Dynamic Scene Understanding: The Role of Orientation Features in Space and Time in Scene Classification

published in CVPR 2012

Konstantinos G. Derpanis,
Matthieu Lecce,
Kostas Daniilidis,
Richard P. Wildes

Overview

- Natural scene classification is a fundamental challenge in the goal of automated scene understanding.

Overview

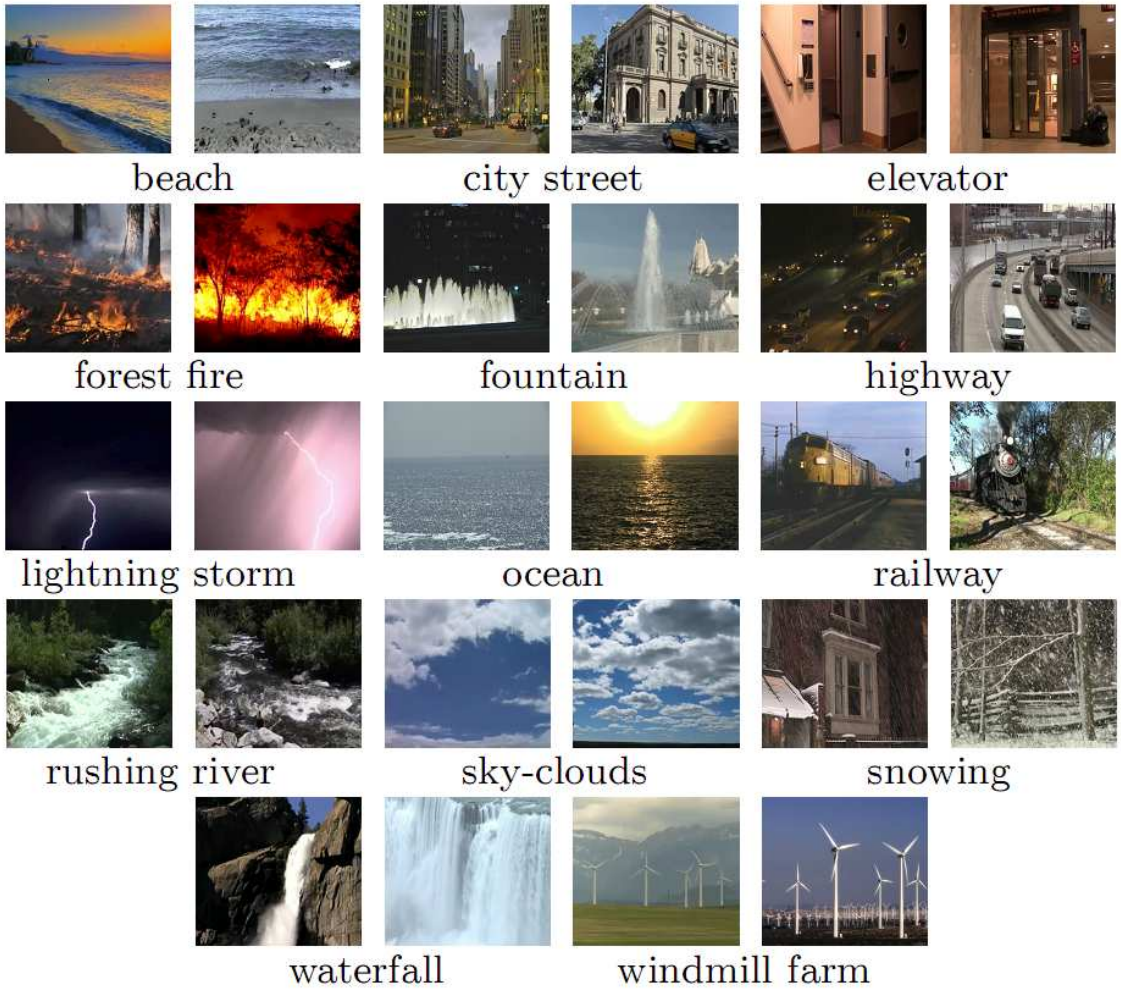
- The majority of studies have limited their scope to scenes from single image stills
  - Ignore potentially informative temporal cues

- This work is concerned with determining the degree of performance gain in considering short videos for recognizing natural scenes.

