# Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency

**Abstract**—Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors. *Modality* refers to the way in which something happens or is experienced and a research problem is characterized as *multimodal* when it includes multiple such modalities. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. *Multimodal machine learning* aims to build models that can process and relate information from multiple modalities. It is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential. Instead of focusing on specific multimodal applications, this paper surveys the recent advances in multimodal machine learning itself and presents them in a common taxonomy. We go beyond the typical early and late fusion categorization and identify broader challenges that are faced by multimodal machine learning, namely: representation, translation, alignment, fusion, and co-learning. This new taxonomy will enable researchers to better understand the state of the field and identify directions for future research.

**Index Terms**—Multimodal, machine learning, introductory, survey.

## 1 Introduction

The world surrounding us involves multiple modalities — we see objects, hear sounds, feel texture, smell odors, and so on. In general terms, a *modality* refers to the way in which something happens or is experienced. Most people associate the word modality with the *sensory modalities* which represent our primary channels of communication and sensation, such as vision or touch. A research problem or dataset is therefore characterized as *multimodal* when it includes multiple such modalities. In this paper we focus primarily, but not exclusively, on three modalities: natural *language* which can be both written or spoken; *visual* signals which are often represented with images or videos; and *vocal* signals which encode sounds and para-verbal information such as prosody and vocal expressions.

In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret and reason about multimodal messages. *Multimodal machine learning* aims to build models that can process and relate information from multiple modalities. From early research on audio-visual speech recognition to the recent explosion of interest in language and vision models, multimodal machine learning is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential.

The research field of Multimodal Machine Learning brings some unique challenges for computational researchers given the heterogeneity of the data. Learning from multimodal sources offers the possibility of capturing correspondences between modalities and gaining an in-depth understanding of natural phenomena. In this paper we identify and explore five core technical challenges (and related sub-challenges) surrounding multimodal machine learning. They are central to the multimodal setting and need to be tackled in order to progress the field. Our taxonomy goes beyond the typical early and late fusion split, and consists of the five following challenges:

1) **Representation** A first fundamental challenge is learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representations. For example, language is often symbolic while audio and visual modalities will be represented as signals.
2) **Translation** A second challenge addresses how to translate (map) data from one modality to another. Not only is the data heterogeneous, but the relationship between modalities is often open-ended or subjective. For example, there exist a number of *correct* ways to describe an image and and one perfect translation may not exist.
3) **Alignment** A third challenge is to identify the direct relations between (sub)elements from two or more different modalities. For example, we may want to align the steps in a recipe to a video showing the dish being made. To tackle this challenge we need to measure similarity between different modalities and deal with possible long-range dependencies and ambiguities.
4) **Fusion** A fourth challenge is to join information from two or more modalities to perform a prediction. For example, for audio-visual speech recognition, the visual description of the lip motion is fused with the speech signal to predict spoken words. The information coming from different modalities may have varying predictive power and noise topology, with possibly missing data in at least one of the modalities.
5) **Co-learning** A fifth challenge is to transfer knowledge between modalities, their representation, and their predictive models. This is exemplified by algorithms of co-training, conceptual grounding, and zero shot learning. Co-learning explores how knowledge learning from one

• *T. Baltrušaitis is with Microsoft Corporation, Cambridge, UK. C. Ahuja and L-P. Morency are with the Language Technologies Institute, at Carnegie Mellon University, Pittsburgh, Pennsylvania*
*E-mail: tbaltrus, cahuja, morency@cs.cmu.edu*

*Manuscript received May 18, 2017*

Table 1: A summary of applications enabled by multimodal machine learning. For each application area we identify the core technical challenges that need to be addressed in order to tackle it.

