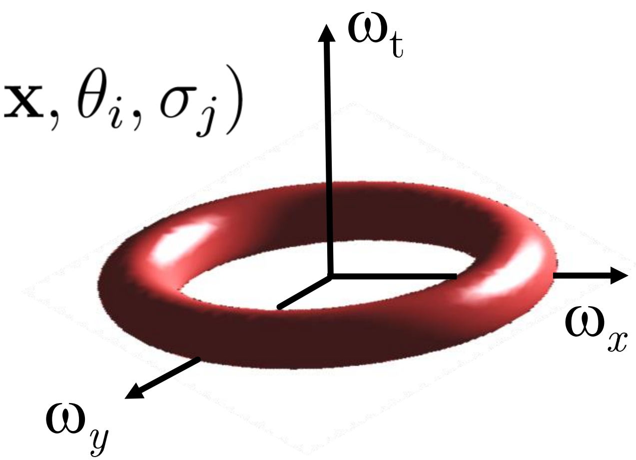
$$E_{ST}(\mathbf{x};\theta_i,\sigma_j) = \sum_{\Omega} |G_{3D}^{(3)}(\theta_i,\sigma_j) * \mathcal{V}(\mathbf{x})|^2$$

$$E_{MST}(\mathbf{x};\hat{\mathbf{n}},\sigma_j) = \sum_{i=0}^{N} E_{ST}(\mathbf{x},\theta_i,\sigma_j)$$



N+1 motion direction consistent energy samples

Sum across frequency domain plane $\hat{\mathbf{n}}$



planar energy samples

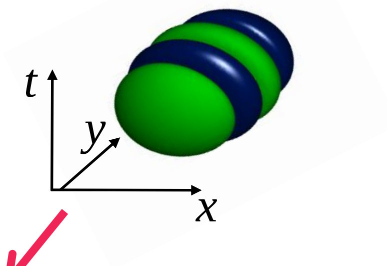
[Derpanis & Wildes, PAMI'12]

Complementary Spacetime Orientation (CSO) descriptor: Temporal information

$(x,y,t)^{\mathrm{T}}$

3D orientation

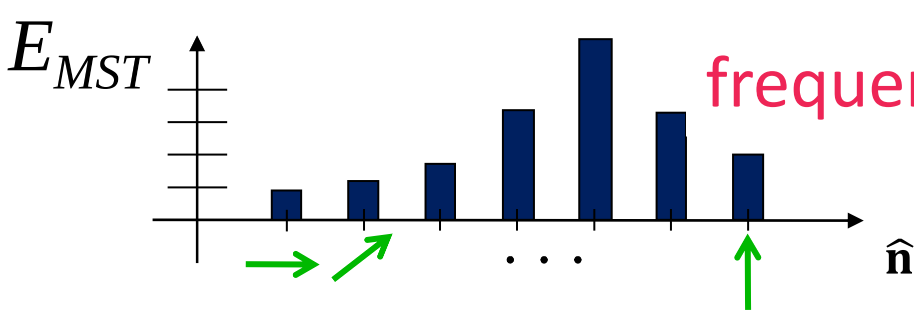
scale



spacetime volume

$$E_{ST}(\mathbf{x};\theta_i,\sigma_j)=\sum_{\Omega}|G_{3D}^{(3)}(\theta_i,\sigma_j)*\mathcal{V}(\mathbf{x})|^2$$

$$E_{MST}(\mathbf{x};\hat{\mathbf{n}},\sigma_j)=\sum_{i=0}^{N}E_{ST}(\mathbf{x},\theta_i,\sigma_j)$$



frequency domain plane

Temporal energy across various directions

Local contrast normalization

- Filter responses are a joint function of space(time) orientation and **contrast**

$$\hat{E}_S(\mathbf{x}, \theta_i, \sigma_j) = \frac{E_S(\mathbf{x}, \theta_i, \sigma_j)}{\sum_i^N E_S(\mathbf{x}, \theta_i, \sigma_j) + \epsilon}$$

