Now You *Shake* Me: Towards Automatic 4D Cinema

Yuhao Zhou[1], Makarand Tapaswi[1,2], Sanja Fidler[1,2]

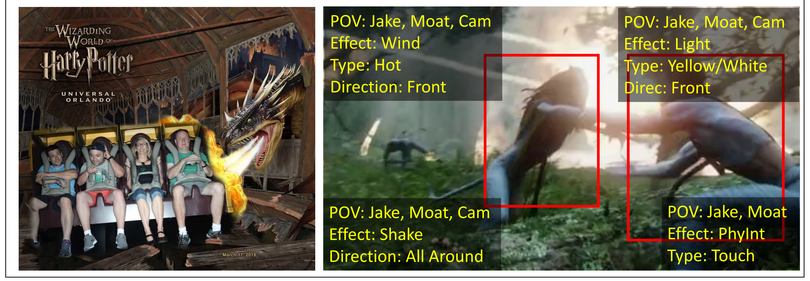
[1]University of Toronto, [2]Vector Institute,

{henryzhou, makarand, fidler}@cs.toronto.edu

http://www.cs.toronto.edu/~henryzhou/movie4d/

Introduction

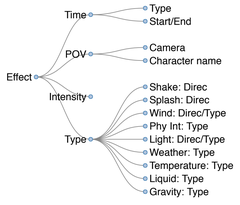
Goal: Automate 4D effect detection in movies



Movie4D Dataset

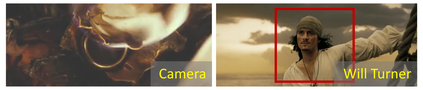
- Effect classes

Each annotation has



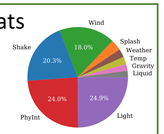
- Each effect has details such as
  - direction, intensity
  - type (e.g., hot/cold for temperature)

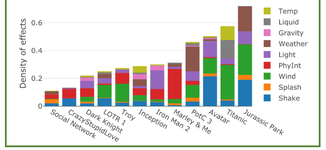
- Point of View: camera and main characters



Stats

Movies	63
# Effects	9286
Duration	6.1





Model

Effect Classification

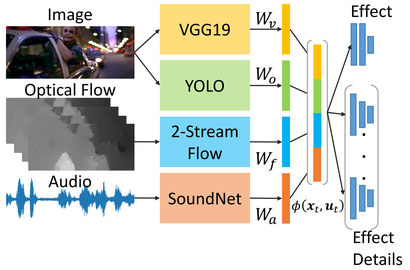
Given a short clip (max. 3s), predict the corresponding effect and all its details.

Audio/visual representation

$$\phi(\mathbf{x}_t, \mathbf{u}_t) = [W_v \mathbf{v}_t, W_f \mathbf{f}_t, W_a \mathbf{a}_t, W_o \mathbf{o}_t]$$

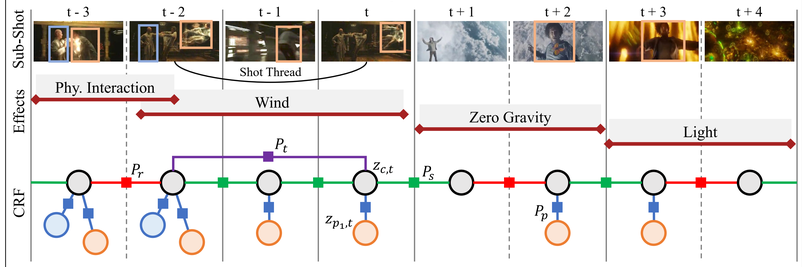
Effect labels and details prediction

$$\mathbf{y}_t = W_2 \cdot \mathrm{ReLU}(W_1 \cdot \phi(\mathbf{x}_t, \mathbf{u}_t))$$



Effect Detection

Find all effects in a 5-minute video clip.



- Effect predictions for camera and person tracks: Unaries
- Model dependency between sub-shots, shots, threads, and camera-track

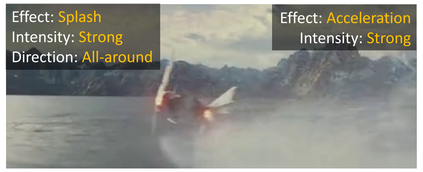
$$E(z, \mathbf{w}) = \sum_t \Big(w_{u_c} \overbrace{U(z_{c,t})}^{\text{camera}} + \sum_l w_{u_p} \overbrace{U(z_{p_l,t})}^{\text{track}} \Big) + w_q \overbrace{\sum_{(t,t') \in \mathcal{Q}} P_q(z_{c,t}, z_{c,t'})}^{\text{video pairwise}} + w_p \overbrace{\sum_{(t,l)} P_p(z_{c,t}, z_{p_l,t})}^{\text{camera-track}}$$

- Combine sub-shot predictions using watershed (SSN, Zhao, et al. ICCV 2017)
- Training: using 0-1 loss and primal-dual approach (PD, Hazan, et al. NIPS 2010)
- Inference: distributed convex belief propagation (Schwing, et al. CVPR 2011)

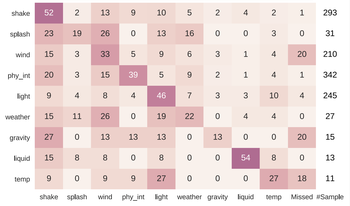
Results

Effect Classification

- Ablation study on modalities, visual strong
- Prediction with effect-specific details



Model	E	D-PR
Random	11.1	2.6
Ours	**43.7**	**15.8**
All – visual	26.9	10.2
All – optflow	32.1	10.9
All – audio	40.7	15.2
All – objdet	36.7	15.5



Effect Detection

- Prediction at each sub-shot

- Ablation on CRF pairwise terms

POV	Model	Exist AP	Effect Acc
Camera	Unaries	55.4	43.6
	CRF: U + shot	51.4	50.4
	CRF: U + thread	53.8	44.6
	CRF: U + all	**57.3**	**52.4**
Camera + Tracks	Unaries	29.4	45.3
	CRF: U + video	27.9	48.0
	CRF: U + all	**28.8**	**48.8**

- Final detection result

POV	Model	PRC	REC	F1
Camera	Unaries	15.0	35.1	21.1
	CRF	25.2	35.6	29.5
Camera + Tracks	Unaries	14.2	37.2	20.5
	CRF	16.7	28.7	21.1

Dataset Quality

- Annotators agreement F1 score 62.6%
- AMT effect classification experiment: when 4/5 turkers agree, acc = 88.4%

Thanks to NSERC, NVIDIA for GPUs, Relu for compute support, and Upwork annotators.