IRISA at TrecVid2015: Leveraging Multimodal LDA for Video Hyperlinking

Rémi Bois [(a)], Anca-Roxana Şimon [(b)], Ronan Sicre [(b)],
Guillaume Gravier [(a)] Pascale Sébillot [(c)]
(a) CNRS, (b) Univ. Rennes 1, (c) INSA Rennes,
IRISA & Inria Rennes
firstname.lastname@irisa.fr

Abstract

This paper presents the runs that we submitted in the context of the TRECVid 2015 Video Hyperlinking task. The task aims at proposing a set of video segments, called *targets*, to complement a query video segment defined as *anchor*. We used automatic transcripts and automatically extracted visual concept as input data. Two out of four runs use cross-modal LDA as a means to jointly make use of visual and audio information in the videos. As a contrast, one is based solely on visual information, and a combination of the cross-modal and visual runs is considered. After presenting the approaches, we discuss the performance obtained by the respective runs, as well as some of the limitations of the evaluation process.

1 Introduction

Recently, the automatic generation of hyperlinks in videos became a subject of growing interest. Most approaches were proposed in the context of the Search and Hyperlinking benchmark at MediaEval [3, 2] and, more recently, at TRECVid [5]. Video hyperlinking consists in establishing links between video fragments that share the same, or similar, topics, within a video collection. A link relates a source, called an anchor, and a target, both being video segments within large video streams. Starting from a set of anchors given by users, targets are determined in a large collection of video streams for each anchor, based on similarity criteria. Application wise, the main goal of creating hyperlinks is to offer information seeking and browsing capabilities in addition to standard search features. Furthermore, instead of retrieving full documents the aim is to retrieve only the video segments that are relevant to the predefined anchors, with specific jump-in points.

The creation of hyperlinks usually implements two steps: first, potential target segments are extracted over the entire video database; second, the most relevant targets are selected for each anchor, relying on content analysis and similarity measures. In this last step, most of the existing approaches for video hyperlinking focus on direct pairwise content-based similarity, varying the weighting schemes and the information used (e.g., named entities, metadata, transcripts, visual features), and end up favoring links between anchors and very similar targets. We believe however that an important aspect of video hyperlinking is to be able to offer diverse targets, favoring serendipity, i.e., unexpected targets that are deemed relevant. The cross-modal approaches that we propose focus on offering increased variety in the links, by searching links that would not be captured with classical direct content comparison.

We propose a new way to increase diversity in targets with multimodal information, relying on a cross-modal generative model inspired by the bilingual LDA model (BiLDA) [8] (RUN-2 and RUN-3). We compare those runs with a single-modal run based on visual description (RUN-1) and a rank-aggregation run (RUN-4). We use BiLDA to create a probabilistic translation model between visual concepts and words pronounced in the videos. The links are created using content-based comparisons between anchor and target pairs, translating from one modality to the other. This translation offers a richer context and allows the creation of diverse links. Additionally, an anchor that has no visual concept associated will have the chance to be linked to targets that visually “show” information related to what is “talked about” in the anchor segment. The same analogy can be made for anchors with no associated transcripts.

The rest of the paper is organized as follows. First, we briefly present the data used in the respective runs as well as the segmentation step. Next, we describe the four runs that were submitted, two of them taking advantage of the novel cross-modal approach. Finally, we discuss the results and show that diversity in the links proposed was not rewarded by evaluators in the TRECVid context.

2 Data and Settings

The design choice that we made was to be as close as possible to a real-life situation. We thus rely only on data that could be obtained automatically. In particular, we discarded reference transcription to the benefit of the automatic transcripts offered by LIMSI [1]. These automatic transcripts correspond to the audio content description on which the cross-modal runs are built. We also use the visual concepts extracted by Leuven for visual content description. We chose to ignore the metadata, which were significantly poorer than in previous years.

That task definition states that for each anchor, we have to propose a set of short-length targets, thus requiring segmentation of long videos. For efficiency reasons, we used (quasi) fixed-length segments, independent of the anchor. We chose a 90 seconds split of the video, growing segments after 90 s up to the next pause in the video as indicated by breath intakes in the transcripts. This simple approach may cut coherent segments in several parts, but we judged this

$$\begin{bmatrix} t_1 \quad t_2 \quad ... \quad t_n \\ anchor\ concepts \end{bmatrix} \begin{bmatrix} t_{11} \quad t_{21} \quad ... \quad t_{m1} \\ t_{12} \quad t_{22} \quad ... \quad t_{m2} \\ ... \quad ... \quad ... \quad ... \\ t_{1n} \quad t_{2n} \quad ... \quad t_{mn} \\ transposed\ targets \end{bmatrix}$$

$$= \begin{bmatrix} score_1 & score_2 & ... & score_m \end{bmatrix}$$

Figure 1: Matrix multiplication for visual scoring

issue not to be crucial as we are looking for short enough segments as targets (a maximum of two minutes was allowed), that are often considered as jump-in points.