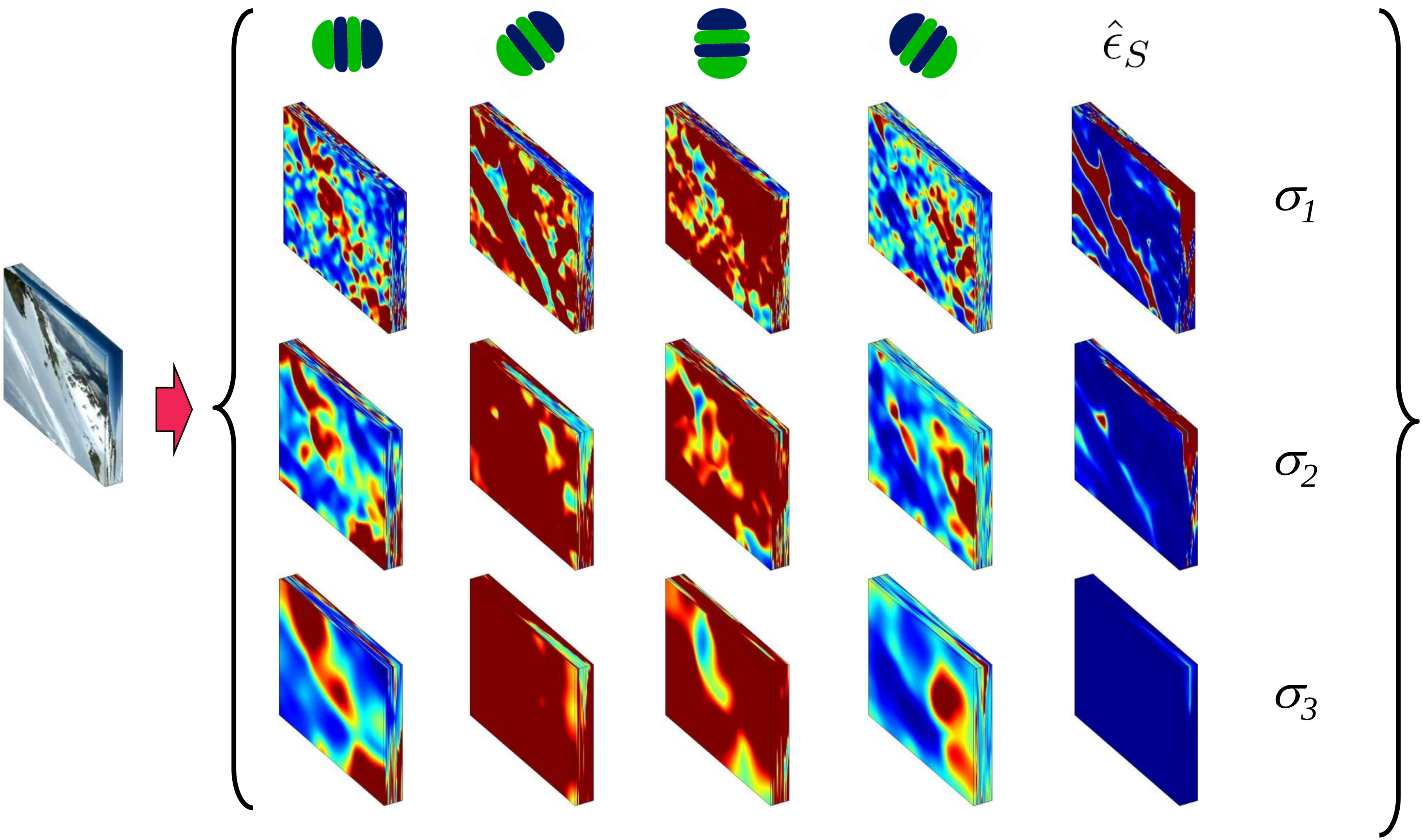
set of oriented energy measurements

small bias added for stability

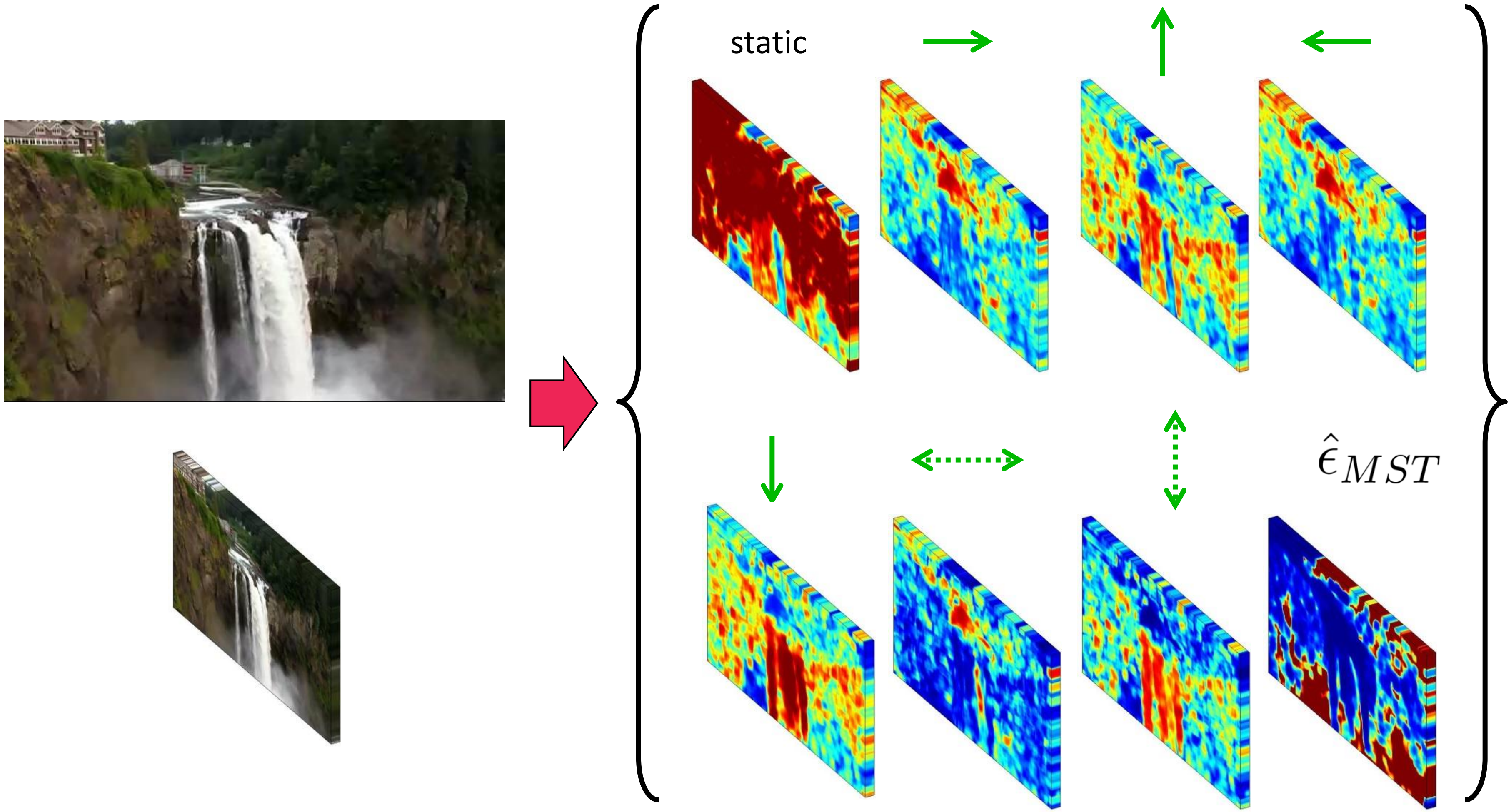
- Unstructuredness indicated by

