



**HAL**  
open science

# A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking

Vedran Vukotić, Christian Raymond, Guillaume Gravier

► **To cite this version:**

Vedran Vukotić, Christian Raymond, Guillaume Gravier. A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking. IEEE MultiMedia, 2018, 25 (2), pp.11-23. 10.1109/MMUL.2018.023121161 . hal-01848539v2


**HAL Id: hal-01848539**

**<https://inria.hal.science/hal-01848539v2>**

Submitted on 20 Aug 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



“THEME ARTICLE”, “FEATURE ARTICLE”, or “COLUMN” goes here: The theme topic or column/departmentname goes after the colon.

# A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking

**Vedran Vukotić**  
INSA Rennes, INRIA/IRISA

**Christian Raymond**  
INSA Rennes, INRIA/IRISA

**Guillaume Gravier**  
CNRS, INRIA/IRISA

Multimodal representations are typically obtained with autoencoders by focusing on reconstruction of multimodal data. We propose an alternative possible approach that focuses on crossmodal translations

between initially disjoint modalities to perform multimodal fusion in a new, common, representation space and evaluate it in multimodal retrieval and video hyperlinking tasks.

With the recent resurgence of neural networks and the proliferation of massive amounts of multimodal unlabeled data, recommendation systems and multimodal retrieval systems based on continuous representation spaces and deep learning methods are becoming of great interest. In this work, we present a method to perform high-level multimodal fusion by focusing on crossmodal translation by means of symmetrical encoders cast into a bidirectional deep neural network (BiDNN).

We analyze different continuous single-modal representations and evaluate BiDNNs in a multimodal retrieval setup. Using the notions learnt from multimodal retrieval we craft a system based on BiDNNs to perform video hyperlinking and recommend interesting video segments to a viewer. Results established within the TRECVID's 2016 video hyperlinking benchmarking initiative show that our method obtained the best score, thus defining the state of the art.

## INTRODUCTION

Dealing with data originating from multiple modalities often requires to either combine them, fuse them into a joint multimodal representation or translate from one modality into another. While combining initially disjoint modalities, e.g., late score fusion by combining the scores obtained with each modality or early feature fusion by concatenating features from the different modalities, is the simplest approach, results obtained in such a manner are usually underperforming. State-of-the-art results are today typically obtained by defining a new representation space that fuses the initially disjoint modalities, incorporating a multimodal autoencoder in the pipeline.<sup>1</sup> In this setting, multimodal autoencoders focus on reconstructing the initially disjoint modalities through a common representation space of lower dimension. In addition, they typically increase robustness by adding noise to their inputs and by learning to reconstruct the various modalities even one is zeroed out. Instead of focusing on reconstruction of initially disjoint modalities, it is possible to focus on the task of crossmodal translation.<sup>2</sup> In this case, one learns mappings from one initial modality to another and vice versa. The importance of focusing on crossmodal translation rather than modality reconstruction, appears to be the key to the improvement in multimodal fusion. By creating translations between modalities, it becomes easier to tweak the system in such a manner that a common multimodal representation space is created where all the initially disjoint modalities are projected and later fused.

When dealing with a video collection, there are two main possible tasks: retrieval and hyperlinking. The task of video retrieval aims at retrieving similar videos or video segments given a query, the latter also being a video or video segment. In this scenario, all the videos or video segments are predefined and stored in a database. Contrary to retrieval, video hyperlinking does not start with a collection of video segments stored in a database but rather with full-length videos in which segments of interest must be found. The task thus requires a segmentation step to create video segments that differ from their neighbouring segments in either one of the modalities before casting hyperlinking into a retrieval task. While video retrieval can easily be assessed by evaluating retrieval results against groundtruth relevance annotation, video hyperlinking evaluations are typically done by human evaluators who assess the importance of a proposed hyperlink and state how likely they would be to follow such a hyperlink if they were watching the source video of the hyperlink.

## CONTENT REPRESENTATION

Multimodal feature fusion necessarily builds in singlemodal representations that are to be combined. We thus start by analyzing different single-modal representations for both the visual and speech modalities. After choosing the best-performing ones, we progress to methods for either combining them or fusing them into a joint multimodal representation. We do so by firstly introducing multimodal autoencoders that focus on multimodal reconstruction and that we use as a baseline, and secondly by introducing our proposed BiDNN architecture that focuses on crossmodal translation

### Initial Single-modal Representations

In this work, two modalities are used: i) automatic transcripts of the speech contained in the audio track and ii) video keyframes. We do not utilize any human-generated information available in the datasets (e.g., metadata, subtitles, etc.) in order to keep our systems fully autonomous and without a human-in-the-loop element in the pipeline. Recent results show that metadata information is beneficial, however, used as a filtering step to filter out non-relevant matches.<sup>3-4</sup> Regarding speech transcripts, we evaluate two different representations of texts in a continuous representation space: paragraph vectors and Word2Vec. Paragraph vectors provide directly a representation of textual segments. Contrary to paragraph vectors, Word2Vec is used to embed single words and thus it is necessary to aggregate the vectors of each word within the speech segment into one representation. We classically perform aggregation by taking the average of the vectors over each words.<sup>5</sup>