| | Challenges | | | | |
|---|---|---|---|---|---|
| Applications | Representation | Translation | Alignment | Fusion | Co-learning |
| **Speech recognition** | | | | | |
| Audio-visual speech recognition | ✓ | | ✓ | ✓ | ✓ |
| **Event detection** | | | | | |
| Action classification | ✓ | | | ✓ | ✓ |
| Multimedia event detection | ✓ | | | ✓ | ✓ |
| **Emotion and affect** | | | | | |
| Recognition | ✓ | | ✓ | ✓ | ✓ |
| Synthesis | ✓ | ✓ | | | |
| **Media description** | | | | | |
| Image description | ✓ | ✓ | ✓ | | ✓ |
| Video description | ✓ | ✓ | ✓ | ✓ | ✓ |
| Visual question-answering | ✓ | | ✓ | ✓ | ✓ |
| Media summarization | ✓ | ✓ | | ✓ | |
| **Multimedia retrieval** | | | | | |
| Cross modal retrieval | ✓ | ✓ | ✓ | | ✓ |
| Cross modal hashing | ✓ | | | | ✓ |
| **Multimedia generation** | | | | | |
| (Visual) speech and sound synthesis | ✓ | ✓ | | | |
| Image and scene generation | ✓ | ✓ | | | |

modality can help a computational model trained on a different modality. This challenge is particularly relevant when one of the modalities has limited resources (e.g., annotated data).

For each of these five challenges, we defines taxonomic classes and sub-classes to help structure the recent work in this emerging research field of multimodal machine learning. We start with a discussion of main applications of multimodal machine learning (Section 2) followed by a discussion on the recent developments on all of the five core technical challenges facing multimodal machine learning: representation (Section 3), translation (Section 4), alignment (Section 5), fusion (Section 6), and co-learning (Section 7). We conclude with a discussion in Section 8.

## 2 Applications: A Historical Perspective

Multimodal machine learning enables a wide range of applications: from audio-visual speech recognition to image captioning. In this section we present a brief history of multimodal applications, from its beginnings in audio-visual speech recognition to a recently renewed interest in language and vision applications.

One of the earliest examples of multimodal research is audio-visual speech recognition (AVSR) [251]. It was motivated by the McGurk effect [143] — an interaction between hearing and vision during speech perception. When human subjects heard the syllable /ba-ba/ while watching the lips of a person saying /ga-ga/, they perceived a third sound: /da-da/. These results motivated many researchers from the speech community to extend their approaches with visual information. Given the prominence of hidden Markov models (HMMs) in the speech community at the time [99], it is without surprise that many of the early models for AVSR were based on various HMM extensions [25], [26]. While research into AVSR is not as common these days, it has seen renewed interest from the deep learning community [157].

While the original vision of AVSR was to improve speech recognition performance (e.g., word error rate) in all contexts, the experimental results showed that the main advantage of visual information was when the speech signal was noisy (i.e., low signal-to-noise ratio) [78], [157], [251]. In other words, the captured interactions between modalities were supplementary rather than complementary. The same information was captured in both, improving the robustness of the multimodal models but not improving the speech recognition performance in noiseless scenarios.

A second important category of multimodal applications comes from the field of multimedia content indexing and retrieval [11], [196]. With the advance of personal computers and the internet, the quantity of digitized multimedia content has increased dramatically [2]. While earlier approaches for indexing and searching these multimedia videos were keyword-based [196], new research problems emerged when trying to search the visual and multimodal content directly. This led to new research topics in multimedia content analysis such as automatic shot-boundary detection [128] and video summarization [55]. These research projects were supported by the TrecVid initiative from the National Institute of Standards and Technologies which introduced many high-quality datasets, including the multimedia event detection (MED) tasks started in 2011 [1].

A third category of applications was established in the early 2000s around the emerging field of multimodal interaction with the goal of understanding human multimodal behaviors during social interactions. One of the first landmark datasets collected in this field is the AMI Meeting Corpus which contains more than 100 hours of video recordings of meetings, all fully transcribed and annotated [34]. Another important dataset is the SEMAINE corpus which allowed to study interpersonal dynamics between speakers and listeners [144]. This dataset formed the basis of the first audio-visual emotion challenge (AVEC) organized in 2011 [186]. The fields of emotion recognition and affective computing bloomed in the early 2010s thanks to strong technical advances in automatic face detection, facial landmark detection, and facial expression recognition [48]. The AVEC challenge continued annually afterward with the

later instantiation including healthcare applications such as automatic assessment of depression and anxiety [217]. A great summary of recent progress in multimodal affect recognition was published by D'Mello et al. [52]. Their meta-analysis revealed that a majority of recent work on multimodal affect recognition show improvement when using more than one modality, but this improvement is reduced when recognizing naturally-occurring emotions.

