# IRISA at TRECVid2016: Crossmodality, Multimodality and Monomodality for Video Hyperlinking

Rémi Bois* Vedran Vukotić** Ronan Sicre
Christian Raymond** Guillaume Gravier*
Pascale Sébillot**

IRISA & INRIA Rennes,
*CNRS, **INSA Rennes,
E-mail: firstname.lastname@irisa.fr

**Abstract**

This paper presents the runs that were submitted to the TRECVid Challenge 2016 for the Video Hyperlinking task. The task aims at proposing a set of video segments, called targets, to complement a query video segment defined as anchor. The 2016 edition of the task encouraged participants to use multiple modalities. In this context, we chose to submit four runs in order to assess the pros and cons of using two modalities instead of a single one and how crossmodality differs from multimodality in terms of relevance. The crossmodal run performed best and obtained the best precision at rank 5 among participants.

## 1 Introduction

The automatic creation of hyperlinks in videos is being investigated for the second time in TRECVid [1, 2]. This task consists in establishing links between video fragments that share a similar topic in a large video collection. Links are created between a source, called anchor and targets. Both anchors and targets are video segments.

Anchors are generated by users and provided for the Hyperlinking task. Participants design systems, which indicates video segments that are considered relevant with respect to each individual anchor. This relevance criterium is often implemented as similarity criteria.

The challenge is then to offer targets that are similar enough to be considered as rightfully linked to the anchor, but dissimilar enough to not be redundant. Offering diversity by proposing dissimilar enough targets is a good way to offer some serendipity [3, 4], *i.e.* unexpected yet relevant links.

The creation of hyperlinks consists in two steps: a segmentation step, in which potential target segments are extracted over the entire video database and a ranking step, in which most relevant targets are selected for each anchor, relying on content analysis and similarity measures. Both steps are subject to many decisions.

One could use a naive segmentation approach compensated by widely overlapping segments, or a smart segmentation allowing for the automatic removal of low interest video parts. In the first case, overlapping segments offer more opportunities to find a good matching video part for the anchor, while the second approach allows for more costly comparisons at the ranking step due to the lower amount of video pairs to compare.

As for the ranking step, many aspects of the videos can be taken into account, from what is shown to what is said, from how it is said to how it is shown. The Hyperlinking task participants have many resources available, including automatic transcriptions, keyframes, visual concepts extracted from the keyframes and user-created metadata.

Our goal in this task is to compare monomodal, multimodal and crossmodal approaches.

The two modalities that we experimented with are the audio modality, more specifically what is said and the visual modality. We did not use any of the user-created metadata. Our crossmodal approach is state of the art and obtained the best results in terms of precision at rank 5 for the task. Our two monomodal runs use the models used to build the crossmodal run. Our multimodal run uses a shallow approach in order to retrieve very similar targets.

The rest of the paper is organized as follows: First, we talk about the data that we used in our respective runs and the data segmentation that is used. Second, we describe our four runs. Finally, results are presented.

## 2 Data and Segmentation

The dataset under scrutiny this year is the BlipTV dataset [2], composed of 14.838 videos with a mean duration of 13 minutes. These videos span multiple languages including English, Chinese, Arabic, etc.

Our four runs used the automatic transcriptions provided by LIMSI [5] and one of our runs (the shallow similarity run) used the automatically extracted concepts provided by EURECOM. The other available data (metadata and shot boundaries) are not used. We chose not to use the metadata as these are user-generated since we only want automatically available data.

We chose not to use shot boundaries as we exploit a speech-based rather than a shot-based segmentation, in order to not cut a segment in the middle a sentence.

Previous years experiments showed that assessors mostly evaluated relevance by watching the beginning of the target videos. To account for this behavior, we chose to segment videos by taking only 30 seconds of contiguous speech (down from last years' 90 seconds) and then cut at the following breath. We run this

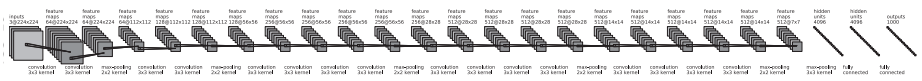

Figure 1: simplified illustration of the VGG-19 architecture - it contains 19 "weighted" (convolutional or fully connected) layers from where the last fully connected one was used to obtain higher level visual representations

segmentation process twice, using an offset of one speech segment at the second pass, in order to obtain an overlapping segmentation. The resulting 307.403 segments have a mean duration of 45 seconds.

We extracted corresponding transcripts and keyframes for each of the segments. Transcripts are then preprocessed with a tokenization step and a stopwords removal step. Stopwords lists were either gathered online[1] or directly from NLTK[6]. These preprocessed transcripts were used in all of our runs.

