

CS6501/4501: Vision and Language

Video and Language

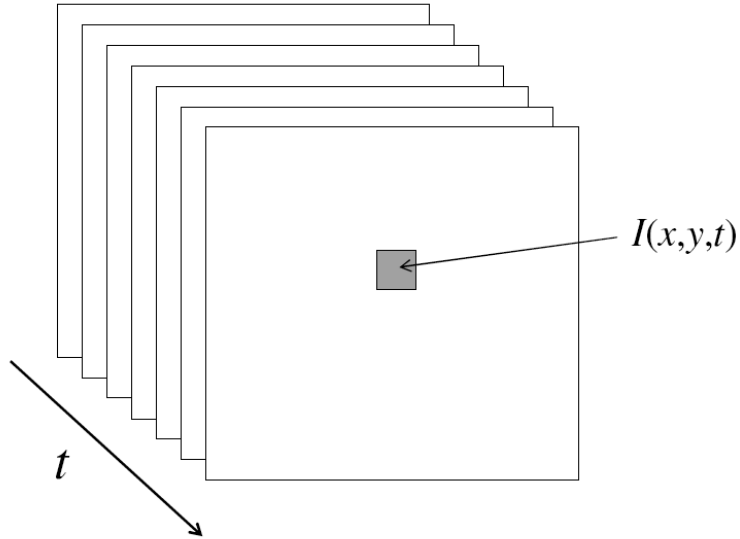


Today's Class

- Video:
 - Optical flow
 - Two-Stream Networks
 - CNN + LSTM
 - CNN + Temporal Pooling
 - 3D CNNs

From images to videos

- A video is a sequence of frames captured over time
- Now our image data is a function of space (x, y) and time (t)



Why is motion useful?

