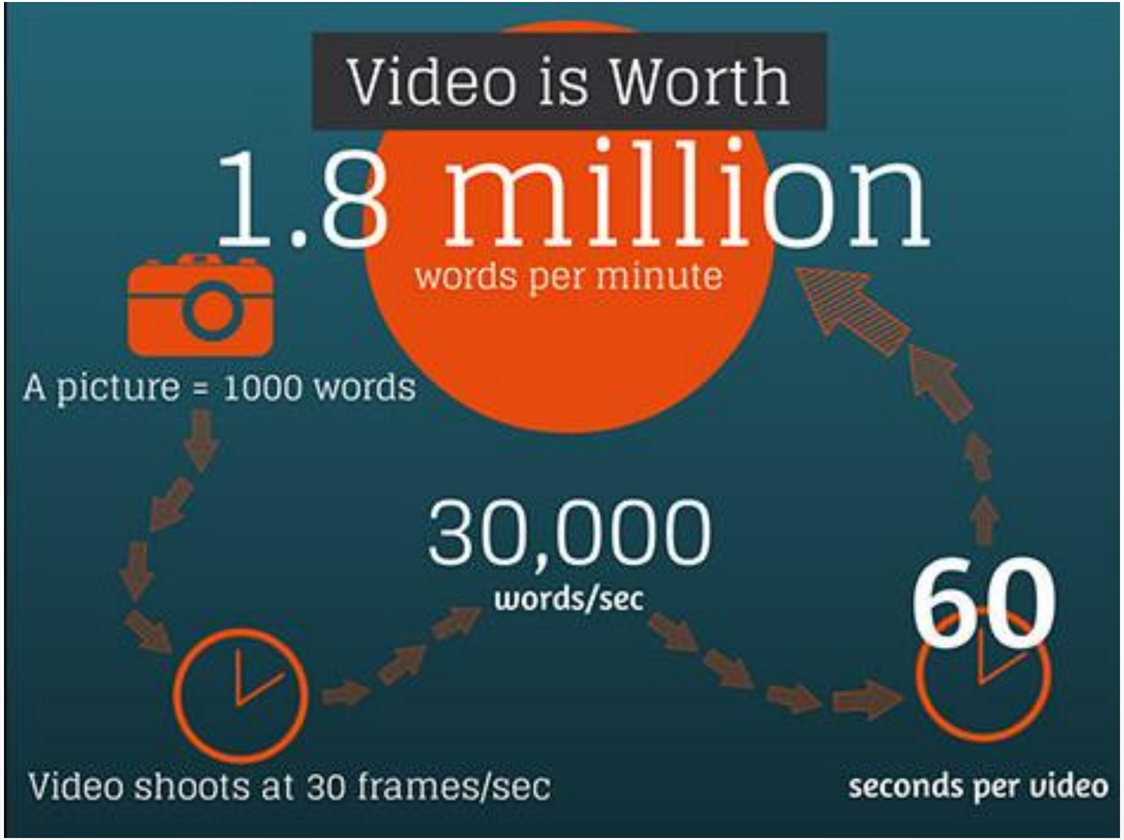
A Framework towards Domain Specific Video Summarization

Vishal Kaushal[1], Sandeep Subramanian[1], Suraj Kothawade[1], Rishabh Iyer[2], Ganesh Ramakrishnan[1]

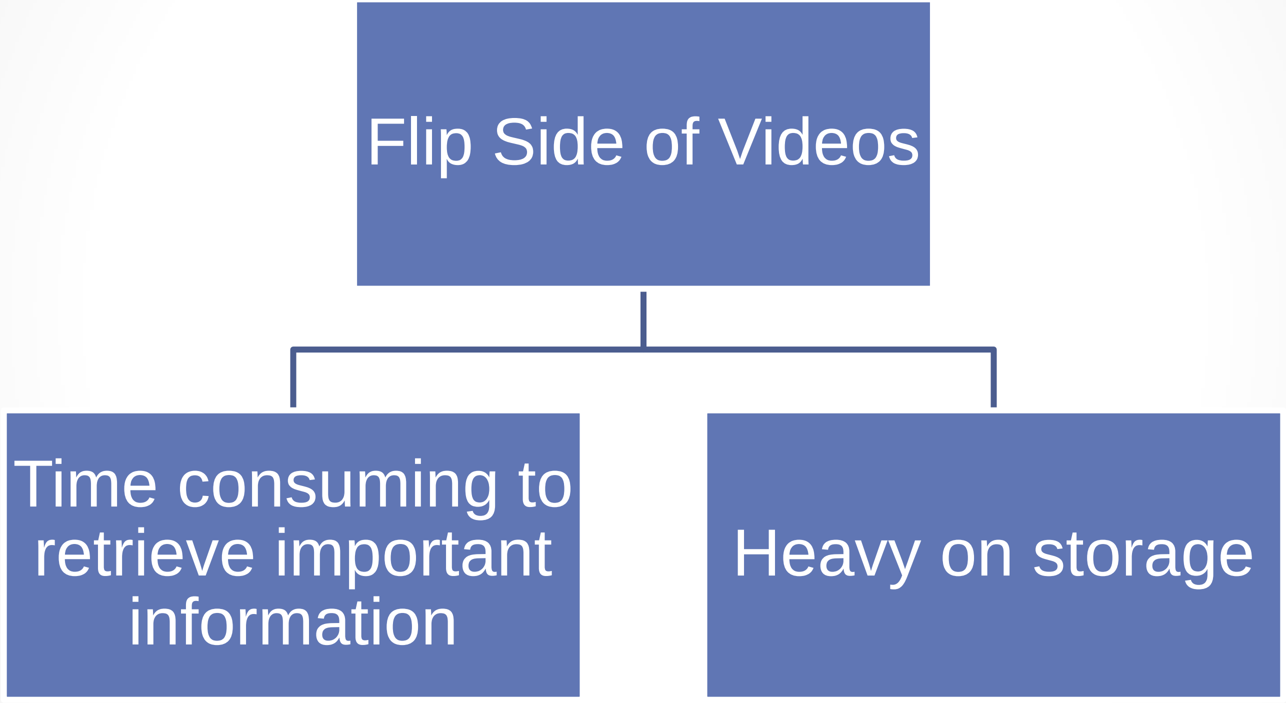
Indian Institute of Technology Bombay[1]
Microsoft Corporation[2]

