

# Selecting semantically relevant video thumbnails

Gal Chechik

Tomas Izo

Samy Bengio

Google, Mountain View, CA, USA {gal,tizo,bengio}@google.com

Extracting the essential information from complex signals is a fundamental machine learning problem. A particularly complex case is that of audio-visual signals as in videos, where this problem has a very practical application: selecting thumbnails to represent videos in response to a search query. In this context, video thumbnails serve as a visual summary of the video's content, similar to text snippets that accompany the results of a web search. A well selected thumbnail should therefore tell users about the central events in the video, and summarize a whole video in a single image.

Since automatic understand and annotation of video scenes is extremely hard, existing approaches for video summarization focus mostly on selecting video frames that are *visually typical* in that they share visual characteristics with other frames in that video. For example, video frames could be clustered, and the most typical frames are used to represent the different shots.

The current abstract describes a very different approach to thumbnail selection, aiming to choose a thumbnail that captures the *semantics* of the video. Our approach, named *ThumbRank*, is based on obtaining a short textual description of the video, taken from the meta data that accompanies the video or from user queries that are known to retrieve the video. Then, we use an image search engine to collect images that match that textual description, and learn a model of the visual characteristics of those images. Finally, we use these models to select video frames that best represent the textual description according to the model learned on still images.

Our learning approach is based on PAMIR - a scalable online algorithm that ranks images for a given text query [Grangier and Bengio, 2006]. It was already shown to scale to handle tens of thousands of queries, and can be trained quickly on a single machine.

Human evaluation experiments with  $\sim 1000$  ratings of popular YouTube videos, show that 40% of human raters prefer *ThumbRank* thumbnails over 30% that prefer current YouTube thumbnails. This suggests that query-dependent thumbnails could be effectively learned even with large-scale and semantically heterogeneous video collections.

## The learning model and algorithm

To learn a model of the visual characteristics for a particular semantic category, we use a scalable online algorithm from the passive-aggressive family. Given a set of images, each represented as a vector of features  $\mathbf{p}_i \in \mathbb{R}^d$ , we assume that for every text query  $\mathbf{q}_i$ , we have access to images that are relevant to that query  $\mathbf{q}_i$  and images that are less relevant. Formally, a relevant image  $\mathbf{p}_i^+$  will have a higher relevance score than an irrelevant image  $\mathbf{p}_i^-$  for query  $\mathbf{q}_i$ :  $rel_{\mathbf{q}_i}(\mathbf{p}_i^+) > rel_{\mathbf{q}_i}(\mathbf{p}_i^-)$ . In practice, it is easy to collect such images by taking the set of images that are highly ranked in response to a text query on any web image search engine. Importantly, we only assume a weak level of supervision, and formalize the learning problem as a ranking problem, that correctly captures ordering among images.

Our goal is to learn a relevance score for each query  $\mathbf{q}_i$   $S_{\mathbf{q}_i}$  with the form:

$$S_{\mathbf{q}_i}(\mathbf{p}_i; \mathbf{w}_{\mathbf{q}_i}) \equiv \mathbf{w}_{\mathbf{q}_i} \cdot \mathbf{p}_i \quad (1)$$

with parameters  $\mathbf{w}_{\mathbf{q}_i} \in \mathbb{R}^d$ . Importantly, if an image is represented as a sparse vector, then  $S_{\mathbf{q}_i}$  can be computed very efficiently even when  $d$  is large. We propose an online algorithm based on the Passive-Aggressive (PA) family of learning algorithms introduced by [Crammer et al, JMLR 2006]. Here we consider an algorithm that uses pairs of images  $\mathbf{p}_i^+, \mathbf{p}_i^-$  for a given query  $\mathbf{q}_i$ , which obey  $rel_{\mathbf{q}_i}(\mathbf{p}_i^+) > rel_{\mathbf{q}_i}(\mathbf{p}_i^-)$ . For each query  $\mathbf{q}_i$ , we define the following hinge loss function:

$$L_{\mathbf{q}_i} = \sum_{(\mathbf{p}_i^+, \mathbf{p}_i^-)} l_{\mathbf{q}_i}(\mathbf{p}_i^+, \mathbf{p}_i^-) \quad \text{with} \quad l_{\mathbf{q}_i}(\mathbf{p}_i^+, \mathbf{p}_i^-) = \max \{0, 1 - S_{\mathbf{q}_i}(\mathbf{p}_i^+; \mathbf{w}_{\mathbf{q}_i}) + S_{\mathbf{q}_i}(\mathbf{p}_i^-; \mathbf{w}_{\mathbf{q}_i})\}. \quad (2)$$