Most recently, a new category of multimodal applications emerged with an emphasis on language and vision: media description. One of the most representative applications is image captioning where the task is to generate a text description of the input image [86]. This is motivated by the ability of such systems to help the visually impaired in their daily tasks [21]. Recently, progress has been made in the inverse task media generation from text [37], [178]. The main challenges media description and generation is evaluation: how to evaluate the quality of the predicted descriptions and media. The task of visual question-answering (VQA) was recently proposed to address some of the evaluation challenges [9] by providing a correct answer.

In order to bring some of the mentioned applications to the real world we need to address a number of technical challenges facing multimodal machine learning. We summarize the relevant technical challenges for the above mentioned application areas in Table 1. One of the most important challenges is multimodal representation, the focus of our next section.

## 3 Multimodal Representations

Representing data in a format that a computational model can work with has always been a challenge in machine learning. Following Bengio et al. [19] we use the term feature and representation interchangeably, with each referring to a vector or tensor representation of an entity, be it an image, audio sample, individual word, or a sentence. A multimodal representation is a representation of data using information from multiple such entities. Representing multiple modalities poses many difficulties: how to combine the data from heterogeneous sources; how to deal with different levels of noise; and how to handle missing data. The ability to represent data in a meaningful way is crucial to multimodal problems, and forms the backbone of any model.

Good representations are important for the performance of machine learning models, as evidenced behind the recent leaps in performance of speech recognition [82] and visual object classification [114] systems. Bengio et al. [19] identify a number of properties for good representations: smoothness, temporal and spatial coherence, sparsity, and natural clustering amongst others. Srivastava and Salakhutdinov [206] identify additional desirable properties for multimodal representations: similarity in the representation space should reflect the similarity of the corresponding concepts, the representation should be easy to obtain even in the absence of some modalities, and finally, it should be possible to fill-in missing modalities given the observed ones.

The development of unimodal representations has been extensively studied [4], [19], [127]. In the past decade there has been a shift from hand-designed for specific applications to data-driven. For example, one of the most popular ways to represent an image in the early 2000s was through a bag of visual words representation of hand designed features, such as the scale invariant feature transform (SIFT) [132]. However, currently most images (or their parts) are represented using descriptions are learned from data using neural architectures such as convolutional neural networks (CNN) [114]. Similarly, in the audio domain, acoustic features such as Mel-frequency cepstral coefficients (MFCC) have been superseded by data-driven deep neural networks in speech recognition [82] and recurrent neural networks for para-linguistic analysis [216]. In natural language processing, the textual features initially relied on counting word occurrences in documents, but have been replaced data-driven word embeddings that exploit the word context [146]. While there has been a huge amount of work on unimodal representation, up until recently most multimodal representations involved simple concatenation of unimodal ones [52], but this has been rapidly changing.

To help understand the breadth of work, we propose two categories of multimodal representation: *joint* and *coordinated*. Joint representations combine the unimodal signals into the same representation space, while coordinated representations process unimodal signals separately, but enforce certain similarity constraints on them to bring them to what we term a coordinated space. An illustration of different multimodal representation types can be seen in Figure 1.

Mathematically, the joint representation is expressed as:

$$\mathbf{x}_m = f(\mathbf{x}_1, \ldots, \mathbf{x}_n), \tag{1}$$

where the multimodal representation $\mathbf{x}_m$ is computed using function $f$ (e.g., a deep neural network, restricted Boltzmann machine, or a recurrent neural network) that relies on unimodal representations $\mathbf{x}_1, \ldots \mathbf{x}_n$. While coordinated representation is as follows:

$$f(\mathbf{x}_1) \sim g(\mathbf{x}_2), \tag{2}$$

where each modality has a corresponding projection function ($f$ and $g$ above) that maps it into a coordinated multimodal space. While the projection into the multimodal space is independent for each modality, but the resulting space is coordinated between them (indicated as $\sim$). Examples of such coordination include minimizing cosine distance [64], maximizing correlation [7], and enforcing a partial order [220] between the resulting spaces.