$$\hat{\epsilon}_S = \frac{\epsilon}{\sum_{i=1}^N E_S(\mathbf{x}, \theta_i, \sigma_j) + \epsilon}$$

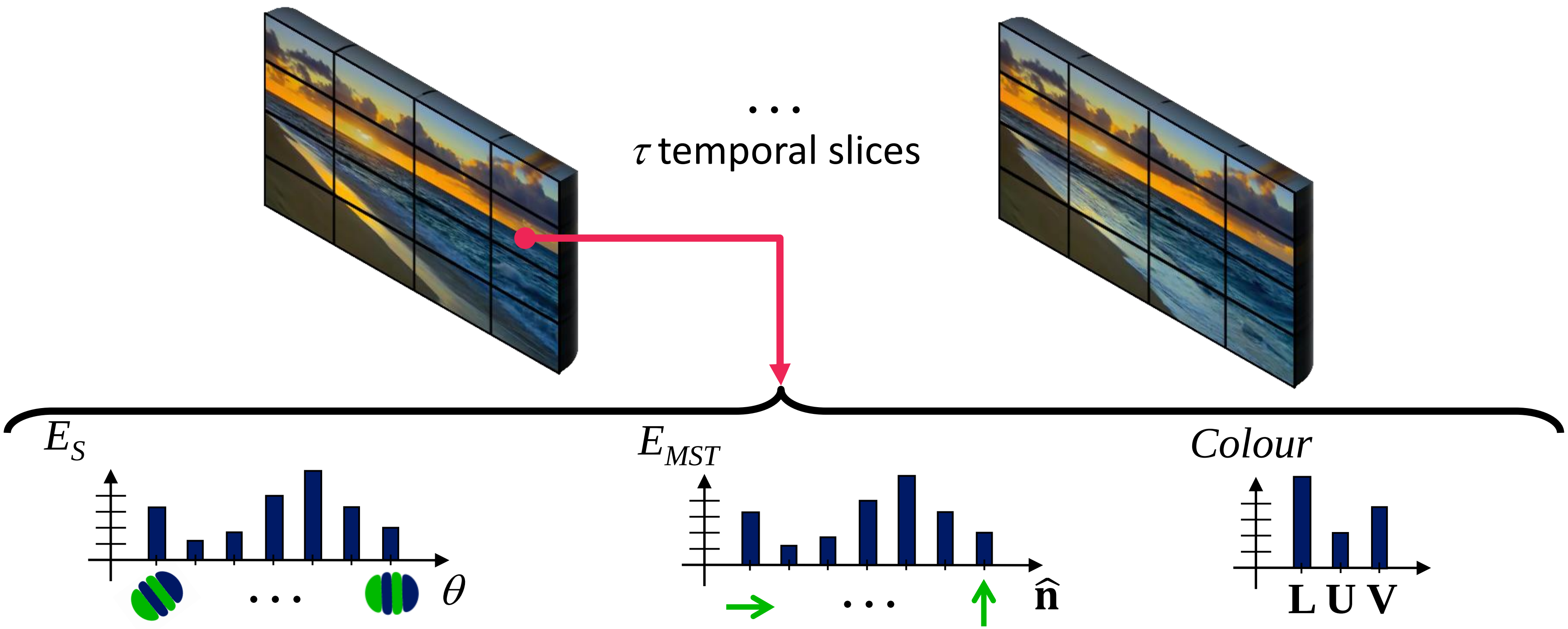
Complementary Spacetime Orientation (CSO) descriptor: Spatial information