To skirt issues with very short anchors, we considered the context of the anchors, i.e., considering 30 seconds before and 30 seconds after the start and end points respectively.

3 Runs

3.1 Visual similarity (RUN-1)

For the experiments relying on visual information, we use the visual concepts provided by Leuven. Each video is represented as a set of keyframes for which visual concepts scores are available. These 1,537 concepts are composed of the 1,000 classes of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2010 and of 537 classes indicated as "popular" by ImageNet [10]. For concept detection, images were described with dense SIFT features encoded with Fisher vectors and classified with a SVM trained using 10K Flickr images as negative examples. For every keyframe, the classifiers were applied and only normalized scores higher than 0.7 were kept.

We used the visual concepts as a way to represent anchors and potential targets. For each video fragment, a concept matrix was built. A concept matrix is a matrix of size $N * M$ where N is the number of total concepts that appear in the collection and M is the number of keyframes available for the video segment. In this case, $N = 1,537$ and M varies between 1 and a few dozens. As in multiple query image retrieval [6], keyframe concept descriptors were combined by an average over all keyframes followed by *l2*-normalization, yielding a N-dimensional descriptor for a segment. The similarity between two video segments is finally computed as the dot product between their respective descriptors, as shown in Figure 1 where all potential targets are considered.

3.2 Multimodal LDA (RUN-2 and RUN-3)

There exists video segments that are about the same, or similar subject, without sharing much vocabulary or visual information. They can however share concepts in two distinct modalities (e.g. someone talking about castles, and a different segment showing castles). A way to connect such segments is to translate the representation of one segment from one modality to another via a shared representation space reachable from either one of the modality. Thus, segments can be linked based on the shared information in the new representation space. To be able to translate from one modality to another, we propose to decompose the video collection into multimodal topics. A bilingual LDA (BiLDA) topic model is leveraged to learn cross-modal translation for each topic.

In order to create a bi-modal LDA model, we assume that instead of translating from a source language (e.g, French) to a target language (e.g., English), as achieved by BiLDA, we do a translation between a source modality (e.g., audio) to a target modality (e.g., visual). To do so, we generate a collection of parallel documents: audio is obtained by considering the automatic transcripts, while visual information is obtained through learned visual concepts scores computed for each keyframe in the video. Based on this collection, the BiLDA model is trained and the resulting probabilistic translation model is used to create links between anchors and targets.

3.2.1 Building the cross-modal topical structure

The BiLDA topic model is a bilingual extension of the classical LDA model. These two latent topic models are based on the idea that there exist latent variables, i.e., topics, which determine how words in documents have been generated. Fitting such a generative model means finding the best set of those latent variables in order to explain the observed data. As a result documents are seen as mixtures of latent topics, while topics are represented as probability distributions over the words in the vocabulary.

In the classic LDA model, each document is assumed to have a specific distribution over topics. Meanwhile, BiLDA assumes that each document pair (from each languages or modalities) shares the same distribution over topics θ. Therefore, the latent topics learned are language-independent, but each language has a language-specific association to topics. In other words, each latent topic is characterized by two probability distributions, one over each language.

In our approach, we learn cross-modal topics instead of cross-lingual ones. Therefore, instead of having a comparable corpora of documents in two languages, we have paired bimodal (i.e., audio and visual) comparable documents. Learning the cross-modal topics is achieved by training BiLDA using Gibbs sampling with standard values for the hyper-parameters $\alpha = 50/K$ and $\beta = 0.01$ [9], where K denotes the number of latent topics. After training, a set of word distributions ϕ, one for each topic, and of visual concepts distributions ψ are obtained. These distributions enable to measure the contribution of each word/visual concept for a particular multimodal topic z_j. Given the documents in the text

Topic 3	text	love, home, feel, day, life, baby, made
$K = 700$	visual concepts	singer, microphone, sax, concert, flute
Topic 25	text	years, technology, find, computer, key, future
$K = 700$	visual concepts	tape-player, computer, equipment, machine

Table 1: Representation of two cross-modal topics, with their top-words and visual concepts.