Regarding the monomodal continuous representation spaces used in three runs, we chose to represent the transcripts of each anchor and target with a *Word2Vec* skip-gram model with hierarchical sampling [7], a representation size of 100 and a window size of 5. This was shown to work best in similar setups [8]. Visual embeddings are obtained from a very deep convolutional neural network (CNN) VGG-19 [9], pretrained on *ImageNet*, by extracting the last fully-connected layer, as shown in Figure 1. Therefore, we obtain a 4096 dimensional embedding for each keyframe. Embeddings are simply averaged over the video segments as it performed well in a similar setup [10].

# 3 Multimodal Shallow Run

Our multimodal run is based on shallow similarity. The rationale for using only shallow similarity was that assessors are sensible to repetition when rating for relevance. A posteriori experiments on previous years datasets showed that near-duplicates were systematically considered as relevant. This run was designed to find such near-duplicates, in order to have a clear idea of the amount of near-duplicates available in this year's dataset.

For each anchor/target pair, we compute two distinct scores, one based what is heard *i.e.* transcripts from LIMSI [5] and one based on what is seen *i.e.* visual concepts from EURECOM. Transcripts are preprocessed and used to obtain the target representation as a TF∗IDF vector [11]. Inverse document frequencies are computed on the whole dataset, considering each segment as a document. We then compute a cosine similarity between the audio representations of each anchor/target pair to obtain the audio score.

A single video segment can contain multiple keyframes. We chose to sum the scores of the concepts present in the keyframes of the segment to obtain a visual representation. The vector we obtained has the size of the number of

[1] `https://github.com/6/stopwords-json` for Chinese, Romanian, Greek and Arabic

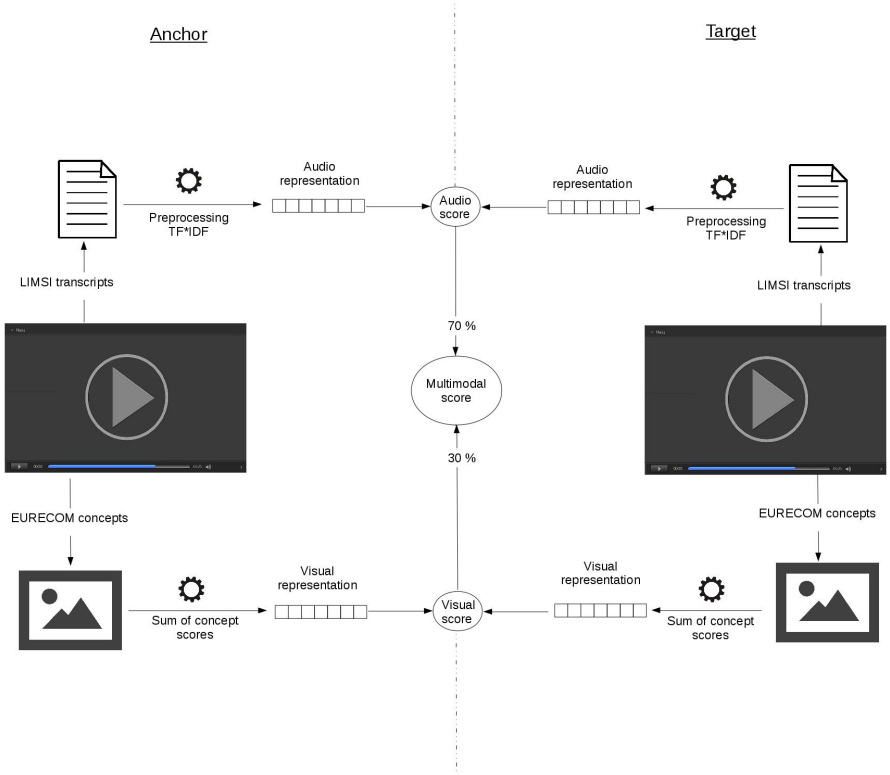


Figure 2: Multimodal score between an anchor and a potential target

concepts available in the whole dataset. For each anchor/target pair, we then compute a visual score by using a cosine similarity.

Those two scores are then combined with a linear combination. We use a weight of 0.7 for the audio modality and a weight of 0.3 for the visual modality. These weights were obtained empirically upon testing on last year's TRECVid dataset. Figure 2 illustrates the full process.