For the visual modality, we rely on keyframes representation, considering two approaches. A first possibility consists in describing a keyframe with visual concepts which are further embedded and aggregated into a continuous representation space. The alternative is to directly embed the image into a continuous representation space. We used the ImageNet concepts to describe keyframes. When directly embedding video keyframes into a continuous representation space, we again use deep convolutional neural networks that have been shown to perform well in a multitude of computer vision tasks.<sup>6</sup> More precisely, we use a less deep convolutional neural network, namely AlexNet and two very deep convolutional deep neural networks, namely VGG-16 and VGG-19.<sup>7-8</sup> Aggregation, both for visual concepts and multiple keyframes, is performed by averaging.

## Multimodal and Crossmodal Approaches

Multiple modalities can be used without actually fusing them. Two very common methods of using multiple modalities are concatenation, of the representations and score fusion.<sup>9-10</sup> When performing score fusion, each modality is processed separately, yielding a classification or decision score for each, and the final score is computed by weighting the scores obtained with each modality.

### Multimodal Autoencoders

Autoencoders are now widely used for multimodal fusion, with approaches that can be broken down into two main families: i) concatenating the different modalities and utilizing a standard autoencoder or ii) keeping the modalities separated and presenting them to a multimodal autoencoder with a modified architecture that contains separate input and output branches for each modality, as illustrated in Figure 1 in the left part. This last architecture also allows the autoencoder to provide a better crossmodal translation if necessary.<sup>1</sup> In both cases, the central layer, typically of low dimension, is used to obtain a multimodal representation of the input modalities. To ensure robustness, noise is often added to the inputs and one modality can be sporadically zeroed while expecting a complete reconstruction of both modalities. We, however, believe autoencoders to have a few downsides:

- Whatever the family of approaches considered (concatenated or separate input), all modalities are mixed and must be present: they all influence the central layer, even when zeroed out.
- Autoencoders have to produce the same output when both modalities are presented to their input and when one modality is zeroed out. These two tasks might not necessarily point towards the same local optimum and might be detrimental for training.
- Autoencoders can perform crossmodal translation by taking one modality and a zero vector at their inputs and reconstructing both modalities, including the missing one, at their output. This is however not as optimal as a direct crossmodal translation.

These potential downsides were tackled by introducing bidirectional (symmetrical) deep neural networks,<sup>11</sup> which we discuss next.

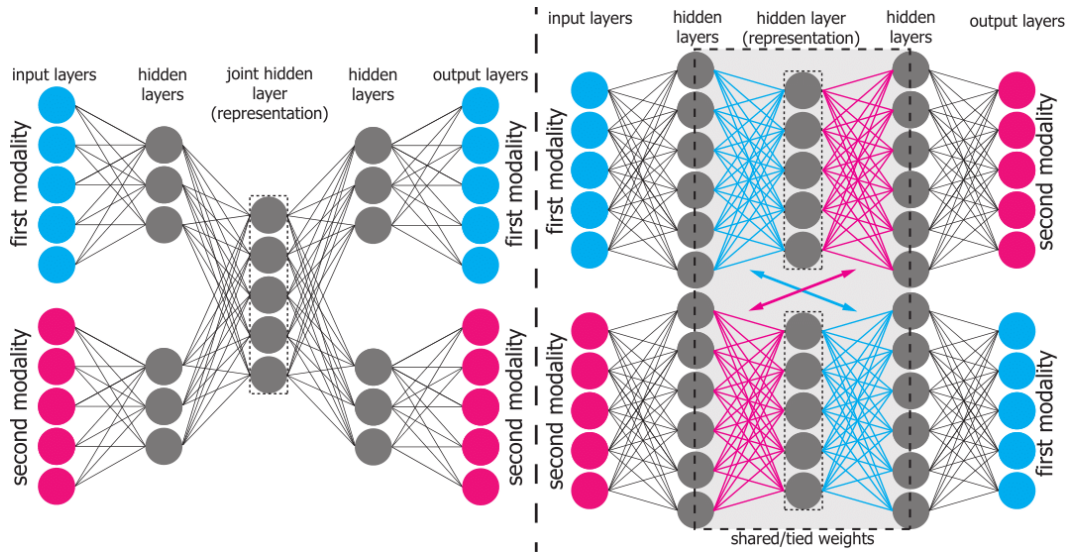


Figure 1. Illustration of the architectures: left - classical multimodal autoencoder; right - our proposed bidirectional symmetrical deep neural network.