Complementary Spacetime Orientation (CSO) descriptor: Temporal information

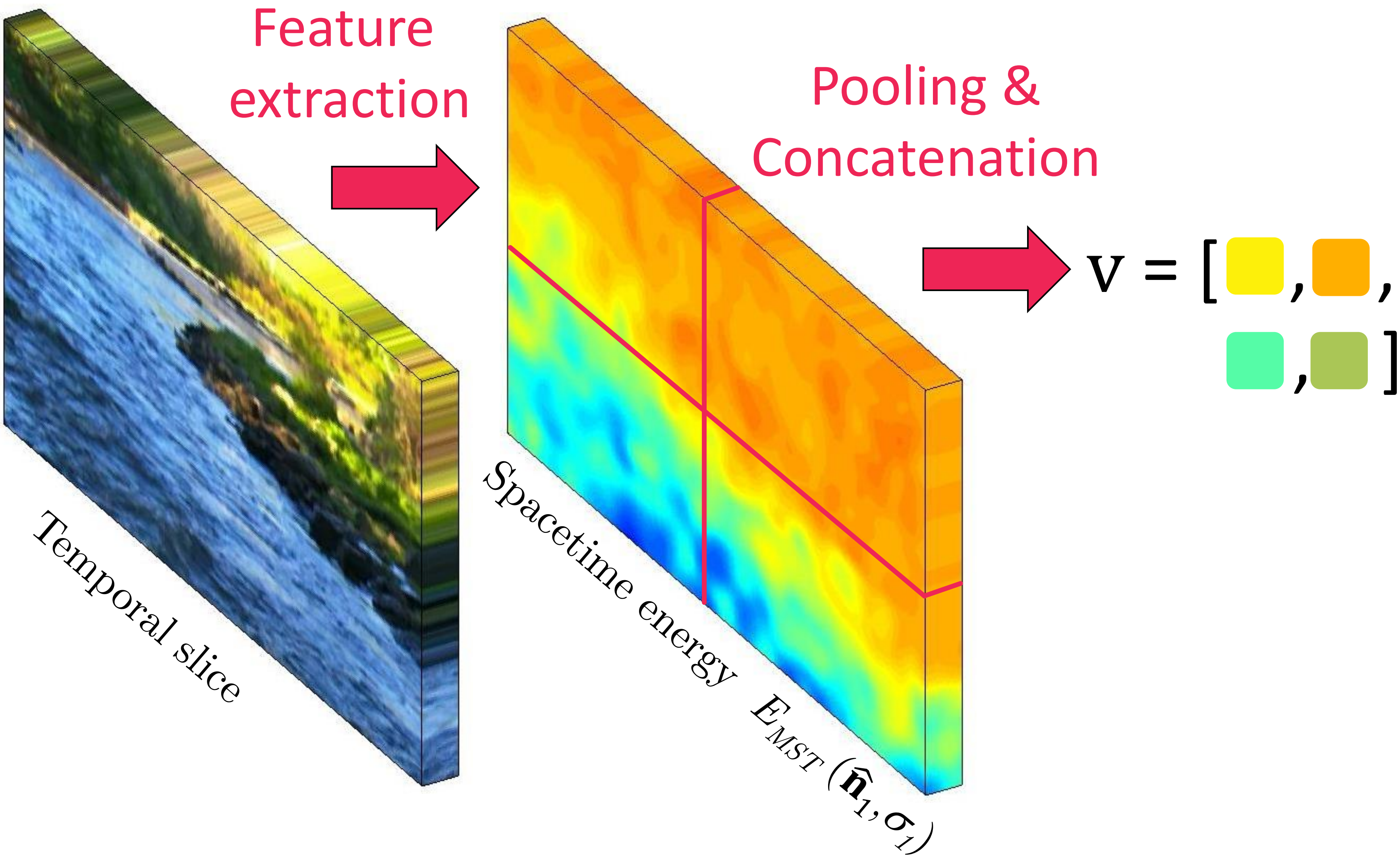


Complementary Spacetime Orientation (CSO) descriptor: Colour information



- 3 bin histogram of the LUV colour channels
- The complementary features are aggregated into histograms to form a spatial pyramid

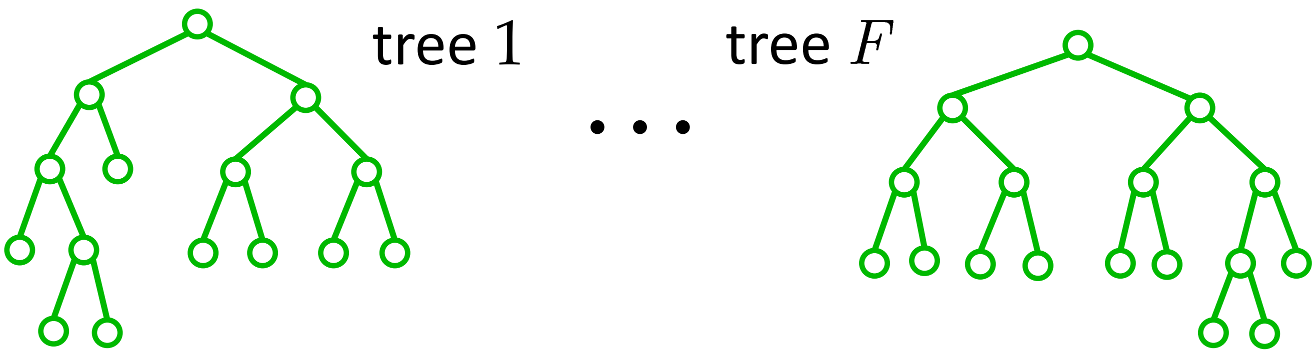
Spacetime energy pooling



Random forest classifier

[Breiman 01]