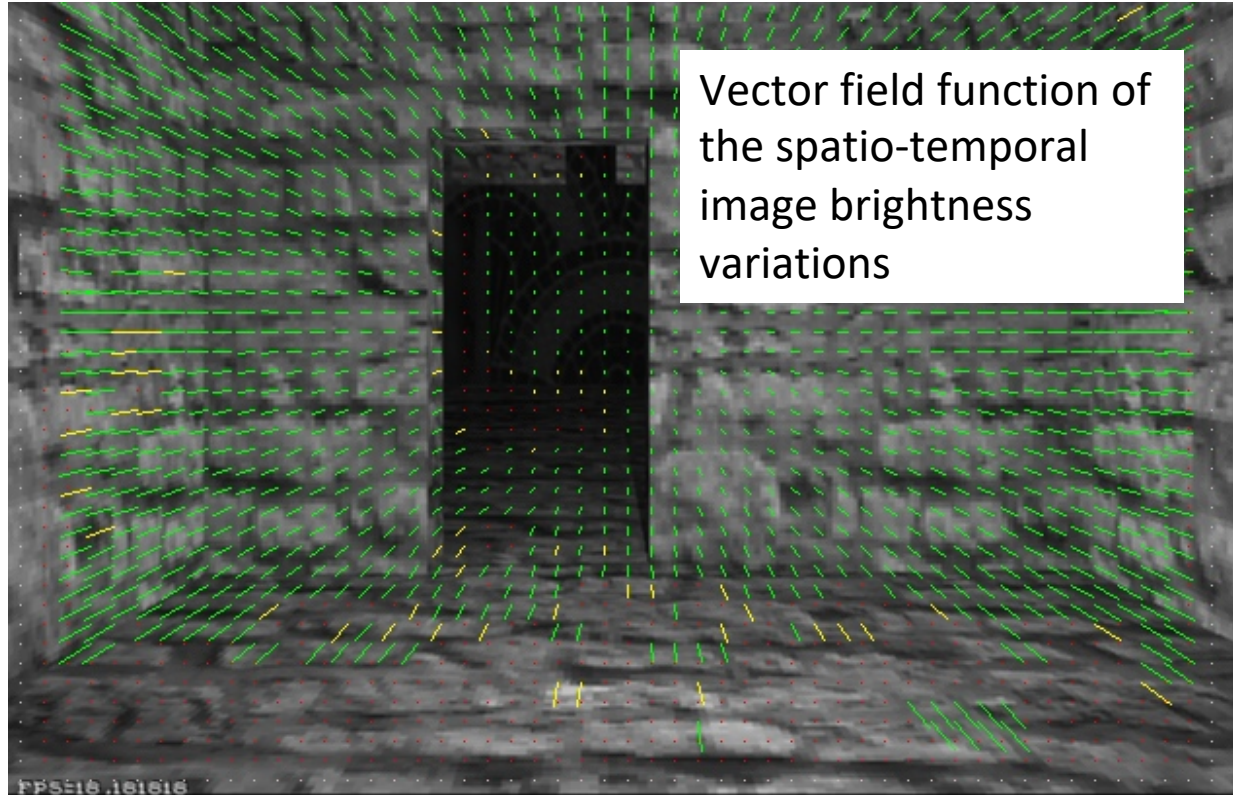


Optical flow

- Definition: optical flow is the *apparent* motion of brightness patterns in the image
- Note: apparent motion can be caused by lighting changes without any actual motion
 - Think of a uniform rotating sphere under fixed lighting vs. a stationary sphere under moving illumination

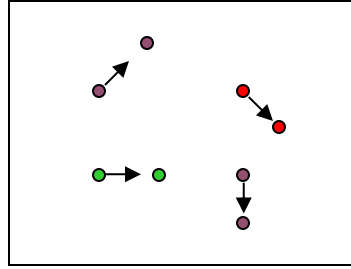
GOAL: Recover image motion at each pixel from optical flow

Optical flow

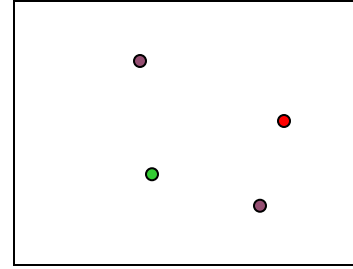