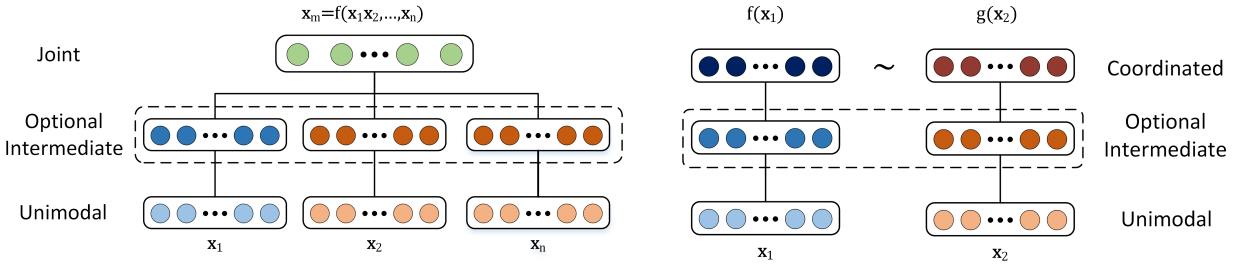
### 3.1 Joint Representations

We start our discussion with joint representations that project unimodal representations together into a multimodal space (Equation 1). Joint representations are mostly (but not exclusively) used in tasks where multimodal data is present both during training and inference steps. The simplest example of a joint representation is a concatenation of individual modality features (also referred to as early fusion [52]). In this section we discuss more advanced methods for creating joint representations starting with neural networks, followed by graphical models and recurrent neural networks (representative works can be seen in Table 2).

**Neural networks** have become a very popular method for unimodal data representation [19]. They are used to represent visual, acoustic, and textual data, and are increasingly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2798607, IEEE Transactions on Pattern Analysis and Machine Intelligence






(a) Joint representation (b) Coordinated representations

Figure 1: Structure of *joint* and *coordinated* representations. Joint representations are projected to the same space using all of the modalities as input. Coordinated representations, on the other hand, exist in their own space, but are coordinated through a similarity (e.g. Euclidean distance) or structure constraint (e.g. partial order).

used in the multimodal domain [157], [163], [225]. In this section we describe how neural networks can be used to construct a joint multimodal representation, how to train them, and what advantages they offer.

In general, neural networks are made up of successive building blocks of inner products followed by non-linear activation functions. In order to use a neural network as a way to represent data, it is first trained to perform a specific task (e.g., recognizing objects in images). Due to the multilayer nature of deep neural networks each successive layer is hypothesized to represent the data in a more abstract way [19], hence it is common to use the final or penultimate neural layers as a form of data representation. To construct a multimodal representation using neural networks each modality starts with several individual neural layers followed by a hidden layer that projects the modalities into a joint space [9], [150], [163], [235]. The joint multimodal representation is then be passed through multiple hidden layers itself or used directly for prediction. Such models can be trained end-to-end — learning both to represent the data and to perform a particular task. This results in a close relationship between multimodal representation learning and multimodal fusion when using neural networks.

As neural networks require a lot of labeled training data, it is common to pre-train such representations using either unsupervised data (e.g., using autoencoder models [12], [83]) or supervised data from a different but related domain [9], [221]. The model proposed by Ngiam et al. [157] extended the idea of using autoencoders to the multimodal domain. They used stacked denoising autoencoders to represent each modality individually and then fused them into a multimodal representation using another autoencoder layer. Similarly, Silberer and Lapata [191] proposed to use a multimodal autoencoder for the task of semantic concept grounding (see Section 7.2). In addition to using a reconstruction loss to train the representation they introduce a term into the loss function that uses the representation to predict object labels.