## Bidirectional Deep Neural Networks

BiDNNs, contrary to multimodal autoencoders, focus on crossmodal translation. The key idea of the architecture relies on the use of two deep neural networks, working in opposite directions: one for translating from the first modality to the second and one for translating from the second modality to the first one. Each network thus performs a crossmodal translation. To create a representation space in the middle that is as similar as possible for the two crossmodal translations, symmetry is enforced in the central layers. Implementation-wise, this is done by sharing the same variables over both networks, as illustrated in Figure 1. One weight matrix from one network is the same weight matrix on the other network, only transposed. Symmetry is implemented solely in the central part as fully symmetrical networks would be too rigid for real, imperfect data which would negatively affect the architecture's ability to converge. The multimodal representation is formed by taking the activations of the central layers and concatenating them. Projecting the initially disjoint single-modal representations to their new representation spaces (formed by learning two crossmodal translations) brings them closer. The two representation spaces are now very similar (although not exactly the same due to imperfections in the data and training) and similarities can be computed between the initially disjoint modalities. When both modalities are present, the two new representations are concatenated to form a new multimodal representation. This allows the same modalities of two different datapoints to be compared in same representation space. If one modality is missing, it is again projected to its new representation space, formed by the corresponding crossmodal translation, but is now duplicated to form a new multimodal representation of the same size. Since the two representation spaces, formed by the two crossmodal translations, are very close, different modalities can be now compared. This allows for one modality of a datapoint to be compared against both modalities of another datapoint.

Formally, let  $h_i^{(j)}$  denote a hidden layer at depth  $j$  in network  $i$  ( $i = 1, 2$ ; one for each modality),  $x_i$  the feature vector for modality  $i$  and  $o_i$  the output of the network for modality  $i$ . In turn, for each network,  $W_i^{(j)}$  denotes the weight matrix and  $b_i^{(j)}$  the bias vector of layer  $j$ . Finally, we assume that each layer admits  $f$  as an activation function. The architecture is then defined by:

$$\begin{aligned} h_1^{(1)} &= f(W_1^{(1)} \times x_1 + b_1^{(1)}) \\ h_2^{(1)} &= f(W_2^{(1)} \times x_2 + b_2^{(1)}) \end{aligned}$$

$$\begin{aligned} \mathbf{h}_1^{(2)} &= f(\mathbf{W}^{(2)} \times \mathbf{h}_1^{(1)} + \mathbf{b}_1^{(2)}) \\ \mathbf{h}_2^{(2)} &= f(\mathbf{W}^{(3)T} \times \mathbf{h}_2^{(1)} + \mathbf{b}_2^{(2)}) \end{aligned}$$

$$\begin{aligned} \mathbf{h}_1^{(3)} &= f(\mathbf{W}^{(3)} \times \mathbf{h}_1^{(2)} + \mathbf{b}_1^{(3)}) \\ \mathbf{h}_2^{(3)} &= f(\mathbf{W}^{(2)T} \times \mathbf{h}_2^{(2)} + \mathbf{b}_2^{(3)}) \end{aligned}$$

$$\begin{aligned} \mathbf{o}_1 &= f(\mathbf{W}_1^{(4)} \times \mathbf{h}_1^{(3)} + \mathbf{b}_1^{(4)}) \\ \mathbf{o}_2 &= f(\mathbf{W}_2^{(4)} \times \mathbf{h}_2^{(3)} + \mathbf{b}_2^{(4)}) \end{aligned}$$

In the above equations, the weight matrices  $\mathbf{W}^{(2)}$  and  $\mathbf{W}^{(3)}$  are used twice due to weight tying, for computing  $\mathbf{h}_1^{(2)}$ ,  $\mathbf{h}_2^{(3)}$  and  $\mathbf{h}_2^{(2)}$ ,  $\mathbf{h}_1^{(3)}$  respectively. Training is performed 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, where both representations can be projected.

Crossmodal translation is performed by presenting a single modality to its respective input  $\mathbf{x}_i$  and generating the output  $\mathbf{o}_i$  of the appropriate network, that represents the projection of the given modality into the representation space of the other modality. Multimodal fusion is performed by presenting one or both modalities,  $(\mathbf{x}_1$  and/or  $\mathbf{x}_2)$ , to their respective inputs and by taking the activation outputs of the central layers  $\mathbf{h}_1^{(2)}$  and/or  $\mathbf{h}_2^{(2)}$ . If the two modalities are available, they are both presented to their respective inputs and the activations of the two central layers are taken and concatenated to form a fused multimodal representation. In the case where only one modality is available, it is presented to its respective input and the activation of the corresponding central layer is taken and replicated to replace the missing central layer activations. This is made possible by the symmetrical nature of the BiDNN architecture and allows to have representations of the same dimension when all modalities are present or when one is missing.

Given the multimodal embedding defined by the BiDNN architecture, the similarity of two video segments is obtained using a cosine similarity on the embedded representation of each segment as defined by the respective activations at the central layer, as illustrated in Figure 2.