Overview

- The impact of multi-scale orientation measurements on scene classification is systematically investigated
  - Spatial appearances
  - Temporal dynamics
  - Joint spatial appearance and dynamics
- A new dataset (YUPENN Dynamic Scenes) is introduced
  - Fourteen scene categories
  - 420 image sequences
  - Temporal scene information due to objects and surfaces

Challenges


beach
city street
elevator
forest fire
fountain
highway
lightning storm
ocean
railway
rushing river
sky-clouds
snowing
waterfall
windmill farm

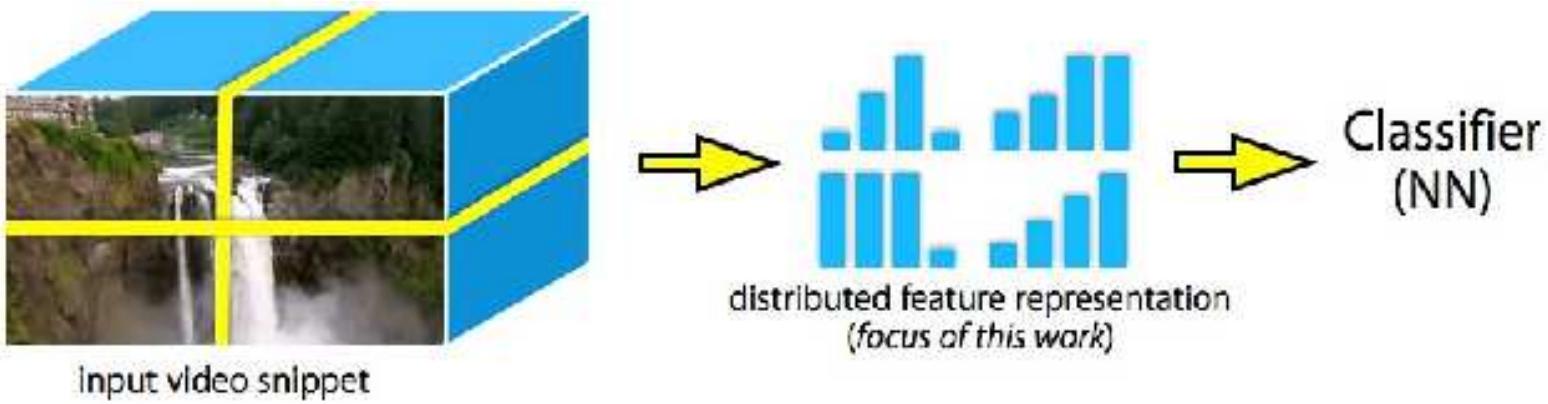
Challenges

- Image motion
  - Cars in “highway”,opening doors in “elevator”
- Nontextured regions
  - Sky in “sky-cloud”
- Flicker
  - Fire in “forest fire”, lightning in “lightning storm”
- Dynamic texture
  - Turbulance in “rushing river”, waves in “ocean”
- Classical Challenges
  - Variation in the appearance including scale, view, illumination, background.

Related Work

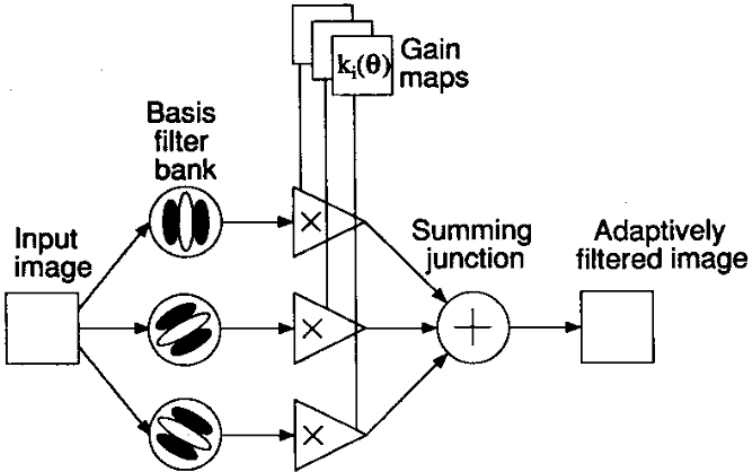
- Gradient-based features
  - GIST, SIFT, HOG …
- Color, texture, frequency (DFT)
  - Computed over entire image or fixed image subregions.
- Linear Dynamic Systems (LDS)
  - Used in recognizing dynamic textures.
  - Perform poorly on recognizing dynamic scenes.
- Histogram of Optical Flow (HOF)
  - multiple motions, temporal flicker, dynamic texture violates the underlying assumptions of flow computation.
- Co-occurrence of objects and actions and their pairwise relationships