The major advantage of neural network based joint representations comes from their ability to pre-train from unlabled data when labeled data is not enough for supervised learning. It is also common to fine-tune the resulting representation on a particular task at hand as the representation constructed with unsupervised data is generic and not necessarily optimal for a specific task [225]. One of the disadvantages comes from the model not being able to handle missing data naturally — although there are ways to alleviate this issue [157], [225]. Finally, deep networks are often difficult to train [72], but the field is making progress with new techniques such improved regularization [204], batch normalization [92] and adaptive gradient algorithms [109].

**Probabilistic graphical models** can be used to construct representations through the use of latent random variables [19]. In this section we describe how probabilistic graphical models are used to represent unimodal and multimodal data. One such way to represent data is through deep Boltzmann machines (DBM) [183], that stack restricted Boltzmann machines (RBM) [84] as building blocks. Similar to neural networks, each successive layer of a DBM is expected to represent the data at a higher level of abstraction. The appeal of DBMs comes from the fact that they do not need supervised data for training [183]. As they are graphical models the representation of data is probabilistic, however it is possible to convert them to a deterministic neural network — but this loses the generative aspect of the model [183].

Work by Srivastava and Salakhutdinov [205] introduced multimodal deep belief networks and multimodal DBMs [206] as multimodal representations. Kim et al. [108] used a deep belief network for each modality and then combined them into joint representation for audiovisual emotion recognition. Huang and Kingsbury [89] used a similar model for AVSR, and Wu et al. [233] for audio and skeleton joint based gesture recognition. Ouyang et al. [163] explored the use of multimodal DBMs for the task of human pose estimation from multi-view data. They demonstrated that integrating the data at a later stage — after unimodal data underwent nonlinear transformations — was beneficial for the model. Similarly, Suk et al. [207] used multimodal DBM representation to perform Alzheimer's disease classification from positron emission tomography and magnetic resonance imaging data.

One of the big advantages of using multimodal DBMs for learning multimodal representations is their generative nature, which allows for an easy way to deal with missing data — even if a whole modality is missing, the model has a natural way to cope. It can also be used to generate samples of one modality in the presence of the other one, or both modalities from the representation. Similar to autoencoders the representation can be trained in an unsupervised

Table 2: A summary of multimodal representation techniques. We identify three subtypes of joint representations (Section 3.1) and two subtypes of coordinated ones (Section 3.2). For modalities + indicates the modalities combined.

| REPRESENTATION | MODALITIES | REFERENCE |
|---|---|---|
| **Joint** | | |
| Neural networks | Images + Audio | [150], [157], [235] |
| | Images + Text | [191] |
| Graphical models | Images + Text | [206] |
| | Images + Audio | [108] |
| Sequential | Audio + Video | [100], [158] |
| | Images + Text | [173] |
| **Coordinated** | | |
| Similarity | Images + Text | [64], [110] |
| | Video + Text | [166], [239] |
| Structured | Images + Text | [33], [220], [256] |
| | Audio + Articulatory | [228] |

manner enabling the use of unlabeled data. The major disadvantage of DBMs is the difficulty of training them — high computational cost, and the need to use approximate variational training methods [206].

**Sequential Representation**. So far we have discussed models that can represent fixed length data, however, we often need to represent varying length sequences such as sentences, videos, or audio streams. Recurrent neural networks (RNNs), and their variants such as long-short term memory (LSTMs) networks [85], have recently gained popularity due to their success in sequence modeling across various tasks [13], [222]. So far RNNs have mostly been used to represent unimodal sequences of words, audio, or images, with most success in the language domain. Similar to traditional neural networks, the hidden state of an RNN can be seen as a representation of the data, i.e., the hidden state of RNN at timestep $t$ can be seen as the summarization of the sequence up to that timestep. This is especially apparent in RNN encoder-decoder frameworks where the task of an encoder is to represent a sequence in the hidden state of an RNN in such a way that a decoder could reconstruct it [13], [244].