(resp. visual) modality, with vocabulary V_1 (resp. V_2), the probability that a word $w_i \in V_1$ (resp. a visual concept $vc_i \in V_2$) is generated by topic z_j is given by

$$p(w_i|z_j) = \phi_{j,i} = \frac{n_{z_j}^{w_i} + \beta}{\sum\limits_{x=1}^{|V_1|} n_{z_j}^{w_x} + \beta|V_1|} \quad , \tag{1}$$

with $n_{z_j}^{w_i}$ the number of times topic z_j was assigned to an occurrence of w_i in the training documents. The sum in the denominator corresponds to the total number of word occurrences assigned to topic z_j and β is a Dirichlet prior. Similarly, the probability of a visual-concept being generated by topic z_j is denoted $p(vc_i|z_j) = \psi_{j,i}$.

Two examples of multimodal topics learned on our data set are given in Table 1, where the words/visual concepts with the highest probability are given. In the example, topics can be characterized as resp. singing and technology and are adequately represented in both modalities.

3.2.2 Linking anchors and targets

We use the previously introduced model to change the representation space of the anchor and target segments, moving to the space of topics. Two probabilities are computed for each segment-topic pair, one based on topic-wise word distributions, the other based on topic-wise visual concepts distributions. The resulting representations are obtained by computing the probability of each anchor (resp. target) segment given the topics learned in both modalities. We normalize the probabilities over the topics to sum to one. For the textual representation of an anchor a, the probability is computed as

$$p(a|z_j) = \left(\prod_{i=1}^{n_a} p(w_i|z_j) \right)^{1/n_a} \quad , \tag{2}$$

where n_a is the size of the vocabulary in the anchor segment. For the visual concepts representation the probability $p(w_i|z_j)$ becomes $p(vc_i|z_j)$. The two modalities' representations are computed in the exact same way.

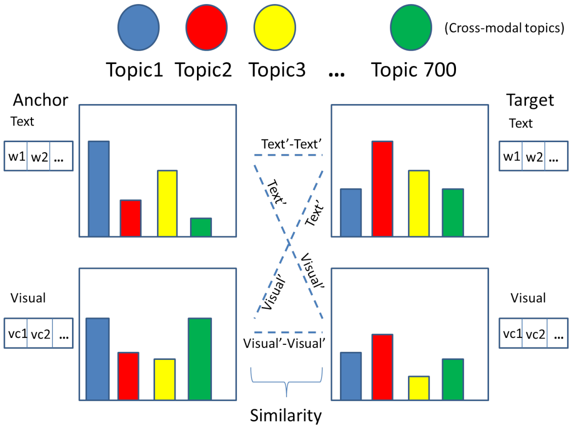

Figure 2: Representation of the translation for the anchor and targets segments and the 4 strategies that can be used for content-based comparisons in this new space. Histograms present the importance of each topic for each anchor and target and are computed both on the audial and visual content of the anchor/target. Anchors and targets sharing the same, or similar, subject (visually or audially) have similar distributions over cross-modal topics.

Figure 2 depicts the change in representation space and the content-based comparisons made within the new space. Taking advantage of the fact that the topics were jointly learned from audio and visual features, one can use any paired representation (audio→audio, visual→visual, audio→visual, visual→audio) to compute a similarity score for an anchor-target pair. We selected audio→visual (RUN-2) and visual→audio (RUN-3) as two of our runs. The scores were computed after a *l2*-normalization of the histograms of topics. For each anchor, the ranking of the targets is achieved based on the obtained scores.

3.2.3 Reranking on RUN-2 and RUN-3

For the two cross-modal runs (RUN-2 and RUN-3), we used reranking on a shortlist of targets (top 50 obtained for each anchor with cross-modal comparison) in order to refine the proposed targets. We used a ngram cosine similarity to rerank visual-audio representations (RUN-2) as it proved efficient in previous works [4]. This ngram cosine was computed with unigrams cosine (weighted 2), bigrams cosine (weighted 3), and trigrams cosine (weighted 5). Weights were obtained empirically on preliminary experiments. Convolutional neural networks (CNN) were used to perform visual reranking on RUN-3. We trained the CNN on the ImageNet ILSVRC classification dataset, using the very deep convolutional network architecture from [7]. The network is composed of 13 convolution layers and 3 fully connected layers. We extracted the output of the second fully connected layer before applying ReLU and performed *l2* normalization. We fused the keyframes descriptors in the same way as RUN-1, by