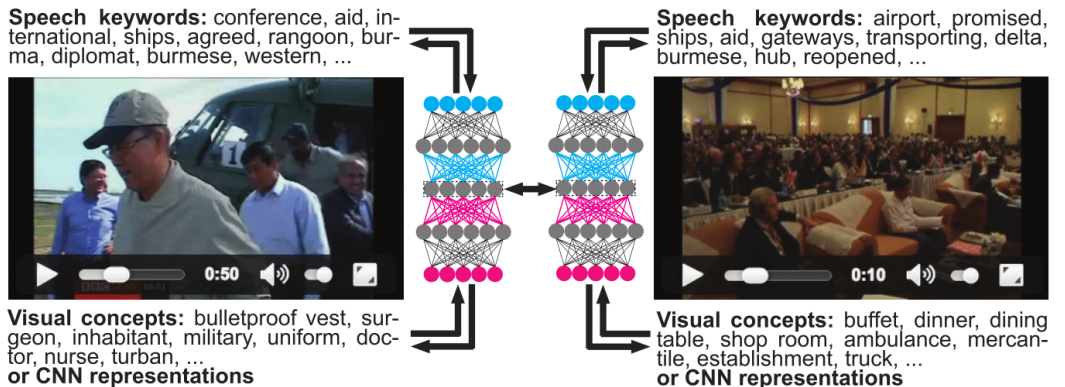


Figure 2. With BiDNNs, the similarity of two video segments is obtained by presenting the respective modalities of each of the two samples to a BiDNN and taking the activations of the respective central layers after propagating the information. Afterward, the obtained representations are used to compute a cosine similarity measure.



## EXPERIMENTS

In this section, we start by describing the datasets used for evaluation. We then proceed to evaluate different single-modal representations, as described theoretically in the previous section. After choosing the best performing single-modal evaluations, we evaluate methods for performing fusion, first preliminarily in multimodal retrieval and then we use the learned notions and evaluate the best methods in video hyperlinking.

### Datasets

#### Video Retrieval

The methods discussed in this work were first evaluated in a multimodal video retrieval setup, based on the groundtruth that was formed after the MediaEval 2014 video hyperlinking task by collecting all the video segments proposed by the participants and their judged relevance.<sup>12</sup> The dataset exploits broadcast videos provided by the BBC and consists of 30 anchors (acting as queries) and 10,809 targets. There are about 34.3 keyframes on each video segment on average. The groundtruth consists of 12,340 anchor-target pairs with the corresponding relevance judgment, i.e., either related or not. Among these video segments, not all of them contain the two modalities.

Although multiple modalities available, each with possible variations, we used two main ones: for speech, we use automatic transcripts embedded in different ways, while for video keyframes, we either use embedded visual concepts (ImageNet classes) provided by KU Leuven,<sup>13</sup> or CNN features we computer ourselves.

#### Video Hyperlinking

Video hyperlinking or, more specifically, the generation of hyperlinks within a collection of video segments is evaluated within the “Search and Hyperlinking” benchmarking initiative, first at MediaEval,<sup>12</sup> and more recently at TRECVID.<sup>14</sup> Our proposed methods were evaluated in a video hyperlinking setup by participating to TRECVID’s 2016 video hyperlinking benchmarking initiative. Contrary to the previous dataset that contained videos provided by BBC, this dataset is formed with videos from the BlipTV video sharing platform and contains different usersubmitted videos with different topics, styles, and languages. The dataset contained 14,838 videos an average duration of 13 minutes. We used automatic speech transcripts provided by LIMSI,<sup>15</sup> and video keyframes provided with the dataset.

In the case of video hyperlinking, we perform segmentation and we recommend relevant video segments out of the multitude of obtained segments, where the big majority of them is irrelevant. Video segmentation must result in a set of 10 to 120 seconds long potential targets, given the limitations imposed by the benchmark organizers. We chose to segment videos by taking only 30 seconds of continuous speech and cut at the following speech pause, as detected by the speech transcription system. Additionally, we run this segmentation another time, using an offset of one speech segment at the second pass, in order to obtain an overlapping segmentation. This results in 307,403 video segments with an average duration of 45 seconds.

Table 1: Results of the evaluated methods over five runs with their respective precision at 10 (%) and standard deviation.

| Modalities                          | Aggregation method | P@10 (%) | $\sigma$ (%) |
|-------------------------------------|--------------------|----------|--------------|
| Single-modal speech representations |                    |          |              |
| Word2Vec                            | average            | 58.67    | -            |
| PV-DM                               | -                  | 45.00    | -            |