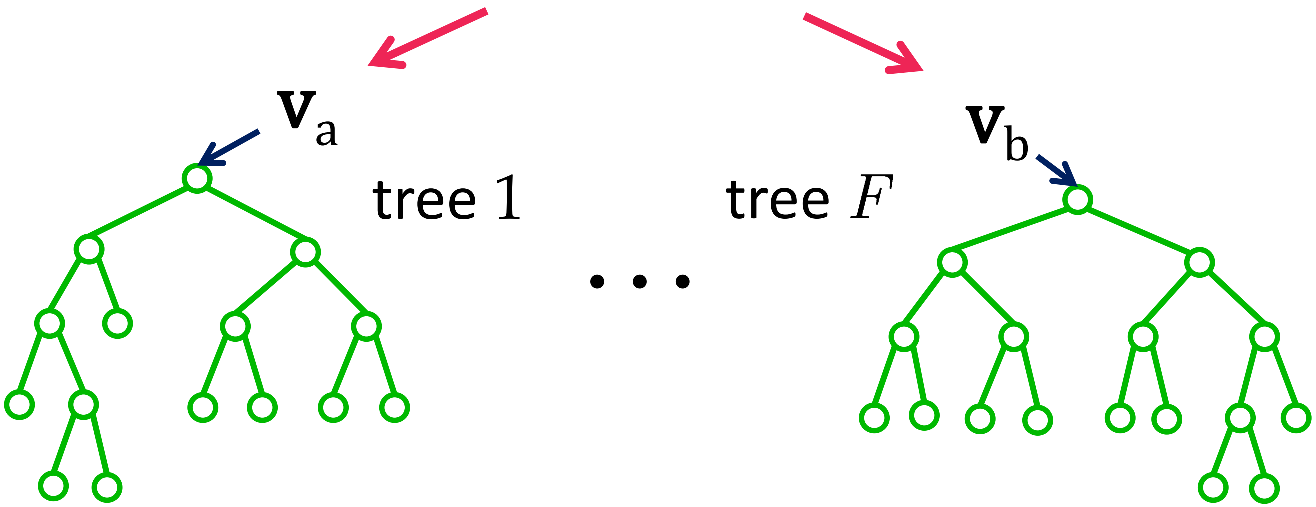
- Two sources of randomness:
  1. Subsample training data for each tree “bagging”



Random forest classifier

[Breiman 01]

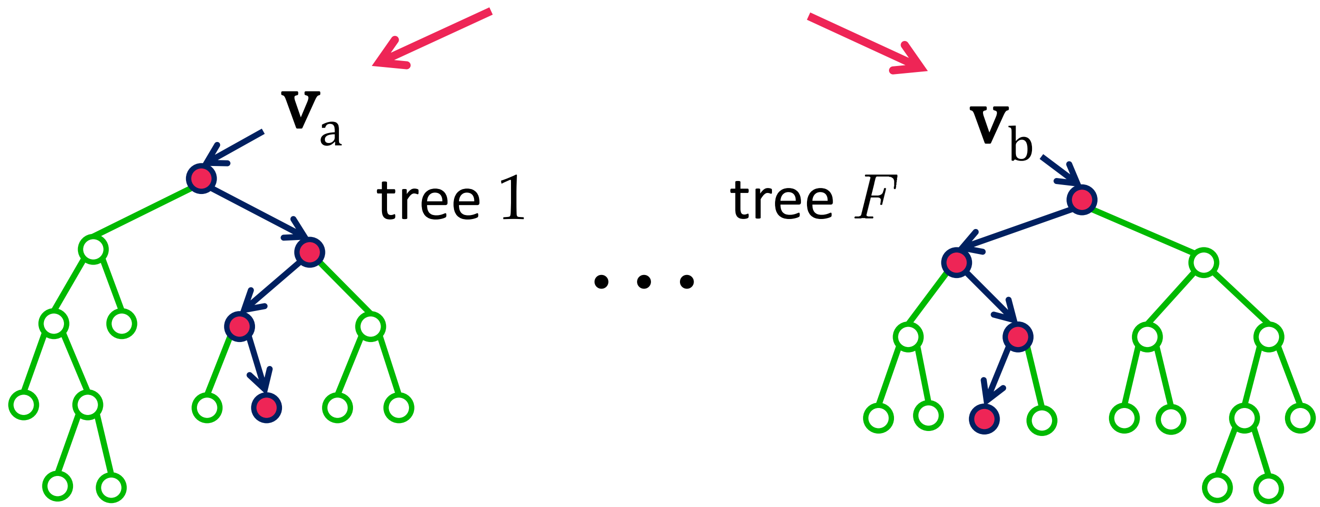
- Two sources of randomness:
  1. Subsample training data for each tree "bagging"



Random forest classifier

[Breiman 01]

- Two sources of randomness:
  1. Subsample training data for each tree "bagging"