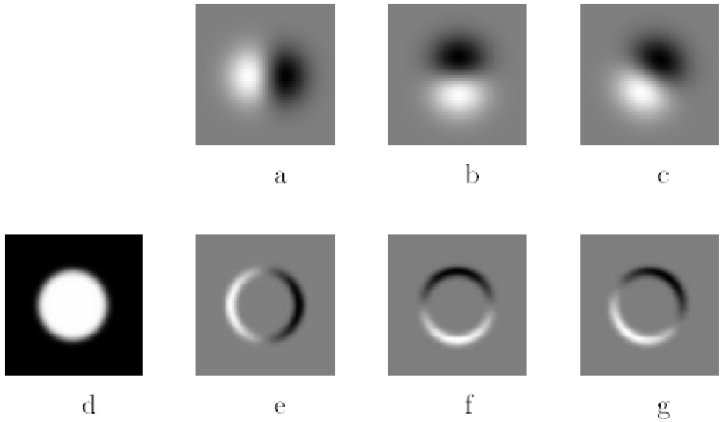
Approach



- The input image sequence is spatially subdivided.
- Relative position of subdivided regions captures scene layout.

Approach – Steerable Filters





- Use steerable filters to compute orientation measurements.
  1. Generate set of basis functions by convolving the input image with the set of basis functions that span the space of all rotations
  2. Then generate the required filtered version of the image by taking appropriate linear combinations of the outputs.

Slide Credit: W.T. Freeman and E.H. Adelson, "The design and use of steerable filters," PAMI, vol. 13, no. 9, pp. 891–906, Sept. 1991

Approach – 3D Spacetime Orientations

- For temporal sequences of images, orientation in space time corresponds to motion.
- Compute orientation energy by using 3D Gaussian third derivative filters.

$$E_{\hat{\theta},\sigma} = \sum_{\mathbf{x}} \Omega(\mathbf{x})[G_{N_{\hat{\theta},\sigma}}(\mathbf{x}) * I(\mathbf{x})]^2$$

$\mathbf{x} = (x,y,t)^T$	, denotes spatiotemporal image coordinates,
$I(\mathbf{x})$	, input image sequence,
$*$	, convolution,
$\Omega(\mathbf{x})$	, a mask defining the aggregation region,
$G_{N_{\hat{\theta},\sigma}}(\mathbf{x})$	, N^{th} derivative of the Gaussian with scale σ and θ the direction of the filter's axis of symmetry

- Ten spacetime orientations were selected as they correspond to the minimal spanning set for G_3.

K. Derpanis and J. Gryn. Three-dimensional nth derivative of Gaussian separable steerable filters. In ICIP, pages III: 553–556, 2005.

Approach – Normalized Energy

- Initial energy formula is confounded by local image contrast that appear as a multiplicative constant in the set of energies.

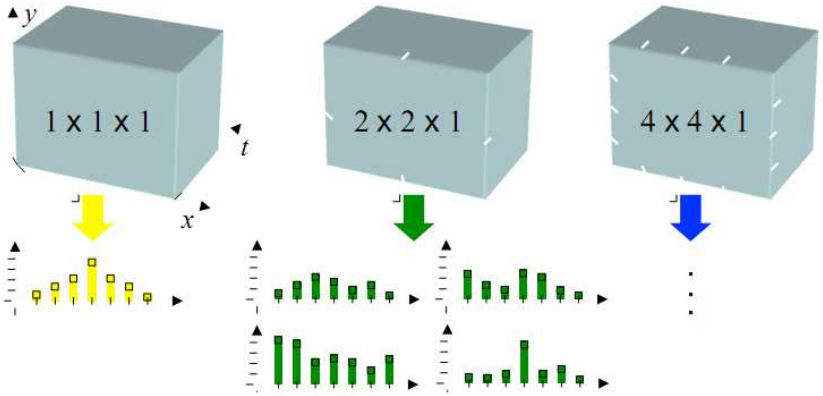
$$\hat{E}_{\hat{\theta}_i,\sigma_j} = E_{\hat{\theta}_i,\sigma_j} \Big/ \left(\epsilon + \sum_{\hat{\theta}\times\sigma\in\mathbb{S}} E_{\hat{\theta},\sigma} \right)$$

where S denotes the set of considered multiscale oriented energies, (1), and ϵ is a constant that serves as a noise floor.

$$\tilde{E}_{\hat{\mathbf{n}},\sigma} = \sum_{i=0}^{N} E_{\hat{\theta}_i,\sigma},$$

Approach – Spatiotemporal Scale

$$E_{\hat{\theta},\sigma} = \sum_{\mathbf{x}} \Omega(\mathbf{x})[G_{N_{\hat{\theta},\sigma}}(\mathbf{x}) * I(\mathbf{x})]^2$$



- The inner scale, corresponding to the Gaussian filter standard deviation, σ, in the energy computation, determines the range of spatiotemporal details captured by the representation.