# 4 Crossmodal Bidirectionnal Joint Learning Run

With the core idea of trying to maximize both relevance and diversity between the anchors and their corresponding targets, we opted to use bidirectional deep neural networks - a novel variation on multimodal/crossmodal autoencoders that seems to improve diversity and relevance in early studies with a fixed groundtruth [8, 10]. The seminal idea of bidirectional deep neural networks is to use separate deep neural networks for each crossmodal translation while tying the weights of the middle layers between the neural networks so as to yield a common multimodal representation. In this setting, the common middle layer

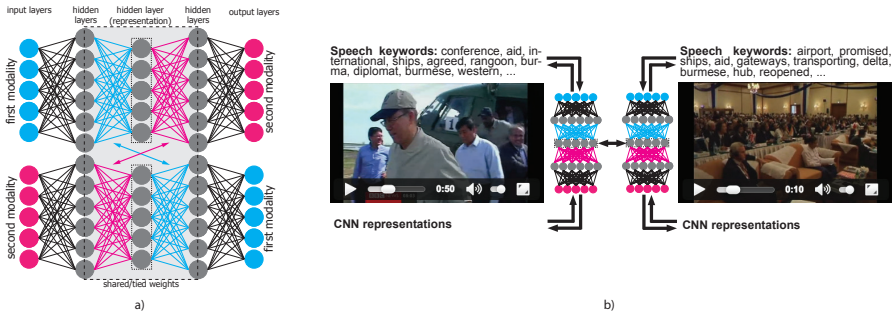


Figure 3: a) BiDNN architecture [8]: training is performed crossmodally and in both directions; a shared representation is created by tying the weights (sharing the variables) and enforcing symmetry in the central part b) video hyperlinking with BiDNNs [10]: both modalities (aggregated CNN embeddings and aggregated speech embeddings) are used used in a crossmodal translation by BiDNNs to form a multimodal embedding. The same is done for both the anchor and the target video segments. The newly obtained embeddings are then used to obtain a similarity score.

acts as a common multimodal representation space that is attainable from either one of the modalities and from which we can attain either one of the modalities.

In bidirectional deep neural networks, learning is performed in both directions: one modality is presented as an input and the other as the expected output while at the same time the second one is presented as input and the first one as expected output. This is equivalent to using two separate deep neural networks and tying them (sharing specific weight variables) to make them symmetrical, as illustrated in Figure 3 a). Implementation-wise the variables representing the weights are shared across the two networks and are in fact the same variables. Learning of the two crossmodal mappings is then performed simultaneously and they are forced to be as close as possible to each other's inverses by the symmetric architecture in the middle. A joint representation in the middle of the two crossmodal mappings is also formed while learning.

Formally, let $\mathbf{h_i^{(j)}}$ denote the activation of a hidden layer at depth $j$ in network $i$ ($i = 1, 2$, one for each modality), $\mathbf{x_i}$ the feature vector for modality $i$ and $\mathbf{y_i}$ the output of the network for modality $i$. Networks are defined by their weight matrices $\mathbf{W_i^{(j)}}$ and bias vectors $\mathbf{b_i^{(j)}}$, for each layer $j$, and admit $f$ as activation function. The entire architecture is then defined by:

$$\mathbf{h_i^{(1)}} = f(\mathbf{W_i^{(1)}} \times \mathbf{x_i} + \mathbf{b_i^{(1)}}) \qquad i = 1, 2 \tag{1}$$

$$\mathbf{h_1^{(2)}} = f(\mathbf{W^{(2)}} \times \mathbf{h_1^{(1)}} + \mathbf{b_1^{(2)}}) \tag{2}$$

$$\mathbf{h_1^{(3)}} = f(\mathbf{W^{(3)}} \times \mathbf{h_1^{(2)}} + \mathbf{b_1^{(3)}}) \tag{3}$$

$$\mathbf{h_2^{(2)}} = f(\mathbf{W^{(3)^T}} \times \mathbf{h_2^{(1)}} + \mathbf{b_2^{(2)}}) \tag{4}$$

$$\mathbf{h_2^{(3)}} = f(\mathbf{W^{(2)^T}} \times \mathbf{h_2^{(2)}} + \mathbf{b_2^{(3)}}) \quad (5)$$

$$\mathbf{o_i} = f(\mathbf{W_i^{(4)}} \times \mathbf{h_i^{(3)}} + \mathbf{b_i^{(4)}}) \qquad i = 1, 2 \quad (6)$$

It is important to note that the weight matrices $\mathbf{W^{(2)}}$ and $\mathbf{W^{(3)}}$ are used twice due to weight tying, respectively in Eqs. 2, 5 and Eqs. 3, 5. Training is performed by applying batch gradient descent to minimize the mean squared error of $(\mathbf{o_1}, \mathbf{x_2})$ and $(\mathbf{o_2}, \mathbf{x_1})$ thus effectively minimizing the reconstruction error in both directions and creating a joint representation in the middle.

Multimodal embeddings are obtained in the following manner:

- When the two modalities are available (automatic transcripts and CNN features), both are presented at their respective inputs and the activations are propagated through the network. The multimodal embedding is then obtained by concatenating the outputs of the middle layer.
- When one modality is available and the other is not (either only transcripts or only visual information), the available modality is presented to its respective input of the network and the activations are propagated. The central layer is then used to generate an embedding by being duplicated, thus still generating an embedding of the same size while allowing to transparently compare video segments regardless of modality availability (either with only one or both modalities).