Picture courtesy of Selim Temizer - Learning and Intelligent Systems (LIS) Group, MIT

Estimating optical flow



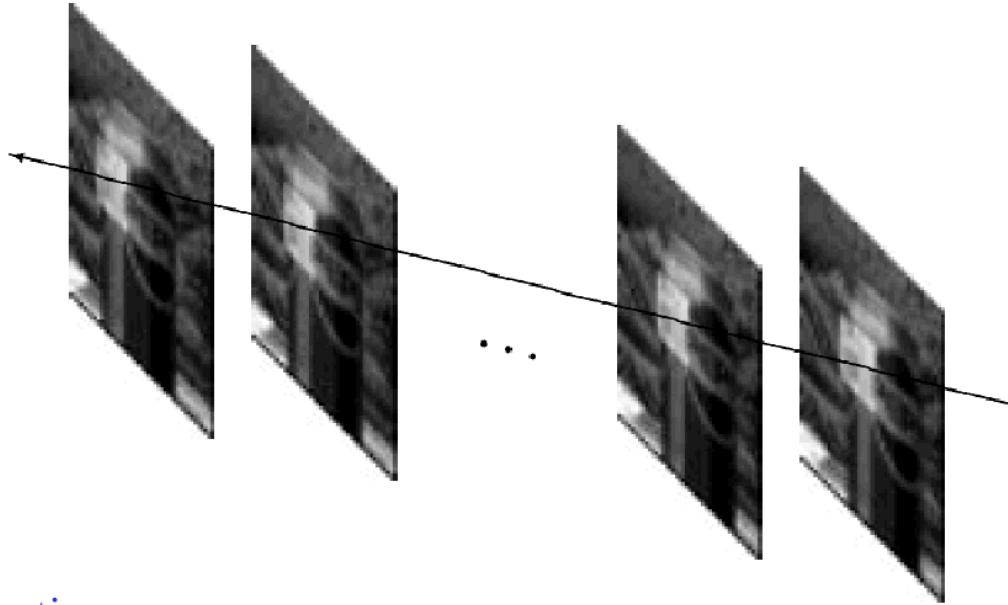
$I(x,y,t-1)$



$I(x,y,t)$

- Given two subsequent frames, estimate the apparent motion field $u(x,y)$, $v(x,y)$ between them
- Key assumptions
 - **Brightness constancy:** projection of the same point looks the same in every frame
 - **Small motion:** points do not move very far
 - **Spatial coherence:** points move like their neighbors