  2. Random split selection

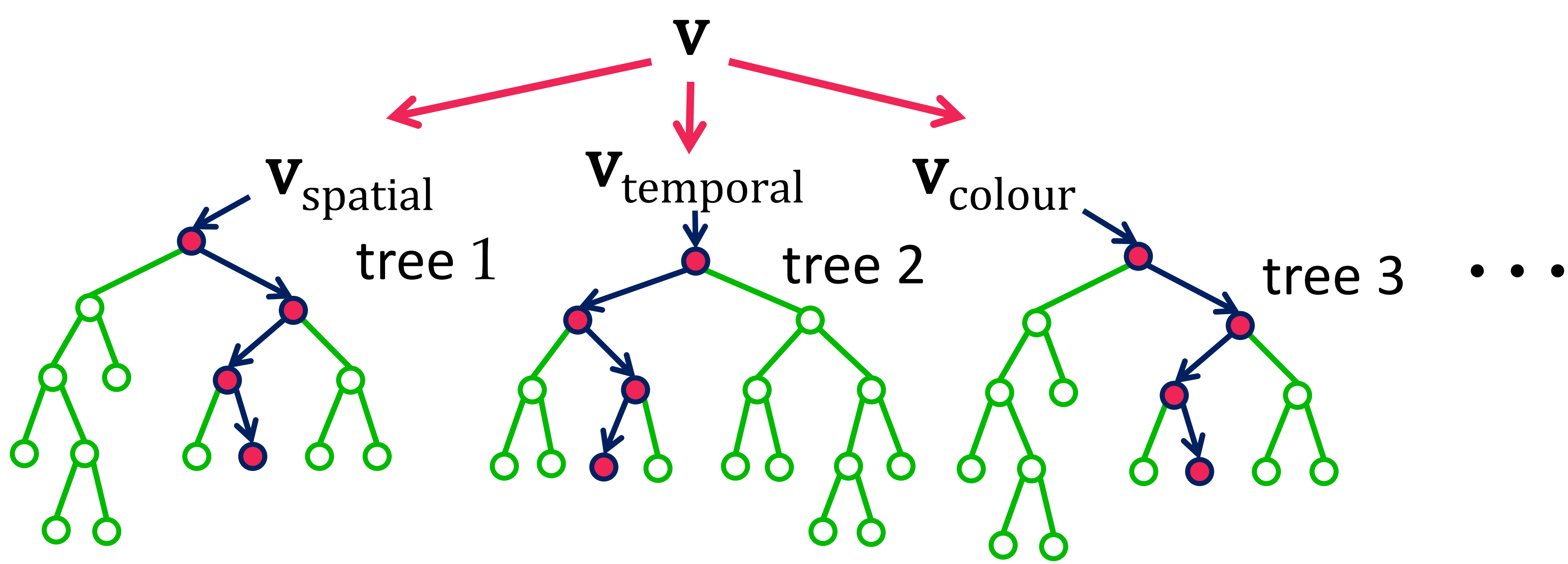
     Use a random number of features to determine best split based on maximum information gain I

$$I = H(Q) - \sum_{i \in \{L,R\}} \frac{|Q^i|}{|Q|} H(Q^i)$$

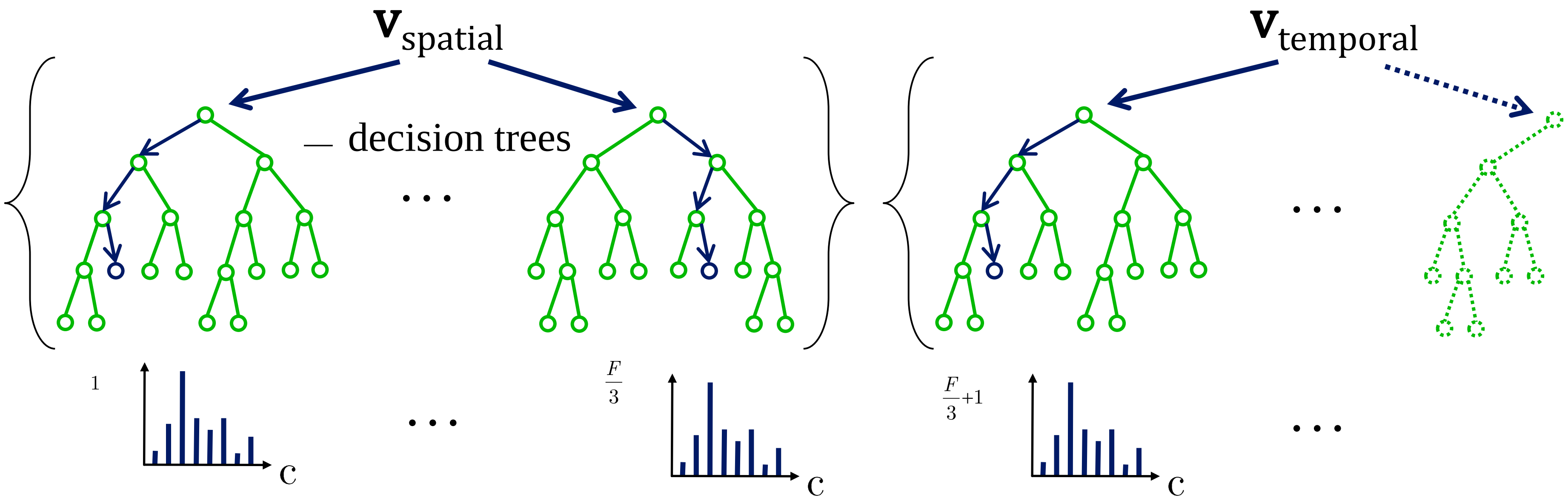
H Shannon entropy

Spacetime Random Forest (STRF)

- Restrict the node optimization process in each tree is to a single feature type
- Some classes are better represented by specific feature types