- The outer scale, given by Ω(x, y, t), specifies the spatiotemporal scale of the support region for aggregating measurements.
  - Limiting the outer scale to the entire image sequence itself disregards the spatiotemporal layout of dynamic structure.

Classification

- Nearest Neighbour (NN) classifier was used in all evaluations.
- The set of normalized oriented energy measurements, within each outer scale form a histogram.
- Leave-one-video-out validation is used.
- Variety of (dis)similarity measures, e.g., Bhattacharyya, L1, L2 and $\chi 2$, that yielded little difference in classification performance.
  - The Bhattacharyya coefficient provided slightly better overall performance

$$s(u;v) = \sum_i \sum_j \sqrt{u_{i;j} v_{i;j}}$$

Evaluation

- Spatial appearance: color and GIST
- Temporal dynamics:
  - Histogram of flow (HOF)
  - Chaos
  - Appearance marginalized spatiotemporal oriented energies (MSOE)
- Joint spatial appearance and dynamics:
  - HOF fused with GIST
  - Chaos fused with GIST
  - Spatiotemporal oriented energies (SOE)
  - Colour fused with each of the spatiotemporal features
- For the combination of methods , weighted sum of the similarities are used as similarity measure. Weight is the average classification accuracy for that method.

Chaos – Theory[1]

- Chaos theory is concerned with understanding dynamical systems whose structure is unknown.
- Further, in such systems small changes in initial conditions result in huge variations in the output.
- The underlying process is not entirely random, there exists a deterministic component which can not be characterized in a closed form.
- Applied to gait modeling and action recognition.

[1] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas:Exploiting motion for describing scenes. In CVPR, 2010.

Datasets - YUPENN Dynamic Scenes

- Contains a wide variety of dynamic scenes captured from stationary cameras.
- 14 dynamic scene categories each containing 30 color videos
- Contains signifcant differences in image resolution, frame rate, scene appearance, scale, illumination conditions and camera viewpoint.

Datasets – Maryland "in the wild"

- Contains large camera motions and scene cuts
- 13 dynamic scene categories each containing 10 color videos
- Difficult to understand whether classification performance of approaches depends on their success in capturing underlying scene structure vs. characteristics induced by the camera.

Results

(a)	Spatial		Temporal			Spatiotemporal		
Scene classes	Color [13]	GIST [26]	HOF [25]	Chaos [30]	MSOE	Chaos+ GIST+ Color	HOF+ GIST	SOE
avalanche	50	10 (50)	0	30	10	40	30 (20)	10 (10)
b. water	30	60 (60)	40	30	50	40	50 (50)	60 (50)
c. traffic	20	70 (40)	20	50	90	70	40 (30)	80 (80)
f. fire	70	10 (60)	0	30	10	40	30 (50)	40 (40)
fountain	50	30 (20)	10	20	10	70	20 (20)	10 (10)
i. collapse	0	10 (20)	10	10	10	50	10 (20)	20 (10)
landslide	10	20 (30)	20	10	30	50	20 (20)	50 (50)
s. traffic	50	40 (30)	30	20	70	50	30 (30)	60 (70)
tornado	60	40 (60)	0	60	80	90	40 (40)	60 (60)
v. eruption	30	30 (40)	0	70	10	50	30 (20)	10 (30)
waterfall	20	50 (30)	20	30	30	10	20 (20)	10 (20)
waves	40	80 (80)	40	80	80	90	80 (80)	80 (80)
whirlpool	10	40 (30)	30	30	30	40	20 (30)	40 (40)
Avg. (%)	34	38 (43)	17	36	39	52	32 (33)	41 (42)

Table 1. (a) Comparison of classification rates among the various spatial, temporal and spatiotemporal image representations on the Maryland “in-the-wild”. Parentheses denote classification rates where Color is additionally considered.