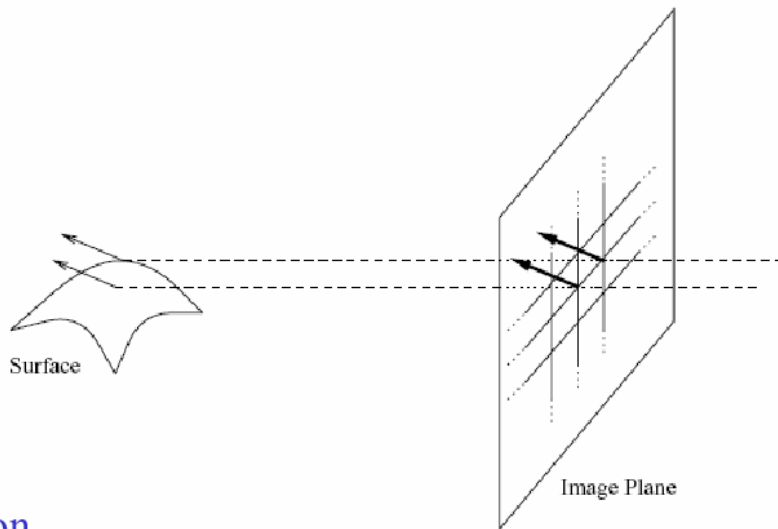
Key Assumptions: small motions



Assumption:

The image motion of a surface patch changes gradually over time.

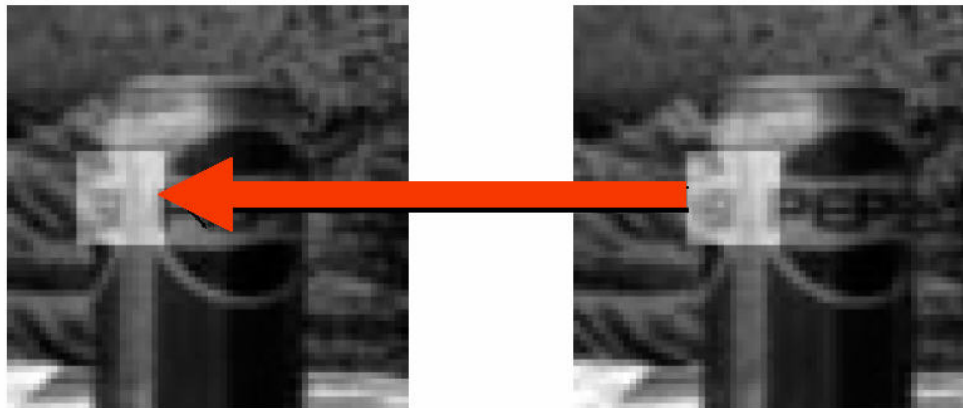
Key Assumptions: spatial coherence



Assumption

- * Neighboring points in the scene typically belong to the same surface and hence typically have similar motions.
- * Since they also project to nearby points in the image, we expect spatial coherence in image flow.

Key Assumptions: brightness Constancy



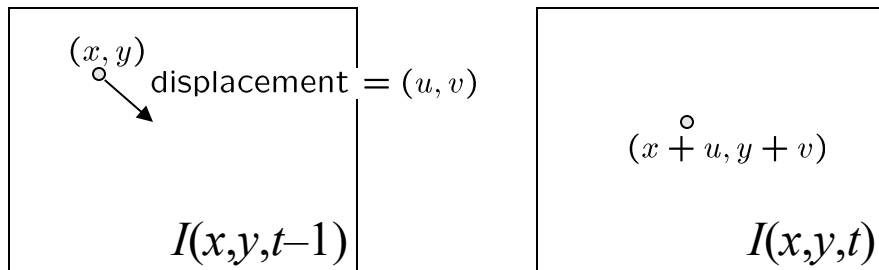
Assumption

Image measurements (e.g. brightness) in a small region remain the same although their location may change.

$$I(x + u, y + v, t + 1) = I(x, y, t)$$

(assumption)