Classification with spacetime forest



- Average posterior probabilities for $\{\mathbf{v}_{\text{spatial}}, \mathbf{v}_{\text{temporal}}, \mathbf{v}_{\text{colour}}\}$

time τ

$$P^{\tau}(c|\mathbf{v}^{\tau}) = \frac{1}{F}\sum_{k=1}^{F} p_k(c|\mathbf{v}^{\tau}) \qquad \underbrace{c^{\tau}}_{\text{class label}} = \arg\max_{c} P^{\tau}(c|\mathbf{v}^{\tau})$$

Maryland “in the wild” dataset


avalanche


boiling water


chaotic traffic


forest fire


fountain


iceberg collapse

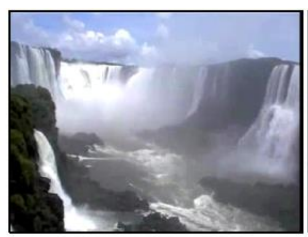

landslide

smooth traffic


tornado


volcanic eruption

- 13 scene categories
- 10 videos each


waterfall


waves


whirlpool

- Unconstrained camera motion

Results on Maryland "in the wild"

Descriptor	HOF+ GIST	Chaos+ GIST		SOE	
Classifier	NN	NN	SVM	NN	RF
Temporal τ	*all*	*all*	*all*	*all*	*all*
Avalanche	0.2	0.4	0.6	0.1	0.4
Bo. Water	0.5	0.4	0.6	0.5	0.5
Ch. Traffic	0.3	0.7	0.7	0.8	0.6
Forest Fire	0.5	0.4	0.6	0.4	0.1
Fountain	0.2	0.7	0.6	0.1	0.5
Iceberg Co.	0.2	0.5	0.5	0.1	0.4
Landslide	0.2	0.5	0.3	0.5	0.2
Sm. Traffic	0.3	0.5	0.5	0.7	0.3
Tornado	0.4	0.9	0.8	0.6	0.7
Volcanic Er.	0.2	0.5	0.7	0.3	0.1
Waterfall	0.2	0.1	0.4	0.2	0.6
Waves	0.8	0.9	0.8	0.8	0.5
Whirlpool	0.3	0.4	0.5	0.4	0.7
Avg. Perf.	**0.33**	**0.52**	**0.58**	**0.42**	**0.43**

[Shroff et al. CVPR'10] Combination of GIST and chaotic invariants

YUPENN dynamic scenes dataset

- 14 scene categories
- 30 videos in each category
- Stabilized camera

Results on YUPENN dynamic scenes

Descriptor	HOF+ GIST	Chaos+ GIST	SOE	
Classifier Temporal τ	NN *all*	NN *all*	NN *all*	RF *all*
Beach	0.87	0.30	0.90	0.93
Elevator	0.87	0.47	0.90	1.00
Forest Fire	0.63	0.17	0.87	0.67
Fountain	0.43	0.03	0.50	0.43
Highway	0.47	0.23	0.73	0.70
Lightning S.	0.63	0.37	0.90	0.77
Ocean	0.97	0.43	0.97	1.00
Railway	0.83	0.07	0.90	0.80
Rushing R.	0.77	0.10	0.90	0.93
Sky-Clouds	0.87	0.47	0.93	0.83
Snowing	0.47	0.10	0.50	0.87
Street	0.77	0.17	0.87	0.90
Waterfall	0.47	0.10	0.47	0.63
Windmill F.	0.53	0.17	0.73	0.83
Avg. Perf.	**0.68**	**0.23**	**0.79**	**0.81**

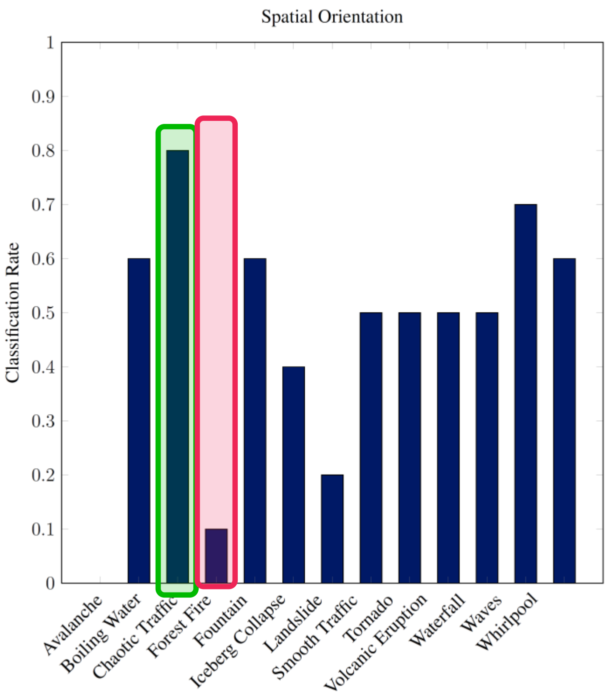
[Derpanis et al. CVPR'12] Spacetime Orientation Features