|                                     |                    |              |             |
|-------------------------------------|--------------------|--------------|-------------|
| PV-DBOW                             | -                  | 41.67        | -           |
| Single-modal visual representations |                    |              |             |
| KU Leuven c., W2V                   | average            | 50.00        | -           |
| KU Leuven c., PV-DM                 | -                  | 45.33        | -           |
| KU Leuven c., PV-DBOW               | -                  | 48.33        | -           |
| AlexNet                             | average            | 63.00        | -           |
| VGG-16                              | average            | 70.67        | -           |
| VGG-19                              | average            | 68.67        | -           |
| Simple multimodal approaches        |                    |              |             |
| Transcripts, visual concepts        | concatenation      | 58.00        | -           |
| Transcripts, AlexNet                | concatenation      | 70.00        | -           |
| Transcripts, VGG-16                 | concatenation      | 75.33        | -           |
| Transcripts, VGG-19                 | concatenation      | 74.33        | -           |
| Transcripts, visual concepts        | linear combination | 61.32        | 3.10        |
| Transcripts, AlexNet                | linear combination | 67.38        | 2.66        |
| Transcripts, VGG-16                 | linear combination | 71.86        | 4.11        |
| Transcripts, VGG-19                 | linear combination | 71.78        | 3.90        |
| Multimodal autoencoders             |                    |              |             |
| Transcripts, visual concepts        |                    | 59.60        | 0.65        |
| Transcripts, AlexNet                |                    | 69.87        | 1.64        |
| Transcripts, VGG-16                 |                    | 74.53        | 1.52        |
| Transcripts, VGG-19                 |                    | 75.73        | 1.79        |
| BiDNN single modality embedding     |                    |              |             |
| Transcripts                         |                    | 66.78        | 1.05        |
| Visual concepts                     |                    | 54.92        | 0.99        |
| AlexNet                             |                    | 66.33        | 0.58        |
| VGG-16                              |                    | 68.70        | 1.98        |
| VGG-19                              |                    | 70.81        | 1.08        |
| BiDNN multimodal embedding          |                    |              |             |
| Transcripts, visual concepts        |                    | 73.74        | 0.46        |
| Transcripts, AlexNet                |                    | 73.41        | 1.08        |
| Transcripts, VGG-16                 |                    | 76.33        | 1.60        |
| <i>Transcripts, VGG-19</i>          |                    | <i>80.00</i> | <i>0.80</i> |
| BiDNN query expansion               |                    |              |             |



|                              |       |      |
|------------------------------|-------|------|
| Transcripts, visual concepts | 62.35 | 0.25 |
| Transcripts, AlexNet         | 70.11 | 1.25 |
| Transcripts, VGG-16          | 75.33 | 0.10 |
| Transcripts, VGG-19          | 74.33 | 0.10 |

## Initial Representations

Speech is represented either with paragraph vectors or averaged Word2Vec<sup>5</sup> — in both cases with a representation of size 100. For Word2Vec we used a skip-gram model with hierarchical sampling, as this work best. Visual concepts were treated in the same manner, except for being sorted before being processed with Word2Vec. All the speech models were trained in an unsupervised manner on the dataset.

For representing video keyframes, we used the output of the last fully connected layer of a convolutional neural network, thus yielding a representation of size 4096. We analyze 3 different CNN architectures, namely AlexNet, VGG-16 and VGG-19, in an implementation called KerasConvnets — a CNN framework based on Keras, offering models already pretrained on ImageNet. Within video segments, representations were obtained for each keyframe and were then averaged.<sup>16</sup> Averaged VGG-16 provide the best visual embedding, yielding a result of 70.67% in precision at 10. The performance of the different methods is shown in Table 1.

## Using Multiple Modalities

Multiple modalities can be used either separately, in their original representation spaces or by fusing them into a new representation. We start by evaluating simple concatenation and score fusion. After that, we progress to multimodal fusion where we first evaluate state-of-the-art multimodal autoencoders, that we use as our baseline and our proposed BiDNN architecture. To have comparable experiments, we use a representation size of 100 for speech and a representation size of 4096 for video keyframes in all experiments. Both multimodal autoencoders and BiDNN were designed to yield a multimodal representation of size 1000. In our experiments and for the given datasets, bigger multimodal representation spaces did not improve the results.

All the weights in the different neural architectures were randomly initialized with an appropriate uniform distribution.<sup>17</sup> Training was performed by stochastic gradient descent (SGD) with Nesterov momentum in mini-batches of 100 samples. For regularization, dropout of 20% was applied. Training was performed for 1000 epochs to ensure convergence was achieved for all systems. No system started to diverge after converging. Every setup was run 5 times, the results were averaged and the respective standard deviation computed, as shown in Table 1.

### Simple Multimodal Approaches

It is possible to utilize multimodal data without fusing the initially disjoint modalities. Two typical methods to do so are concatenation and score fusion.<sup>10</sup> Although these methods are typically underperforming, we still include them for the sake of completeness and to obtain a better overall picture of the improvement brought by multimodal autoencoders and BiDNNs. Although there is no significant improvement when concatenating speech transcripts and visual concepts, combining VGG-16 embeddings with embedded transcripts yields 75.33% (precision at 10) over the initial performance of 70.67% and 58.67% respectively.