The brightness constancy constraint



- Brightness Constancy Equation:

$$I(x, y, t - 1) = I(x + u(x, y), y + v(x, y), t)$$

Linearizing the right side using Taylor expansion:

$$I(x + u, y + v, t) \approx I(x, y, t - 1) + \overset{\text{Image derivative along x}}{I_x} \cdot u(x, y) + I_y \cdot v(x, y) + I_t$$

$$I(x + u, y + v, t) - I(x, y, t - 1) = I_x \cdot u(x, y) + I_y \cdot v(x, y) + I_t$$

$$\text{Hence, } I_x \cdot u + I_y \cdot v + I_t \approx 0 \quad \rightarrow \quad \nabla I \cdot [u \ v]^T + I_t = 0$$

The brightness constancy constraint

(x, y) displacement = (u, v)

B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

$$I(x+u, y+v, t) - I(x, y, t-1) = I_x \cdot u(x, y) + I_y \cdot v(x, y) + I_t$$

$$\text{Hence, } I_x \cdot u + I_y \cdot v + I_t \approx 0 \quad \rightarrow \quad \nabla I \cdot [u \ v]^T + I_t = 0$$

Action Classification from Video

Recommended Paper to Read:

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset

João Carreira[†]

joaoluis@google.com

Andrew Zisserman^{†,*}

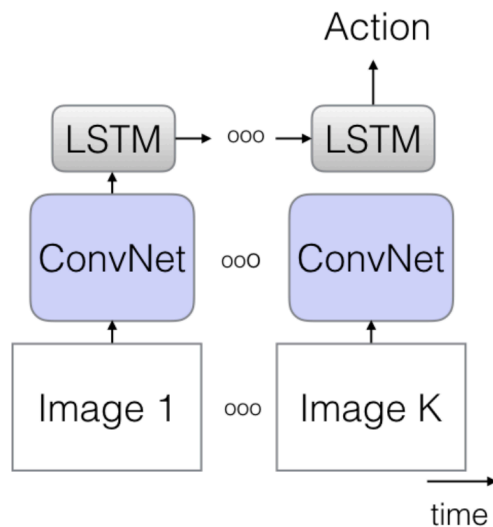
zisserman@google.com

[†]DeepMind

^{*}Department of Engineering Science, University of Oxford

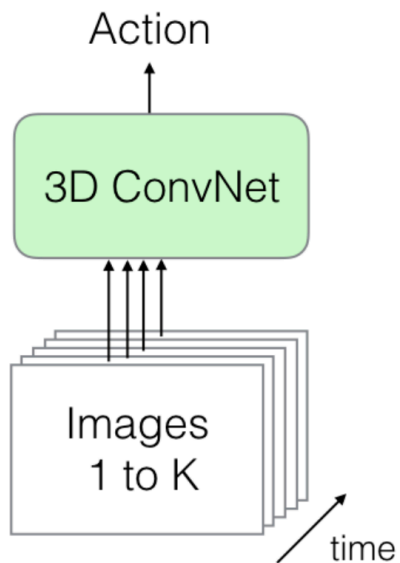
Action Classification from Video

CNN + LSTM over sequence of frames



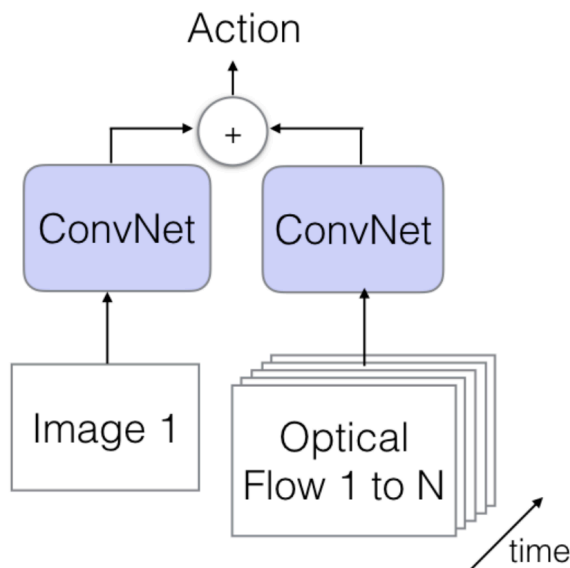
Action Classification from Video

3D CNN of consecutive frames across time



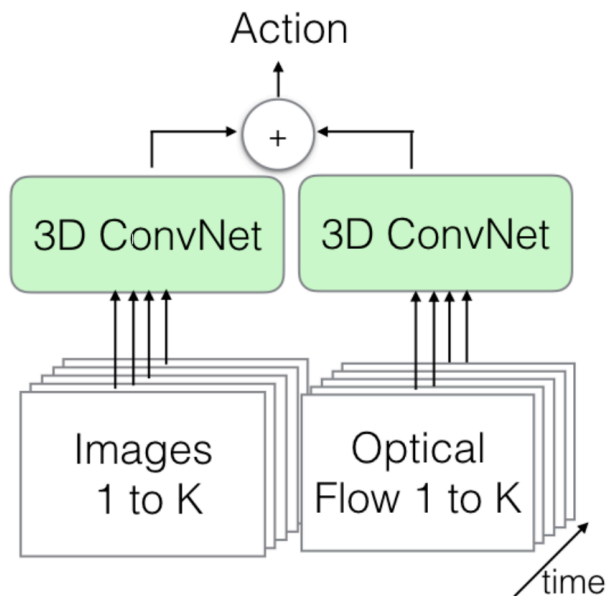
Action Classification from Video

Two Stream CNN: Images + Flow Map



Action Classification from Video

Two Stream 3D CNN: Images + Flow Map



UCF-101 Action Dataset



<https://www.crcv.ucf.edu/data/UCF101.php>

Action Classification from Video

Results on UCF101 actions

Architecture	UCF-101		
	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–
(b) 3D-ConvNet	51.6	–	–
(c) Two-Stream	83.6	85.6	91.2
(d) 3D-Fused	83.2	85.8	89.3
(e) Two-Stream I3D	84.5	90.6	93.4

Movie Trailers

Moviescope: Large-scale Analysis of Movies using Multiple Modalities

Paola Cascante-Bonilla^{1*} Kalpathy Sitaraman^{2†*} Mengjia Luo¹ Vicente Ordonez¹