The use of RNN representations has not been limited to the unimodal domain. An early use of constructing a multimodal representation using RNNs comes from work by Cosi et al. [45] on AVSR. They have also been used for representing audio-visual data for affect recognition [39], [158] and to represent multi-view data such as different visual cues for human behavior analysis [173].

### 3.2 Coordinated Representations

An alternative to a joint multimodal representation is a coordinated representation. Instead of projecting the modalities together into a joint space, separate representations are learned for each modality but are coordinated through a constraint. We start our discussion with coordinated representations that enforce similarity between representations, moving on to coordinated representations that enforce more structure on the resulting space (representative works of such coordinated representations can be seen in Table 2).

**Similarity models** minimize the distance between modalities in the coordinated space. For example such models encourage the representation of the word *dog* and an image of a dog to have a smaller distance between them than distance between the word *dog* and an image of a car [64]. One of the earliest examples of such a representation comes from the work by Weston et al. [229], [230] on the WSABIE (web scale annotation by image embedding) model, where a coordinated space was constructed for images and their annotations. WSABIE constructs a simple linear map from image and textual features such that corresponding annotation and image representation would have a higher inner product (smaller cosine distance) between them than non-corresponding ones.

More recently, neural networks have become a popular way to construct coordinated representations, due to their ability to learn representations. Their advantage lies in the fact that they can jointly learn coordinated representations in an end-to-end manner. An example of such coordinated representation is DeViSE — a deep visual-semantic embedding [64]. DeViSE uses a similar inner product and ranking loss function to WSABIE but uses more complex image and word embeddings. Kiros et al. [110] extended this to sentence and image coordinated representation by using an LSTM model and a pairwise ranking loss to coordinate the feature space. Socher et al. [199] tackle the same task, but extend the language model to a dependency tree RNN to incorporate compositional semantics. A similar model was also proposed by Pan et al. [166], but using videos instead of images. Xu et al. [239] also constructed a coordinated space between videos and sentences using a $\langle$subject, verb, object$\rangle$ compositional language model and a deep video model. This representation was then used for the task of cross-modal retrieval and video description.

While the above models enforced similarity between representations, **structured coordinated space** models go beyond that and enforce additional constraints between the modality representations. The type of structure enforced is often based on the application, with different constraints for hashing, cross-modal retrieval, and image captioning.

Structured coordinated spaces are commonly used in cross-modal hashing — compression of high dimensional data into compact binary codes with similar binary codes for similar objects [226]. The idea of cross-modal hashing is to create such codes for cross-modal retrieval [28], [97], [118]. Hashing enforces certain constraints on the resulting multimodal space: 1) it has to be an $N$-dimensional Hamming space — a binary representation with controllable number of bits; 2) the same object from different modalities has to have a similar hash code; 3) the space has to be similarity-preserving. Learning how to represent the data as a hash function attempts to enforce all of these three requirements [28], [118]. For example, Jiang and Li [96] introduced a method to learn such common binary space between sentence descriptions and corresponding images using end-to-end trainable deep learning techniques. While Cao et al. [33] extended the approach with a more complex LSTM sentence representation and introduced an outlier insensitive bit-wise margin loss and a relevance feedback based semantic similarity constraint. Similarly, Wang et al. [227] constructed a coordinated space in which images (and sentences) with similar meanings are closer to each other.

Another example of a structured coordinated representation comes from order-embeddings of images and language [220], [257]. The model proposed by Vendrov et al.

[220] enforces a dissimilarity metric that is asymmetric and implements the notion of partial order in the multimodal space. The idea is to capture a partial order of the language and image representations — enforcing a hierarchy on the space; for example image of "a woman walking her dog" → text "woman walking her dog" → text "woman walking". A similar model using denotation graphs was also proposed by Young et al. [246] where denotation graphs are used to induce a partial ordering. Lastly, Zhang et al. present how exploiting structured representations of text and images can create concept taxonomies in an unsupervised manner [257].