Motivation



Video is Worth
1.8 million
words per minute
A picture = 1000 words
30,000
words/sec
60
Video shoots at 30 frames/sec
seconds per video

Motivation


Flip Side of Videos
Time consuming to retrieve important information
Heavy on storage

Motivation

- Growing focus on different techniques for Video Summarization
- Good summary?
  - Eliminate motionless chunks
  - Eliminate repetitive chunks
  - Retain what is **important**
- What is important for one domain is different from what is important for another domain
  - **Type of scenes** - Eg. Birthday (blowing candles, cutting cakes, ..), Soccer (kick, penalty, ..)
  - **Nature of summary** – Eg. Surveillance videos require outliers, TV Shows require representation

Different Domains

Surveillance Video

Birthday Video

Soccer Video

- Given a video of a particular domain, our system can produce a summary based on what is important for that domain
- Past related work has focused either on using supervised approaches for ranking the snippets to produce summary or on using unsupervised approaches of generating the summary as a subset of snippets with the above characteristics

Our Contributions

- **Joint problem** of learning domain specific importance of segments as well as the desired summary characteristic for that domain
- **Ratings** more effective as opposed to binary inclusion/exclusion information
  - In capturing the domain specific relevance
  - As unified representation of all possible ground truth summaries of a video, taking us one step closer in dealing with challenges associated with multiple ground truth summaries of a video
- A **novel evaluation measure**, more naturally suited in assessing the quality of video summary for the task at hand than F1 like measures
  - Leverages the ratings information and is richer in appropriately modeling desirable and undesirable characteristics of a summary
- A **gold standard dataset** for furthering research in domain specific video summarization
  - First dataset with long videos across several domains with rating annotations

Approach

Category	Number of Videos	Duration in mins
Cricket	7	276
Birthday	9	136
Soccer	11	609
Entry Exit	21	306
Office	33	687

- Created a training dataset
  - Birthday, Cricket, Soccer, Office, EntryExit
  - Scenes and ratings
- Weighted mixture of modular and submodular terms
  - Modular terms to capture the domain specific importance of snippets
  - Submodular terms like Set Cover, Facility Location etc. for imparting certain desired characteristics to the summary
- For each training video, components of the mixture are instantiated using different features and the weights of the complete mixture for that domain are learnt using max margin learning framework
- For any given test video of that domain, the weighted mixture is then maximized to produce the desired summary video

Formulation

$$y^* = \underset{y \subseteq Y_v, |y| \leq k}{\operatorname{argmax}} \; o(x_v, y)$$

$$o(x_v, y) = w^T f(x_v, y)$$

$$\min_{w \geq 0} \frac{1}{N} \sum_{n=1}^{N} L_n(w) + \frac{\lambda_1}{2} ||w_1||^2 + \frac{\lambda_2}{2} ||w_2||^2$$

$$L_n(w) = \max_{y \subseteq Y_v^n} (w^T f(x_v^n, y) + l_n(y)) - w^T f(x_v^n, y_{gt}^n)$$