¹University of Virginia, ²Microsoft

[pc9za, ml6uk, vicente]@virginia.edu, kasivara@microsoft.com



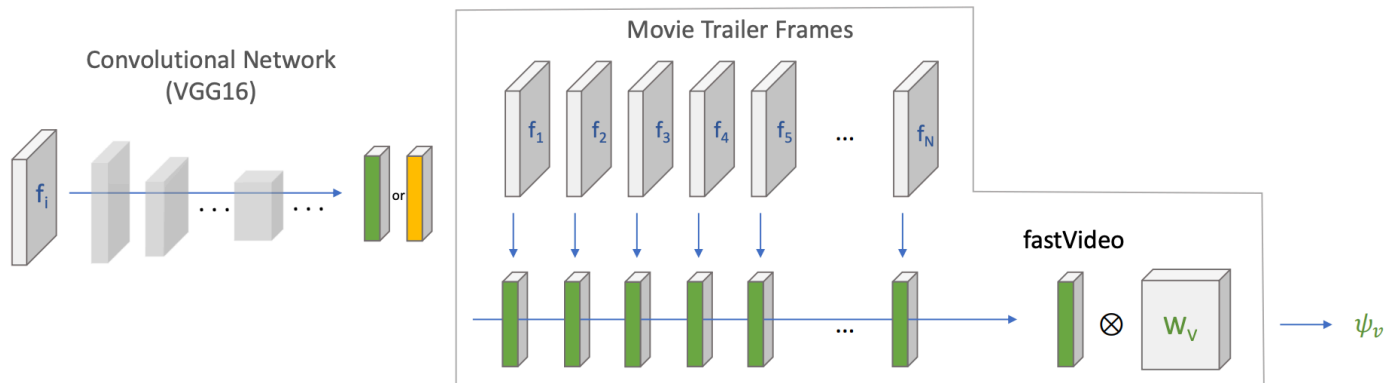
<https://arxiv.org/abs/1908.03180>

Movie Trailers

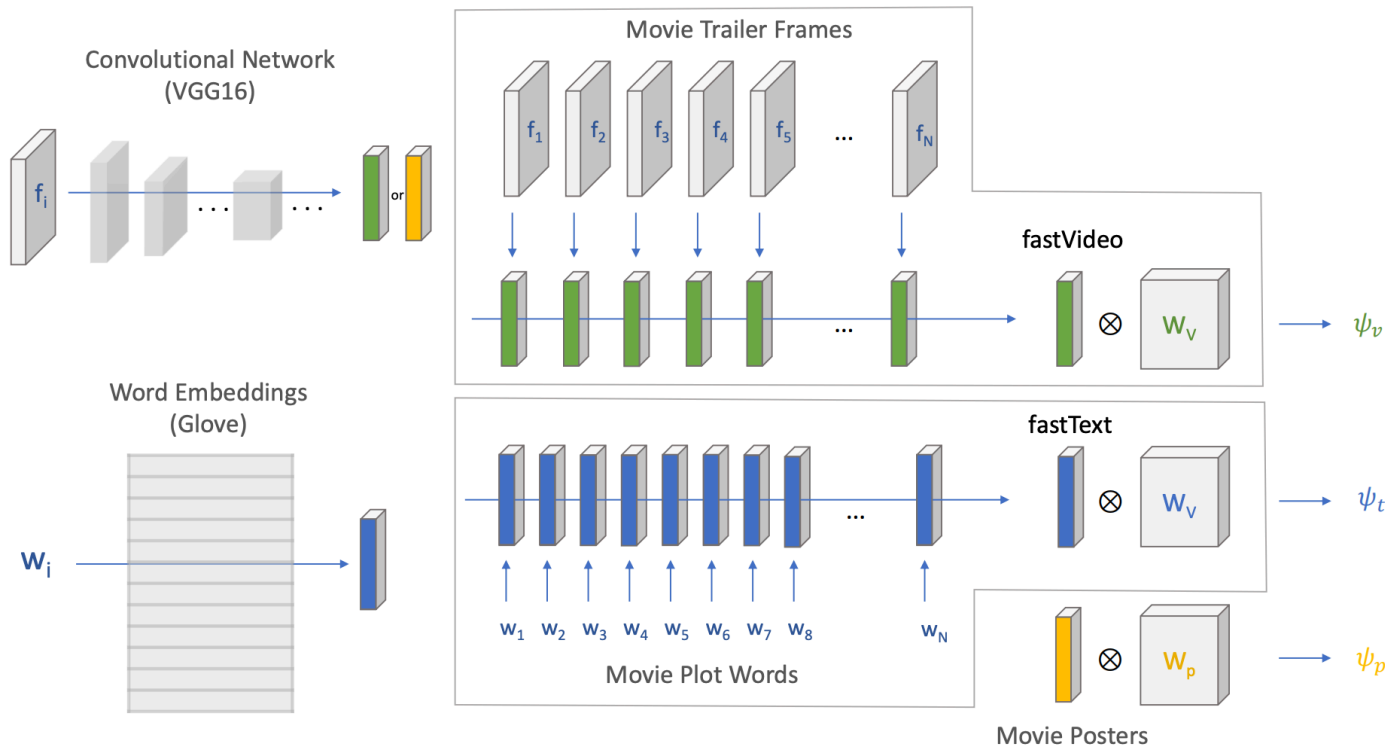


- Movie Trailers
- Movie Plots
- Movie Posters
- Movie Metadata

CNN + Temporal Pooling



CNN + Temporal Pooling



Results

Table 3. Mean Average Precision (mAP) Scores for movie genre prediction.