### Multimodal Embedding with Autoencoders

Multimedia autoencoders are a state-of-the-art method for performing multimodal fusion. Although a simple autoencoder can be used for multimodal fusion by concatenating the different

modalities, we implemented a multimodal autoencoder with separate branches for each modality, as illustrated in Figure 1 in the left part. Simple autoencoders performed worse than multimodal autoencoders and were thus not used as a baseline.<sup>11</sup>

The results are illustrated in Table 1. Multimodal embedding clearly performs better than every single modality separately. Combining embedded transcripts and VGG-19 features yield 74.73%, compared to 58.67% and 68.07% respectively. In some cases, multimodal embedding did not improve the results significantly. We believe this to be the case for already good initial single-modal representations given the fact that autoencoders have to train to represent the correct output with both modalities being present at their input and with one zeroed modality. In cases where the initial embeddings perform less good (e.g., embedded visual concepts combined with embedded transcripts), autoencoders seem to improve in a more significant way.

## BiDNN Multimodal Embedding

To have comparable results, we performed all the BiDNN experiments with an architecture of comparable dimensionality and a number of layers as the architectures of multimodal autoencoders previously evaluated: the input branches were either of 100 or 4096 dimensional and the central representation layer was of size 1000. Training was done for 1000 epochs though convergence was again achieved earlier. Just like before, each model was run for 5 times and the averages with their respective standard deviations were computed and reported in Table 1. When comparing different models, we use this information to determine whether there is a significant improvement by performing a single-tailed t-test. Multimodal fusion created with BiDNNs has the benefit of a common joint representation space where both modalities are projected from their initial representation spaces. This provides multimodal embedding superior to those obtained with multimodal autoencoders and brings significant improvement. For instance, combining embedded transcripts with VGG-19 embeddings yields a precision at 10 of 80.00%, compared to 58.67% and 68.67% respectively. All the other tested combinations also yielded better results and high-quality multimodal embeddings.

## BiDNN Single Modality Embedding

When performing multimodal fusion, both modalities are projected into the new representation space created in the network's central layers. However, it is possible to only project one modality into the new representation space and use it as an improved multimodal representation. The network is still trained in a crossmodal manner but the only a single modality is projected into the newly learned representation space. When doing so, automatic speech transcripts improve from 58.67% to 66.78%, visual concepts from 50.00% to 54.92% and VGG-19 embeddings from 68.67% to 70.81%. Although using only a single modality cannot achieve the same results as multimodal fusion, these experiments show that the newly learned representation space offers better performance and can even improve single-modal representations after they are projected into it.

## BiDNN Crossmodal Query Expansion

BiDNNs are crafted for crossmodal translation in mind, so when a single modality of a datapoint is missing it is possible to easily reconstruct it from the other. By doing so, we expand the queries in such a way that all datapoints have both their modalities present and are compared to both modalities. It is important to note that in case of crossmodal query expansion, we remain into the original representation spaces of each modality and we only fill in the very few missing modalities. When combining transcripts and visual concepts (originally 58.00%), we obtain, e.g., 62.35%. No significant improvement is obtained for pairs computed with highperforming deep convolutional neural networks and automatic transcripts. We believe this to be caused by the fact that only a few datapoints are missing one modality, so by filling in the missing modalities and staying into the same representation spaces, we do not see an improvement if the original representation spaces are already well performing

## Video Hyperlinking Evaluation at TRECVID 2016

To further analyze BiDNNs and evaluate whether they also work in a video hyperlinking setup, we decided to evaluate them through the TRECVID’s 2016 video hyperlinking benchmark.<sup>18</sup> In this case, as explained before, we also perform segmentation and we recommend video segments to a human evaluator out of the multitude of mostly irrelevant video segments. To see not only how well the proposed system performs in regards to the systems proposed by other participants but also how much does it improve over the initial, disjoint single-modal representations, we decided to evaluate the best performing single-modal representations, according to our previous evaluation in multimodal retrieval, together with the two modalities fused in a crossmodal fashion by BiDNNs. Each submission was made by finding the top 10 most similar video segments, for each given anchor, out of the 307,403 overlapping video segments. In case two proposed segments overlapped, only the most similar one was kept and the other was removed from the stack.

The results of the submitted runs, as well as the statistics of all the team that participated, are given in Table 2. The official scores at TRECVID’s 2016 video hyperlinking task were given in precision at 5. It is important to emphasize that, contrary to some participants, we did not use any additional information available with the dataset and we utilized solely the video keyframes and automatic speech transcripts. Also, many participants did not use deep learning methods. The video hyperlinking evaluation confirmed what we have previously shown in multimodal retrieval evaluations: focusing on crossmodal translations with added restrictions (namely enforced symmetry) yields improved multimodal embeddings. Not only did BiDNN improve the two initially disjoint modalities (40% and 45% into a representation that yields 52% in terms of precision at 5) but it also proved that obtaining multimodal fusion in a crossmodal fashion outperforms multimodal fusion methods. This shows how a simple, unsupervised learning system, with no additional information and no handcrafted features, can compete with more complex multimodal systems. It also proves that crossmodal translations with enforced symmetry improve over classic ways of multimodal fusion and define the new state of the art.