Finally, segments are then compared as illustrated in Figure 3 b): for each video segment, the two modalities are taken (embedded automatic transcripts with embedded CNN representations) and a multimodal embedding is created with a bidirectional deep neural network. The two multimodal embeddings are then simply compared with a cosine distance to obtain a similarity measure. An implementation of bidirectional deep neural networks is available publicly and free to use[2].

For this run, we used aggregated CNN features of size 4096 as the first modality and aggregated *Word2Vec* features of size 100 as the second modality. The hidden representation layers were of size 1024, thus yielding a multimodal embedding of size 2048. We trained the network for 10000 epochs using stochastic gradient descent (SGD) with Nesterov momentum, using a learning rate of 0.2 and momentum of 0.9. Also, dropout of 20% was used to improve generalization.

# 5 Monomodal Runs

Our two monomodal runs were based on each modality learned for the cross-modal system described in Section 4. While late fusion of both modalities is expected to perform significantly better, submitting runs based on the original continuous representation may provide significant insight on the difference of the new representation space compared to the two original ones. For these two

[2]https://github.com/v-v/BiDNN

| | Crossmodal | Audio | Visual | Multimodal |
|---|---|---|---|---|
| prec@5 | **0.52** | 0.40 | 0.45 | 0.34 |

Table 1: Precision at rank 5 for crossmodal, monomodal audio, monomodal visual and multimodal runs.

| | Min | 25% | Median | 75% | Max |
|---|---|---|---|---|---|
| prec@5 | 0.24 | 0.32 | 0.35 | 0.41 | **0.52** |

Table 2: Minimum, median, maximum and first and third quartiles scores among participants to the Hyperlinking task.

runs, the original representations are the same as described in Section 4, except that only one modality is used (either aggregated CNN features with a size of 4096 or aggregated *Word2Vec* features with size 100) for obtaining a similarity measure, still with cosine distance.

# 6 Results

The evaluation consists in a manual annotation by Turkers (AMT) of the relevance of the top 5 targets for each anchor. We report the mean precision at rank 5 for our four runs table 1. Table 2 shows the performance of other teams for the hyperlinking task.

Our crossmodal run obtained the best score among participants, with a precision at rank 5 of 0.5244. This demonstrates the interest of crossmodal systems in the hyperlinking setup. More specifically, the crossmodal run performed better than the two monomodal runs.

The visual monomodal run performed a lot better than the audio monomodal run, demonstrating a heavy bias towards the importance of visual information, contrary to last year's dataset where audio similarity was favored. This bias explains in part the low score of the multimodal for which the audio weighted more than the visual information. This emphasizes the benefit of using crossmodality instead of multimodality as a mean to lower the incidence of manually set weights that can differ widely from one dataset to another. Another explanation for the lower results of the multimodal run is the probable absence of near-duplicates that were largely found in last year's dataset, damaging approaches that rely on shallow similarity. This is for the best as near-duplicates are of very little interest to the users.

# 7 Conclusion

This year's experiments showed that crossmodal systems perform particularly well, as our crossmodal run ranked first among participants in terms of precision at 5. Crossmodality also outperforms similar monomodal models. Results also seem to indicate that the visual modality was favored during anchor selection,

or that assessors were more interested in what is shown than what is said, as opposed to last year's dataset. Lastly, there seems to be far less near-duplicates in this year's task, probably due to less redundancy in the dataset, or because selected anchors appeared only once in the dataset. This setup is beneficial as it allows the discovery of more interesting targets, hopefully increasing diversity and serendipity.

# References


[1] Jon Fiscus Martial Michel David Joy Alan F. Smeaton Wessel Kraaij Georges Quenot Roeland Ordelman Robin Aly Paul Over, George Awad. Trecvid 2015: An overview of the goals, tasks, data, evaluation mechanisms, and metrics. In *TRECVID 2015*, page 52. http://www-nlpir.nist.gov, 2015.

[2] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, and Roeland Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[3] Abigail McBirnie. Seeking serendipity: the paradox of control. *Aslib Proceedings*, 60(6):600–618, 2008.

[4] Allen Foster and Nigel Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[5] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[6] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[8] Vedran Vukotic, Christian Raymond, and Guillaume Gravier. Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications. In *Proceedings of ACM International Conference in Multimedia Retrieval (ICMR)*, New York, United States, June 2016. ACM.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[10] Vedran Vukotic, Christian Raymond, and Guillaume Gravier. Multimodal and Crossmodal Representation Learning from Textual and Visual Features

with Bidirectional Deep Neural Networks for Video Hyperlinking. In *ACM Multimedia 2016 Workshop: Vision and Language Integration Meets Multimedia Fusion (iV&L-MM'16)*, Amsterdam, Netherlands, October 2016. ACM Multimedia.

[11] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.