Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Multimedia Data for the Visually Impaired

Niket Tandon
Max Planck Institute for Informatics
Saarbrücken, Germany
ntandon@mpi-inf.mpg.de

Shekhar Sharma
PQRS Research
pqrs-research.org
shekhar.sharmaa@gmail.com

Tanima Makkad
PQRS Research
pqrs-research.org
tanima.vit@gmail.com

Abstract

The Web contains a large amount of information in the form of videos that remains inaccessible to the visually impaired people. We identify a class of videos whose information content can be approximately encoded as an audio, thereby increasing the amount of accessible videos. We propose a model to automatically identify such videos. Our model jointly relies on the textual metadata and visual content of the video. We use this model to re-rank Youtube video search results based on accessibility of the video. We present preliminary results by conducting a user study with visually impaired people to measure the effectiveness of our system.

Introduction

For defining video accessibility, Web Content Accessibility Guidelines (http://www.w3.org/TR/WCAG20/) are commonly used given the easy rendering of audio description within videos and the availability of annotation tools (Champin et al. 2010). However, this involves two problems: First, assuming presence of audio description for the videos is unrealistic. Second, approaches that automatically generate audio description (Kobayashi et al. 2010) wrongly assume any video as a candidate to their approach. For instance, a video showing a robot machine in response to the query : "Artificial Intelligence" is of little use to the visually impaired user even with audio description.

This paper addresses the unexplored problem of finding videos that can be meaningful for the visually impaired users. We propose an accessibility measure for videos based on *how important audio content is to a video*.

We make use of relative changes in the frames depicting actions or events (Liu, Zhang, and Qi 2003) in a video. Small relative changes over a larger duration (e.g. in a speech) imply the importance of audio content in comparison to video content. We present a model to solve this problem.

The contributions of this paper are: (a) Currently videos without audio description are unusable by visually impaired users. Our approach overcomes this limitation to some extent by defining accessibility with a new paradigm of visual information contained in a video. (b) We propose a model to measure visual information in videos.

Model

Our task is to measure the amount of accessible information content in a video, i.e. the amount of information encoded in the audio. An accessible video can be approximately encoded as an audio i.e. the visual content of an accessible video carries little or no information. Typical examples of accessible videos include speeches, news recitation, lecture videos. Intuitively, there is little visual activity i.e. visual changes across frames in accessible videos.

We classify a video as accessible or not accessible based on the visual differences across neighboring frames. Textual metadata of the video like title and description provide cues to the accessibility of a video because inaccessible videos usually consist of keywords related to activities (e.g. a game) while accessible videos consist of keywords like speech, lecture, news. Therefore, we need a joint model that leverages both visual and textual cues.

Model Overview: We propose a joint model to leverage cues provided by both textual and non-textual (visual) features. Our model will have two distinct views of a video: a) text and b) video based features that are capable of classification (accessible or inaccessible) independently also (see Fig.1) . A combination of the two views leads to richer information for the model. Further, we have a scarcity of the expensive training data (c.f. §Experiments). Therefore, our goal is to use minimal supervision and increase prediction accuracy. We describe the two views next, and then propose the joint model over these two views.

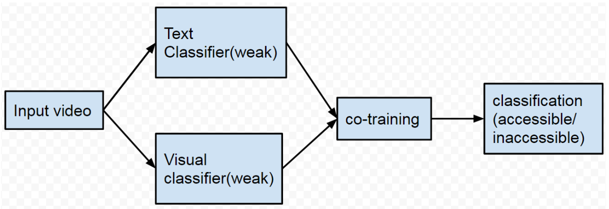

Figure 1: System architecture

Textual features: We identify discriminative function words that appear exclusively with accessible videos and inaccessible videos respectively. We construct a vector of binary features over the metadata of a video like title, description, tags and if available, comments and video category.

Visual features: We use visual features to estimate differences across neighboring video frames. Frame changes either happen due to changing foreground (motion of objects) or changing background (scene change), namely:
(a) Motion based features are used to estimate the rate of change of frames in a video, namely:
Shot change: To detect average length of the shot in video. A shot is a sequence of frames taken from a single camera without any change in color content of consecutive images.
Intensity change: To estimate the difference between two frames, intensity change is measured in the background.

(b) Content based features : To estimate the rate at which new content (new scenes, objects, people) appears in a video, we use content based features, namely:
Scene change: A scene is defined as a sequence of frames in a video that describe one scenario. Higher the scene change, higher the difference between two scenes i.e. background.
Enhanced Hue Saturation Value (HSV): The HSV is computed as the sum of change in color histogram and spatiogram, where the change is aggregated over all pairs of consecutive frames (Kailath 1967) is $\sum_i^n \sqrt{(\sum a_i \cdot \sum b_i)}$, where a and b are one pair of consecutive frames and n bins.