A special case of a structured coordinated space is one based on canonical correlation analysis (CCA) [87]. CCA computes a linear projection which maximizes the correlation between two random variables (in our case modalities) and enforces orthogonality of the new space. CCA models have been used extensively for cross-modal retrieval [79], [111], [176] and audiovisual signal analysis [184], [195]. Extensions to CCA attempt to construct a correlation maximizing nonlinear projection [7], [121]. Kernel canonical correlation analysis (KCCA) [121] uses reproducing kernel Hilbert spaces for projection. However, as the approach is nonparametric it scales poorly with the size of the training set and has issues with very large real-world datasets. Deep canonical correlation analysis (DCCA) [7] was introduced as an alternative to KCCA and addresses the scalability issue, it was also shown to lead to better correlated representation space. Similar correspondence autoencoder [61] and deep correspondence RBMs [60] have also been proposed for cross-modal retrieval.

CCA, KCCA, and DCCA are unsupervised techniques and only optimize the correlation over the representations, thus mostly capturing what is shared across the modalities. Deep canonically correlated autoencoders [228] also include an autoencoder based data reconstruction term. This encourages the representation to also capture modality specific information. Semantic correlation maximization method [256] also encourages semantic relevance, while retaining correlation maximization and orthogonality of the resulting space — this leads to a combination of CCA and cross-modal hashing techniques.

### 3.3 Discussion

In this section we identified two major types of multimodal representations — joint and coordinated. Joint representations project multimodal data into a common space and are best suited for situations when all of the modalities are present during inference. They have been extensively used for AVSR, affect, and multimodal gesture recognition. Coordinated representations, on the other hand, project each modality into a separate but coordinated space, making them suitable for applications where only one modality is present at test time, such as: multimodal retrieval and translation (Section 4), grounding (Section 7.2), and zero shot learning (Section 7.2). Furthermore, while joint representations have been used in situations to construct representations of more than two modalities, coordinated spaces have, so far, been mostly limited to two. Finally, the multimodal networks we discussed are largely static, in the future we may see more work on one modality driving the structure of a network applied to another modality [6].

Table 3: Taxonomy of multimodal translation research. For each class and sub-class, we include example tasks with references. Our taxonomy also includes the directionality of the translation: unidirectional (⇒) and bidirectional (⇔).

| | TASKS | DIR. | REFERENCES |
|---|---|---|---|
| **Example-based** | | | |
| Retrieval | Image captioning | ⇒ | [58], [162] |
| | Media retrieval | ⇔ | [199], [239] |
| | Visual speech | ⇒ | [27] |
| | Image captioning | ⇔ | [102], [103] |
| Combination | Image captioning | ⇒ | [77], [119], [124] |
| **Generative** | | | |
| Grammar based | Video description | ⇒ | [15], [213] |
| | Image description | ⇒ | [53], [126], [147] |
| Encoder-decoder | Image captioning | ⇒ | [110], [139] |
| | Video description | ⇒ | [222], [249] |
| | Text to image | ⇒ | [137], [178] |
| Continuous | Sounds synthesis | ⇒ | [161], [164] |
| | Visual speech | ⇒ | [5], [49], [212] |

## 4 TRANSLATION

A big part of multimodal machine learning is concerned with translating (mapping) from one modality to another. Given an entity in one modality the task is to generate the same entity in a different modality. For example given an image we might want to generate a sentence describing it or given a textual description generate an image matching it. Multimodal translation is a long studied problem, with early work in speech synthesis [91], visual speech generation [141] video description [112], and cross-modal retrieval [176].