Table 2: Results of our submitted runs containing solely the initial modalities, fused modalities with BiDNN and overall statistics for all the participants of TRECVID’s 2016 video hyperlinking benchmark.

| Evaluation                            | P@5 (%) |
|---------------------------------------|---------|
| Speech transcripts only               | 40      |
| VGG-19 features only                  | 45      |
| <i>BiDNN – transcripts and VGG-19</i> | 52      |
| Min (all teams)                       | 24      |
| 1 <sup>st</sup> quartile (all teams)  | 32      |
| Mean (all teams)                      | 35      |
| 3 <sup>rd</sup> quartile (all teams)  | 41      |
| Max (all teams)                       | 52      |

## Post TRECVID 2016 Evaluation

While we have seen that BiDNNs define the new state of the art in video hyperlinking, it is still interesting to further evaluate the overall system and see how would the different parts impact the evaluation if altered. After the TRECVID 2016 video hyperlinking evaluation was performed, a new groundtruth was again collected by the organizers, containing the targets proposed

by all their participants and their relevance, as judged by AMT human evaluators. These targets are not consistent with our evaluated video segments as every participant had their own segmentation. We thus use our proposed system (multimodal representations of each video segment, fused with BiDNNs) to analyze the similarity of the targets proposed by all the participants for each anchor and the rank (determined by similarity) of the targets we proposed. On average, there are 79.19 proposed targets for each anchor.

Figure 3 illustrates the distribution of our proposed targets between all the targets that were proposed for a specific anchor by all the participants, according to their similarity given by multimodal embeddings obtained with a BiDNN model. While most of the targets are within the top 10 by similarity to the given anchor, we proposed many targets that are quite low ranking according to our own system. The average rank for our proposed targets is 27.98 and furthermore, 61.17% of our proposed targets are not within the top 10 video segments when evaluating all the video segments proposed by all participating teams. According to BiDNN similarity, within the top 10 best targets for any given anchor, there are on average only 3.02 targets that we proposed. This indicates that, had we evaluated the same video segments as some other participants, the system could further be improved, at least in regards to similarity between targets and anchors given by BiDNN based representations.

Additionally, by performing a multimodal retrieval analysis on all the proposed targets (without retraining the model) and reranking the targets proposed by all the participants, we obtain a result of 49.56% in precision at 5, compared to 30.8% for the video hyperlinking evaluation, which strongly indicates that although multimodal retrieval evaluations are acceptable for choosing the best performing method, they largely differ from video hyperlinking evaluations in terms of performance.

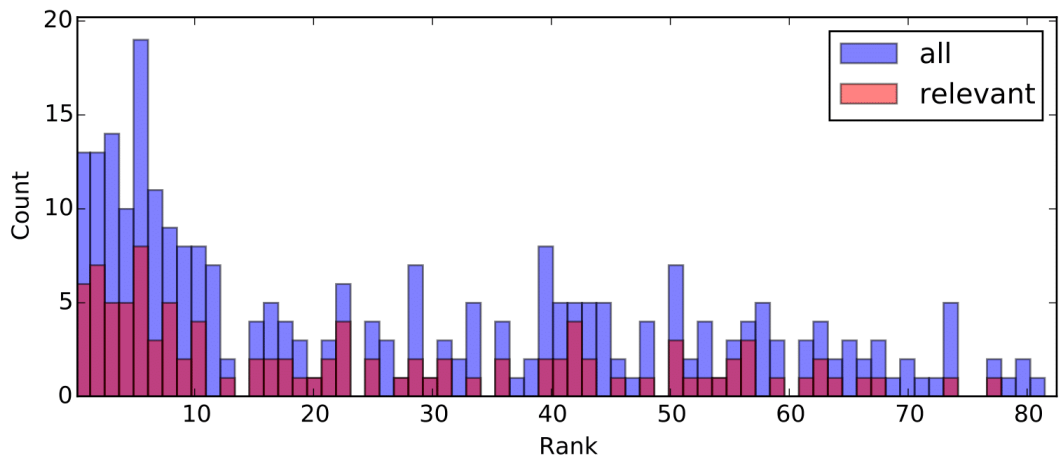


Figure 3. Histogram illustrating the distribution of ranks of our proposed targets in the collection of targets proposed by all the participants, ordered by similarity according to BiDNN fused representations. The blue histogram represents all our proposed targets while the one in red solely those that we proposed and were judged relevant.