To minimize  $L_{\mathbf{q}_i}$ , we apply the Passive-Aggressive algorithm iteratively to optimize  $\mathbf{w}_{\mathbf{q}_i}$ . First,  $\mathbf{w}$  is initialized to some value  $\mathbf{w}^0$ . Then, at each training iteration  $i$ , we randomly select a pair  $(\mathbf{p}_i^+, \mathbf{p}_i^-)$  for that query, and solve the following convex problem with soft margin:

$$\mathbf{w}^i = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{i-1}\|^2 + C\xi \quad \text{s.t.} \quad l_{\mathbf{q}_i}(\mathbf{p}_i^+, \mathbf{p}_i^-) \leq \xi \quad \text{and} \quad \xi \geq 0. \quad (3)$$

At each iteration  $i$ ,  $\mathbf{w}^i$  optimizes a trade-off between remaining close to the previous parameters  $\mathbf{w}^{i-1}$  and minimizing the loss on the current pair  $l_{\mathbf{q}_i}(\mathbf{p}_i^+, \mathbf{p}_i^-)$ . The *aggressiveness* parameter  $C$  controls this trade-off. Eq. 3 can be solved analytically and yields a very efficient parameter update rule.

## Experiments

We first illustrate semantic thumbnail selection using a specific video. The video followed a struggle between a zebra and a lioness (<http://www.youtube.com/watch?v=3HIL2R4cQao>), and had the word “zebra” in its title. We trained a PAMIR model for the word zebra, and applied to all frames in the video. Figure 1 traces the score for the term zebra, as it progresses through out the video. It also shows those frames that obtained the highest and lowest scores for the query zebra. The top scored frame indeed contain a clear image of two zebras.

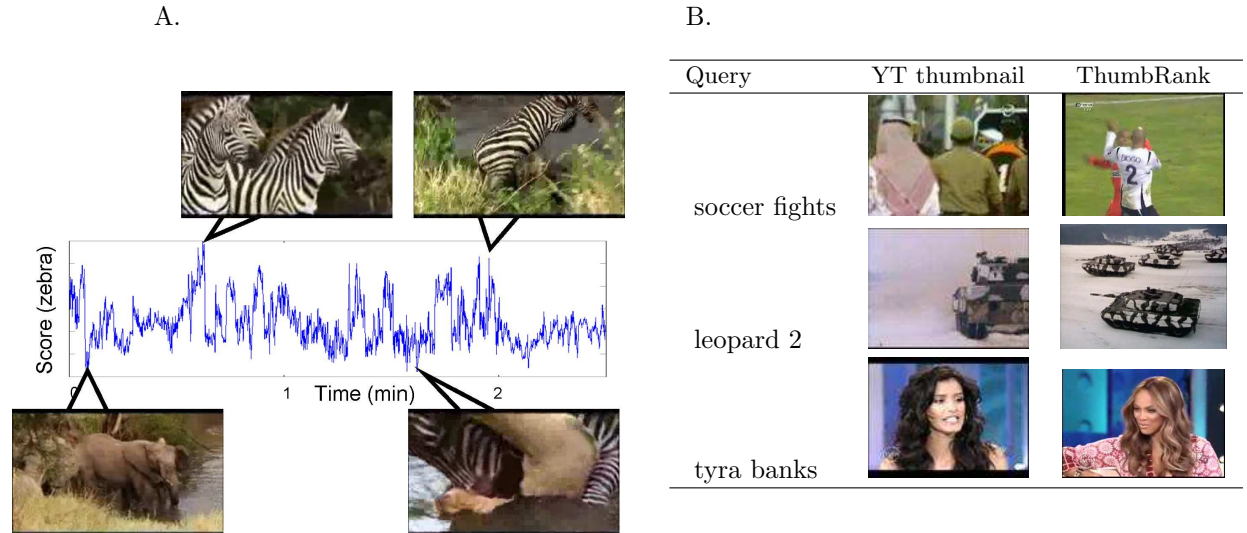


Figure 1: (A) Score for the query *zebra* traced along the video. (B) Examples of three queries and the corresponding thumbnails obtained by Thumbrank as compared to current YouTube thumbnails. The videos we used were [www.youtube.com/watch?v=UyDcJZzLcic](http://www.youtube.com/watch?v=UyDcJZzLcic), [www.youtube.com/watch?v=abqb-DqYXn](http://www.youtube.com/watch?v=abqb-DqYXn), [www.youtube.com/watch?v=jWuYoSNiLqo](http://www.youtube.com/watch?v=jWuYoSNiLqo).

To evaluate the quality of the selected thumbnails we conducted the following experiment: We processed 500 popular videos from YouTube, together with their metadata. For each video, we also had access to the text query that was most commonly used to retrieve that video. We trained a PAMIR model for each of the 500 queries. Finally, for each video, we used the PAMIR model trained for its textual description, to score each frame in the video. To speed the feature extraction process, videos were down sampled to have 2 frames per second. For each of the 500 video, we presented the semantic thumbnails along side with thumbnails that were chosen to be good representatives of the video (details elsewhere). Human raters were asked to inspect the two thumbnails, together with the original video and the text query, and mark their preferences. We found that 40% of the ratings favored the semantic thumbnails, and 30% favored the visually representative thumbnails.