	action	anim	bio	com	crime	drama	fam	fant	horr	myst	rom	scifi	thrlr	<i>mAP</i>	<i>μAP</i>	<i>sAP</i>
% of training samples	8.70	1.84	2.22	14.17	10.56	19.63	4.14	6.97	4.29	3.79	8.36	4.66	10.69	-	-	-
Baseline accuracy	22.1	4.3	6.2	39.3	18.6	53.6	10.8	17.0	10.5	10.9	22.1	13.5	25.8	19.6	13.7	21.0
Video (V)																
C3D [37]	63.8	91.3	16.2	82.3	45.1	71.6	65.3	54.8	50.8	28.2	38.3	21.8	64.8	53.4	57.9	68.8
I3D [5]	37.2	51.8	9.2	72.6	33.9	67.6	43.6	39.0	22.8	21.3	34.3	22.6	48.3	38.8	50.5	65.6
LSTM	47.5	86.8	12.0	79.2	33.0	72.0	64.5	54.4	22.7	24.7	40.4	36.5	54.8	48.4	59.6	70.5
Bidirectional LSTM	49.9	86.3	8.2	77.6	29.9	70.8	65.4	55.3	22.3	21.7	41.6	35.9	51.2	47.4	58.2	69.9
fastVideo	61.4	94.8	23.9	81.5	41.7	77.0	67.0	62.6	36.1	30.4	48.4	48.2	62.0	56.5	64.9	75.6
fastVideo + TempConv	64.7	95.7	21.2	83.5	49.1	78.9	68.6	68.9	42.7	29.2	46.8	51.0	64.8	58.9	65.9	76.3
Audio (A)																
CRNN	56.7	48.0	11.2	86.2	40.0	79.0	49.6	44.7	37.6	22.7	43.0	27.0	56.3	46.3	61.4	72.3
Poster (P)																
VGG16	48.6	60.0	12.1	73.4	33.4	69.8	47.2	41.3	37.0	22.3	38.1	33.9	46.3	43.3	51.9	66.5
Text (T)																
Conv1D	62.5	34.4	24.7	64.8	54.3	73.8	50.3	64.6	50.4	31.5	43.2	70.6	61.5	52.8	57.8	70.4
LSTM	64.8	44.5	25.6	70.1	63.4	78.0	63.3	70.8	63.2	32.6	47.1	75.2	66.5	58.9	63.8	73.8
Bidirectional LSTM	63.7	42.5	31.2	69.3	58.1	76.7	57.9	66.4	61.3	30.7	52.3	76.2	63.2	57.7	63.2	73.5
fastText	72.0	50.7	40.6	81.1	68.7	82.3	69.2	68.8	78.3	47.8	60.3	74.4	72.9	66.7	72.5	81.4
fastText w/ Glove [20]	72.2	51.6	45.2	81.2	69.1	82.3	70.8	68.9	78.8	49.7	61.1	75.2	73.3	67.7	72.8	81.7
Metadata (M)																
XGBoost	61.5	76.8	35.4	74.8	36.7	82.7	83.7	53.7	62.3	22.8	31.4	33.4	50.9	54.3	62.9	73.7
RandomForest	59.3	73.7	33.3	74.9	40.6	82.7	83.2	58.8	62.7	25.4	35.4	37.9	55.0	55.6	63.9	73.7
Score Fusion																
Video-Audio (VA)	69.0	90.8	26.1	88.6	49.0	82.6	74.8	63.8	49.0	34.4	49.8	51.1	70.8	61.5	70.3	78.8
Vid-Aud-Poster (VAP)	68.8	92.5	27.4	88.5	48.9	82.6	74.8	63.7	49.5	34.3	50.1	50.3	70.7	61.7	70.4	78.8
Vid-Aud-Post-Text (VAPT)	73.3	95.2	29.9	91.0	61.2	85.0	77.2	69.0	68.9	38.8	51.8	61.6	74.1	67.5	74.9	82.3
Vid-Aud-Post-Text-Metad (VAPTM)	75.5	88.8	36.6	91.5	60.6	86.8	87.0	70.5	74.6	39.7	49.7	59.4	71.3	68.6	75.3	82.5

Results

Table 4. Mean Average Precision Scores on UCF101.

	mAP
‘Slow Fusion’ spatio-temporal ConvNet [16]	65.4
LSTM composite model (only RGB) [34]	75.8
C3D (fc6) [37]	76.4
iDT+C3D (fc6) [37]	86.7
Two-stream model [28]	88.0
Two-Stream I3D [5]	98.0
fastVideo - 16 Frames	79.2
fastVideo - 200 Frames	79.4
fastVideo - 49 Frames	81.1

Other Video and Language

- Youtube videos with titles
 - <http://aliensunmin.github.io/project/video-language/index.html#VTW>
- YouCook2 Dataset
 - <http://youcook2.eecs.umich.edu/>
- MSRVT: Microsoft Video and Text Dataset
 - <https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/>

Questions?