Results

(b)	Spatial		Temporal			Spatiotemporal		
Scene classes	Color [13]	GIST [26]	HOF [25]	Chaos [30]	MSOE	Chaos+ GIST	HOF+ GIST	SOE
beach	50	90 (90)	37	27	83	30 (30)	76 (87)	87 (90)
c. street	47	50 (63)	83	17	63	17 (17)	80 (77)	83 (87)
elevator	83	53 (80)	93	40	60	40 (47)	90 (87)	67 (90)
f. fire	47	50 (57)	67	50	60	17 (17)	63 (63)	83 (87)
fountain	13	40 (50)	30	7	40	3 (3)	37 (43)	47 (50)
highway	30	47 (53)	33	17	60	23 (23)	53 (47)	77 (73)
l. storm	83	57 (70)	47	37	87	40 (37)	70 (63)	90 (90)
ocean	73	93 (97)	60	43	97	43 (43)	93 (97)	100 (97)
railway	43	50 (53)	83	3	60	7 (7)	87 (83)	87 (90)
r. river	57	63 (80)	37	3	90	10 (10)	73 (77)	93 (90)
sky	30	90 (93)	83	33	80	43 (47)	87 (87)	90 (93)
snowing	53	20 (20)	57	10	17	10 (10)	40 (47)	33 (50)
waterfall	30	33 (40)	60	10	37	10 (10)	50 (47)	43 (47)
w. farm	57	47 (60)	53	17	47	17 (17)	60 (53)	57 (73)
Avg. (%)	50	56 (65)	59	20	63	22 (23)	69 (68)	74 (79)

Table 1. (b) Comparison of classification rates among the various spatial, temporal and spatiotemporal image representations on the “stabilized”. Parentheses denote classification rates where Color is additionally considered.

Results

(a)

Actual \ Classified	avalanche	b. water	c. traffic	f. fire	fountain	i. collapse	landslide	s. traffic	tornado	v. eruption	waterfall	waves	whirlpool
avalanche	1			1		2	1			1	**3**		1
b. water		**6**	1		1		2						
c. traffic			**8**					2					
f. fire	1			**4**			1		2			1	1
fountain		2	1		1		**3**	1			2		
i. collapse			**4**			2	1		1	1			1
landslide				1	1	2	**5**		1				
s. traffic			1				1	**6**			1	1	
tornado							1	1	**6**	2			
v. eruption	1		**3**	1	2					1		1	1
waterfall	**2**			**2**	1	1	**2**		1		1		
waves				1			1					**8**	
whirlpool			2	1		1	1			1			**4**

(b)

Actual \ Classified	Sky	Beach	Ocean	Street	Railway	R. River	Highway	Snowing	Waterfall	Fountain	L. Storm	F. Fire	W. Farm	Elevator
Sky	**27**		1		1	1								
Beach		**26**	3			1								
Ocean			**30**											
Street				**25**		3			1			1		
Railway		1			**26**	2						1		
R. River		1				**28**	1							
Highway		1		1		3	**23**		2					
Snowing				1		6		**10**	4	2		3	1	3
Waterfall						2		2	**13**	7		3	2	1
Fountain								3	9	**14**		4		
L. Storm											**27**	3		
F. Fire								2	1		2	**25**		
W. Farm								3	3	1	1	5	**17**	
Elevator								2	3	4		1		**20**

Table 2. Confusion matrix for SOE (4 × 4 × 1) on the “in the wild” and “stabilized” data sets in (a) and (b), resp.Bold shows top classification for each actual set.

- Table 2. Confusion matrix for SOE (4 × 4 × 1) on the “in the wild” and “stabilized” data sets in (a) and (b), resp.Bold shows top classification for each actual set.

Results

			Inner scale			
			0	1	2	all
Outer scale	$1 \times 1 \times 1$	MSOE [8]	52	53	51	55
		SOE	56	54	56	63
	$2 \times 2 \times 1$	MSOE	55	58	58	61
		SOE	66	67	66	69
	$4 \times 4 \times 1$	MSOE	52	57	60	63
		SOE	64	69	69	74
	all	MSOE	53	60	62	63
		SOE	65	70	70	75

Table 3. Impact of inner and outer scales on overall classification on the "stabilized" scenes data set.

- The "all" row for outer scale is constructed from a weighted sum of the similarities of the individual outer scale levels with the individual weights proportional to the corresponding level's average accuracy.
- The "all" column for inner scale is constructed as a natural consequence of combining the individual inner scales via the normalization process.

Conclusion

- In temporal approaches
  - Chaos are more difficult to explain, it has better results when the camera is moving.
  - The poor performance of HOF on the “in-the-wild” data can be explained by the erratic camera motions and scene cuts that are difficult to capture.
  - MSOE performs best across both data sets
- In stpatiotemporal approaches
  - SOE is the best performer on the stabilized data and the second best on in-the-wild data.
  - SOE is consistently able to characterize dynamic scenes

Questions

?