Joint model: Co-training (Blum and Mitchell 1998) is a semi-supervised approach that uses two independent feature set of training data over two weak classifiers such that high confidence classifier from either classifier becomes training data for the other. Co-training is a good fit for our scenario because we have scarcity of training data and there are two independent views available using textual and visual classifiers that are both weak independently (see Algorithm 1).

Algorithm 1 Co-training algorithm

```
Require: $L$: set of labeled, $U$: set of unlabeled points. Let
    $h_1$ = textual feature based classifier, $h_2$ = Visual feature
    based classifier, $F_1$ = textual and $F_2$ = visual features.
for $K$ iterations do
    Train classifiers $h_1$ and $h_2$ using $F_1$ and $F_2$ views of
      L resp., let $h_1$ and $h_2$ classify the instance in $U$.
    Compute confidences of the classified instances in $U$
      with $h_1$'s and $h_2$s confidence measure.
    If $h_1$ classifies $U$ with higher confidence, move la-
      beled data to training set of $h_2$ and vice-versa.
end for
```

Experiments

Labeled data In order to construct labeled data, we conduct a user study with a group of 20 visually impaired people, over a dataset of 80 videos from various categories like news, games, speeches from Youtube. The task is to label a video as accessible or inaccessible. The users only hear the sound of the video and were asked if the video was easy to follow. As an additional source of implicit labeling, we prepare a questionnaire on the prominent content of the video. For this implicit label, we mark a video as accessible if the majority of the questions were correctly answered. We found a high correlation between the direct and indirect labels. Each video is labeled by a minimum of three users. We use majority vote to label the video. We collected a set of 38 accessible and 42 inaccessible videos through this process. 56 labeled videos were used as training set and rest as test set to measure accuracy of various classifiers.

Weak classifiers We train two weak classifiers, text features based and video features based as explained in §Model. For classification, we use SVM and Decision trees. Textual classifier's precision is 0.61 using SVM and visual classifier achieves 0.67 using SVM.

Baseline As baseline, we consider a SVM over a combination of text features and visual features in a disjoint setting. This achieves a precision of 0.75.

Our model Co-training based joint classifier starts with a very small number of labeled data points and can recursively label large data set. The joint model uses two separate views. Our model achieves a precision of 0.91 and outperforms the baseline and the weak classifiers by a large margin.

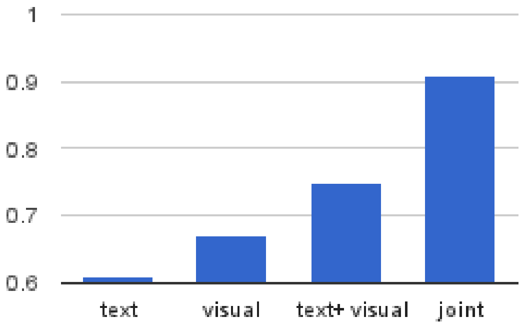

Figure 2: Precision of different approaches

Use case We re-rank Youtube results to reflect accessibility as: $\frac{1}{i+1}.score_{acc}$ where i is the original rank. $score_{acc}$ is our joint classifier's confidence for accessible label. [1]

Conclusion

We presented an approach to increase the accessible video data for the visually impaired people. Our model considers two views of the text and video features and outperforms either of them or their combined feature set. Finally, we use our model to construct a system that re-ranks youtube search results tailored for the visually impaired users.

References

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *COLT*, COLT' 98, 92–100. ACM.

Champin, P.-A.; Encelle, B.; Evans, N. W.; O-Beldame, M.; Prié, Y.; and Troncy, R. 2010. Towards collaborative annotation for video accessibility. In *W4A*, 17. ACM.

Kailath, T. 1967. The divergence and Bhattacharyya distance measures in signal selection. *Communication Technology* 52–60.

Kobayashi, M.; O'Connell, T.; Gould, B.; Takagi, H.; and Asakawa, C. 2010. Are synthesized video descriptions acceptable? In *SIGACCESS*, 163–170. ACM.

Liu, T.; Zhang, H.-J.; and Qi, F. 2003. A novel video key-frame-extraction algorithm based on perceived motion energy model. *Circuits and Systems for Video Technology, IEEE Transactions on* 13(10):1006–1013.

[1] Supplementary material at http://bit.ly/gyanjyoti