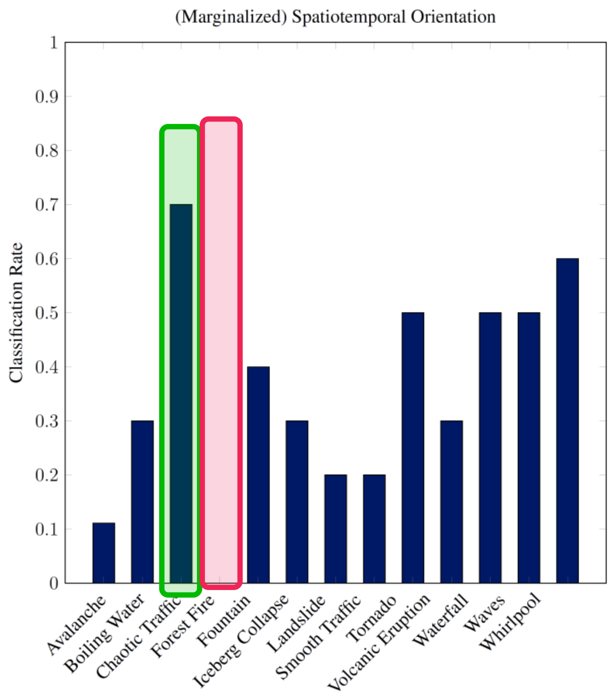
Complementarity of CSO descriptor

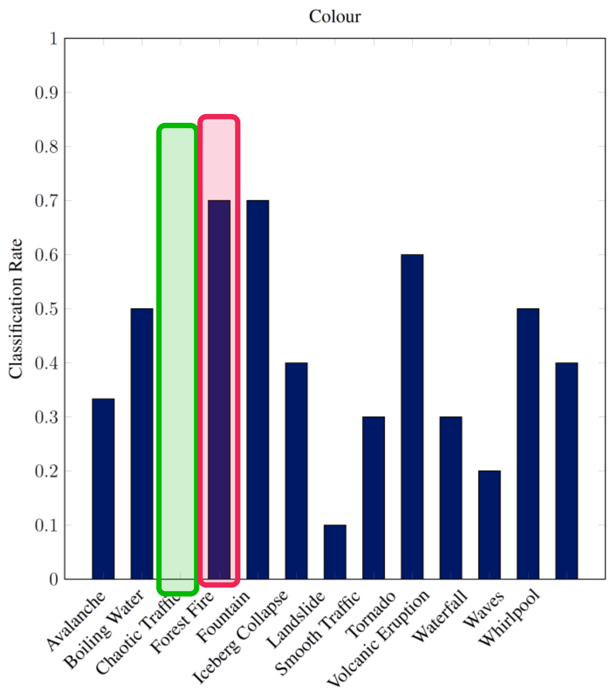
spatial

temporal

colour



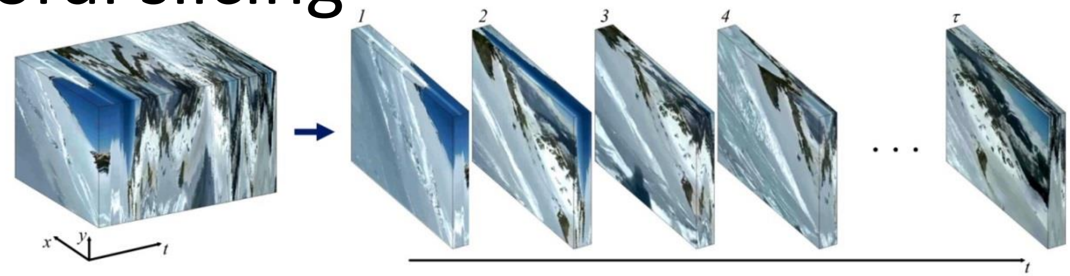




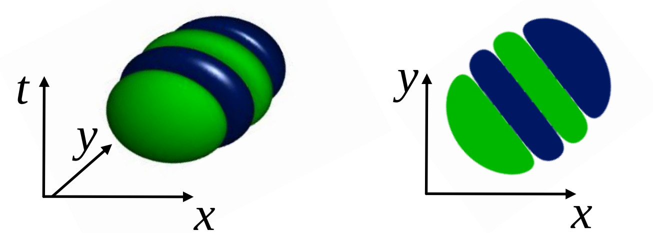
(a) Maryland “In-The-Wild”

In Summary

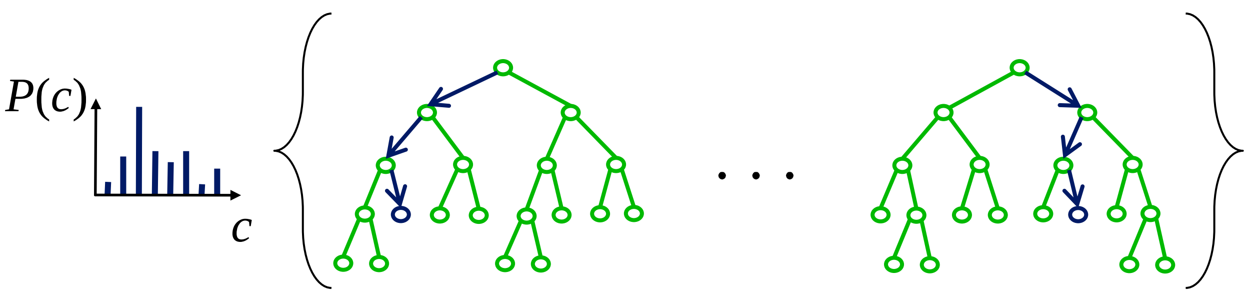
- Temporal slicing



- Complementary spacetime descriptor



- Spacetime random forest



- State of the art recognition rates with only a single slice