Evaluation Measure

Positively Rated: Reward

$$S_V(y) = \sum_{x_i \in X_P} |y \cap x_i| * (1 + \frac{|y \cap x_i|}{|x_i|}) * e^{\alpha * rating(x_i)}$$

$$+ \sum_{x_i \in X_R} \min(|y \cap x_i|, \beta) * (1 + \frac{\min(|y \cap x_i|, \beta)}{\min(|x_i|, \beta)})$$

$$* e^{\alpha * rating(x_i)}$$

$$- \sum_{x_i \in X_N} |y \cap x_i| * k$$

Repetitive: Saturate

Negatively Rated: Penalize

Results

Domain	Method	ScoreLoss
Birthday	All Modular	0.7234
	All Submodular	0.7307
	Full	**0.6625**
	Random	0.7378
	Uniform	0.7569
	Submodular	0.7432
EntryExit	All Modular	0.5967
	All Submodular	0.6306
	Full	**0.5884**
	Random	0.7706
	Uniform	0.7785
	Submodular	0.6306
Cricket	All Modular	0.8140
	All Submodular	0.8275
	Full	**0.7733**
	Random	0.8911
	Uniform	0.8979
	Submodular	0.8275
Office	All Modular	0.3871
	All Submodular	0.4783
	Full	**0.3696**
	Random	0.5743
	Uniform	0.5399
	Submodular	0.5590
Soccer	All Modular	0.8849
	All Submodular	0.7645
	Full	**0.6533**
	Random	0.9217
	Uniform	0.8747
	Submodular	0.9152

Full mixture performs the best, as hypothesized

Results

Model Trained On	Model Tested On	ScoreLoss
Birthday	Birthday	**0.6625**
	Soccer	0.9753
	Cricket	0.9177
EntryExit	EntryExit	**0.5884**
	Soccer	0.9900
	Cricket	0.9710
	Birthday	0.8009
Cricket	Cricket	**0.7733**
	Soccer	0.8284
	Birthday	0.8103

Birthday	Random GTs	**0.6625**
	Same GT	0.6818
EntryExit	Random GTs	**0.5883**
	Same GT	0.6188

Models trained on one domain do not perform well on another – has learnt characteristics specific to that domain

Multiple GTs help!

Results: Top Individual Components

Birthday	**Mod:vgg_features** PSC:googlenet_p_concepts **SC:color_hist_r_features** **Mod:googlenet_features** PSC:yolo_coco_p_concepts PS SeC:yolo_coco_concepts **GC:vgg_features** **SC:color_hist_b_features** **FL:vgg_concepts**	**Mod:vgg_features** **GC:vgg_features** GC:googlenet_features **Mod:googlenet_features** FL:googlenet_features **FL:vgg_features** GC:color_hist_b_features **SC:color_hist_b_features** **SC:color_hist_r_features** GC:color_hist_r_features
Cricket	**SC:googlenet_features** **GC:googlenet_features** **Mod:vgg_features** **FL:color_hist_g_features** PSC:googlenet_p_concepts PSC:vgg_p_concepts **FL:googlenet_features** **GC:vgg_features** Mod:googlenet_features PS	**GC:vgg_features** **Mod:vgg_features** **GC:googlenet_features** **SC:googlenet_features** GC:color_hist_r_features SC:color_hist_r_features **FL:color_hist_g_features** SC:color_hist_s_features GC:color_hist_s_features **FL:googlenet_features**
Office	**Mod:yolo_coco_p_concepts** PSC:googlenet_p_concepts **Mod:vgg_features** **Mod:vgg_concepts** SeC:color_hist_r_features DM:color_hist_b_features **DM:vgg_features** **DM:yolo_coco_features** **DM:color_hist_b_features** **DM:color_hist_r_features**	**DM:googlenet_features** **Mod:yolo_coco_p_concepts** **DM:color_hist_b_features** **DM:color_hist_r_features** DM:color_hist_g_features PSC:vgg_p_concepts **DM:vgg_features** **Mod:vgg_features** GC:color_hist_g_features **Mod:vgg_concepts**

Left Column => Top Components based on learnt weights
Right Column => Top Components with highest individual score when optimized.

We see Strong correlation between the two!

Results: Relevance to Domain

Birthday	**Mod**:vgg_features **SC**:color_hist_r_features **PSC**:yolo_coco_p_concepts **SeC**:yolo_coco_concepts **SC**:color_hist_b_features	**PSC**:googlenet_p_concepts **SC**:googlenet_features PS **GC**:vgg_features **FL**:vgg_concepts
Cricket	**Mod**:googlenet_features **SC**:vgg_features **PSC**:googlenet_p_concepts **FL**:googlenet_features **Mod**:googlenet_features	**GC**:googlenet_features **FL**:color_hist_g_features **PSC**:vgg_p_concepts **GC**:vgg_features PS
Office (Surveillance)	**Mod**:yolo_coco_p_concepts **PSC**:color_hist_b_features **GC**:color_hist_r_features **DM**:yolo_coco_concepts **DM**:yolo_coco_features	**Mod**:vgg_concepts **Mod**:vgg_features **DM**:vgg_features **DM**:color_hist_b_features **DM**:color_hist_r_features

Object Features
Scene Features
Object Detectors (People etc.)
Color Features
Diversity Model
Representation Model
Coverage Model
Importance Model

Results: Best Snippets