	RUN-1	RUN-2	RUN-3	RUN-4
P@5	0.2140	0.0160	0.2580	0.1760
P@10	0.2070	0.0170	0.2240	0.1560

Table 2: Precision at ranks 5 and 10, for Visual (RUN-1), Audio→Visual (RUN-2), Visual→Audio (RUN-3) and Rank Aggregation (RUN-4).

averaging all the concepts in each keyframe and computed the similarity via a dot product.

3.3 Rank Aggregation (RUN-4)

Our last run uses results from the three previous runs, as well as a pure ngram scoring. We combine them by adding the rankings and reordering them in decreasing order. Targets can thus be heavily impacted by a low score on one of the systems, while homogeneous targets, that perform reasonably well with most systems, are favored.

4 Results and discussion

4.1 Results

Results are reported in Table 2. As we can see, we have a surprisingly low score for the audio→visual (RUN-2) method. This result is counterbalanced by the fact that our best run overall is the similar in idea visual→audio (RUN-3) method. We are still investigating the reasons for such a low score, a time-consuming task. We also notice that the rank aggregation (RUN-4) achieves medium scores, and combining ranks does not seem to bring more information.

We can see that the visual→audio (RUN-3) method performs better than the purely visual method (RUN-1), a hint that using both modalities does offer better targets. The precisions at 5 and 10 are similar, indicating that for the method is robust.

4.2 Discussion

We started analysing our best run, the visual→audio (RUN-3) cross-modal run, and found some surprising results. As opposed to previous years, targets that come from the exact same show were discarded for evaluation. The rationale behind this decision was that proposing links that were already watched by the user is of little interest. However, due to the very nature of the data, we can still find duplicates, albeit in distinct shows. That is the case for anchors 82-83-84-85, which all discuss the “No” from Ireland to the Lisbon treaty after a national referendum. This information was important enough to be reported in consecutive news shows, using the same content. Hyperlinking to those near-duplicates, that show the same images and do not bring any new information, has been judged as relevant in this task. The evaluation thus continues encouraging systems to propose near-duplicates.

On the contrary, targets that are not close enough to the anchor were judged not relevant, even when they were related and brought new information. As a matter of example, for anchor 85, we proposed a target that shows a debate within the U.K. parliament, where members argue for pressuring Ireland into making a second vote in a new referendum. Unfortunately, this is not uncommon in the evaluation, and probably the result of having only one annotation per proposed target, with no way to compute an inter-annotator agreement that could emphasize the discrepancies that can exist on such a task.

5 Conclusion

In this paper, we described our four runs for the TRECVid 2015 Hyperlinking task, and introduced a novel method for cross-modal models. We proposed a new method designed to bring more diversity in the targets proposed and that outperforms single-modality comparison. While more in-depth study is necessary to better understand the results of the evaluation, we are confident that the community will push for the removal of near-duplicate targets in further editions, and encourage the design of methods that favorise diversity and serendipity.

References

[1] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In *European Conference on Information Retrieval*, pages 187–198, 2008.

[2] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking task at MediaEval 2014. In *Working Notes Proc. of the MediaEval Workshop*, 2014.

[3] M. Eskevich, G. J. F. Jones, R. Aly, and et al. Multimedia information seeking through search and hyperlinking. In *ACM International Conference on Multimedia Retrieval*, 2013.

[4] C. Guinaudeau, A.-R. Simon, G. Gravier, and P. Sébillot. Hits and irisa at mediaeval 2013: Search and hyperlinking task. In *MediaEval working notes*, page 2, 2013.

[5] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quenot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*, 2015.

[6] R. Sicre and H. Jégou. Memory vectors for particular object retrieval with multiple queries. In *ACM International Conference on Multimedia Retrieval*, pages 479–482. ACM, 2015.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[8] W. D. Smet and M. Moens. Cross-language linking of news stories on the web using interlingual topic modelling. In *2nd ACM Workshop on Social Web Search and Mining*, 2009.

[9] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[10] T. Tommasi, R. B. N. Aly, K. McGuinness, K. Chatfield, and et al. Beyond metadata: searching your archive based on its audio-visual content. In *International Broadcasting Convention*, 2014.