## CONCLUSIONS

We started by analyzing different methods for obtaining single-modal continuous representations and methods to fuse them into a single multimodal representation. After evaluating them in a multimodal retrieval setup, we use the learned notions and the best performing methods in video hyperlinking, where we perform a more exhaustive evaluation. Expectedly, visual embeddings obtained with deep convolutional neural networks outperformed embedded visual concepts and proved to be more relevant than automatic transcripts. Very deep VGG convolutional neural network architectures significantly outperformed the less deep AlexNet architecture. VGG-16 performed best and produced a single-modal visual embedding that yields 70.68% in precision at

10. We have shown that the few downsides of autoencoders can affect their results and that BiDNNs successfully tackle these problems and clearly outperform multimodal autoencoders by a significant margin. Although VGG-16 performed better than VGG-19 in a single-modal setup, the best performance was obtained by multimodal fusion of embedded automatic transcripts and embedded VGG-19 features, yielding a precision at 10 of 80.00%.

The video hyperlinking evaluation confirmed that focusing on crossmodal translations with added restrictions, namely in the form of symmetry, is a feasible method to perform multimodal fusion and BiDNNs defined the new state of the art by achieving the best performance at the challenge. After the post-evaluation groundtruth was formed, we used it to further analyze our model and found that, although our submissions were most often ranked between the top 10 ones, multimodal representations obtained with BiDNNs often favored other targets proposed by other participants. This indicates that further improvements can be made solely by changing the initial segmentation. BiDNNs have already been shown to successfully tackle the downsides of multimodal autoencoders and to provide superior multimodal embeddings. In this work, we extensively tested BiDNNs first in a preliminary evaluation in multimodal retrieval and then more extensively in video hyperlinking. This reinforces the points already made in preliminary research,<sup>11</sup> and shows that BiDNNs defined the new state of the art at the last TRECVID's 2016 video hyperlinking benchmarking initiative.

---

## REFERENCES

1. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning" in Intl. Conf. on Machine Learning, 2011.
2. F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder" in ACM Intl. Conf. on Multimedia, 2014, pp. 7–16.
3. G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quenot et al., "Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking" in Proceedings of TRECVID, 2017.
4. M. Demirdelen, M. Budnik, G. Sargent, R. Bois, and G. Gravier, "IRISA at TRECVID 2017: Beyond crossmodal and multimodal models for video hyperlinking" in Working Notes of the TRECVID 2017 Workshop, 2017.
5. M. Campr and K. Jezek, "Comparing semantic models for evaluating automatic document summarization" in Text, Speech, and Dialogue, 2015.
6. A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks" in Advances in neural information processing systems, 2012, pp. 1097–1105.
8. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition" 2014.
9. T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication" Proceedings of the IEEE, vol. 86, no. 5, pp. 837–852, 1998.
10. C. Guinaudeau, A. R. Simon, G. Gravier, and P. Sebillot, "HITS and IRISA at MediaEval 2013: Search and hyperlinking task" in Working Notes MediaEval Workshop, 2013.

11. V. Vukotić, C. Raymond, and G. Gravier, "Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications" in Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, 2016, pp. 343–346.
12. M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. Jones, "The search and hyperlinking task at MediaEval 2014" in Working Notes MediaEval Workshop, 2014.
13. K. Chatfield and A. Zisserman, "Visor: Towards on-the-fly largescale object category retrieval" in Asian Conference on Computer Vision. Springer, 2012, pp. 432–446.
14. P. Over, G. Awad, J. Fiscus, M. Michel, D. Joy, A. Smeaton, W. Kraaij, G. Quenot, R. Ordelman, and R. Aly, "Trecvid 2015 an overview of the goals, tasks, data, evaluation mechanisms, and metrics" in Proceedings of TRECVID, 2015.
15. L. Lamel and J.-L. Gauvain, "Speech processing for audio indexing" in Advances in Natural Language Processing. Springer, 2008, pp. 4–15.
16. V. Vukotić, C. Raymond, and G. 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: ACM Multimedia, Oct. 2016.
17. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in Aistats, vol. 9, 2010, pp. 249–256.
18. G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quenot, M. Eskevich, R. Aly, and R. Ordelman, "Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking" in Proceedings of TRECVID, vol. 2016.

---

## ABOUT THE AUTHORS

**Vedran Vukotić** is a Ph.D. student at IRISA and INSA Rennes. He received his B.Sc. and M.Sc. in computer science in 2012 and 2014 respectively and his M.Sc. in nautical sciences in 2010. His main areas of interests are unsupervised methods of obtaining multimodal representations with deep learning architectures and in particular multimodal fusion of speech and vision in video hyperlinking.

**Christian Raymond** received the Ph.D. degree in 2005 in computer science, from the University of Avignon, France. He was appointed in 2009 associate professor at INSA Rennes (France). He's member of LinkMedia team, devoted to multimedia document analysis, at the IRISA research unit. His research activities focus mainly on speech understanding and machine learning for natural language processing.

**Guillaume Gravier** is senior research scientist at CNRS, France. He leads the Linkmedia research group focusing on content-based media analysis, indexing and linking with the ultimate goal of enabling better multimedia applications and new innovative services. His current research interests are in media analytics, multimedia collection modeling, deep learning and multimodality, graph-based methods for multimedia content representation.