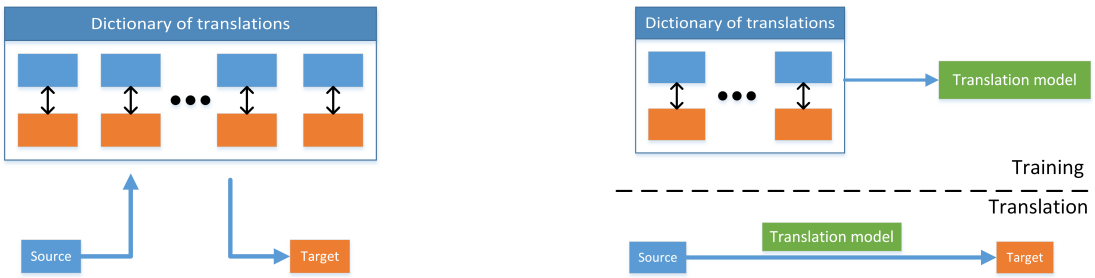
More recently, multimodal translation has seen renewed interest due to combined efforts of the computer vision and natural language processing (NLP) communities [20] and recent availability of large multimodal datasets [40], [214]. A particularly popular problem is visual scene description, also known as image [223] and video captioning [222], which acts as a great test bed for a number of computer vision and NLP problems. To solve it, we not only need to fully understand the visual scene and to identify its salient parts, but also to produce grammatically correct and comprehensive yet concise sentences describing it.

While the approaches to multimodal translation are very broad and are often modality specific, they share a number of unifying factors. We categorize them into two types — *example-based*, and *generative*. Example-based models use a *dictionary* when translating between the modalities. Generative models, on the other hand, construct a *model* that is able to produce a translation. This distinction is similar to the one between non-parametric and parametric machine learning approaches and is illustrated in Figure 2, with representative examples summarized in Table 3.

*Generative* models are arguably more challenging to build as they require the ability to generate signals or sequences of symbols (e.g., sentences). This is difficult for any modality — visual, acoustic, or verbal, especially when temporally and structurally consistent sequences need to be generated. This led to many of the early multimodal translation systems relying on *example-based* translation. However, this has been changing with the advent of deep learning models that are capable of generating images [178], [218], sounds [161], [164], and text [13].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2798607, IEEE Transactions on Pattern Analysis and Machine Intelligence






(a) Example-based (b) Generative

Figure 2: Overview of *example-based* and *generative* multimodal translation. The former retrieves the best translation from a dictionary, while the latter first trains a translation model on the dictionary and then uses that model for translation.

### 4.1 Example-based

Example-based algorithms are restricted by their training data — dictionary (see Figure 2a). We identify two types of such algorithms: retrieval based, and combination based. *Retrieval*-based models directly use the retrieved translation without modifying it, while *combination*-based models rely on more complex rules to create translations based on a number of retrieved instances.

**Retrieval-based models** are arguably the simplest form of multimodal translation. They rely on finding the closest sample in the dictionary and using that as the translated result. The retrieval can be done in *unimodal* space or intermediate *semantic* space.

Given a source modality instance to be translated, unimodal retrieval finds the closest instances in the dictionary in the space of the source — for example, visual feature space for images. Such approaches have been used for visual speech synthesis, by retrieving the closest matching visual example of the desired phoneme [27]. They have also been used in concatenative text-to-speech systems [91]. More recently, Ordonez et al. [162] used unimodal retrieval to generate image descriptions by using global image features to retrieve caption candidates [162]. Yagcioglu et al. [240] used a CNN-based image representation to retrieve visually similar images using adaptive neighborhood selection. Devlin et al. [51] demonstrated that a simple $k$-nearest neighbor retrieval with consensus caption selection achieves competitive translation results when compared to more complex generative approaches. The advantage of such unimodal retrieval approaches is that they only require the representation of a single modality through which we are performing retrieval. However, they often require an extra multimodal post-processing step such as re-ranking of retrieved translations [140], [162], [240]. This indicates a major problem with this approach — similarity in unimodal space does not always imply a good translation.

An alternative is to use an intermediate semantic space for similarity comparison during retrieval. An early example of a hand crafted semantic space is one used by Farhadi et al. [58]. They map both sentences and images to a space of ⟨object, action, scene⟩, retrieval of relevant caption to an image is then performed in that space. In contrast to hand-crafting a representation, Socher et al. [199] learn a coordinated representation of sentences and CNN visual features (see Section 3.2 for description of coordinated spaces). They use the model for both translating from text to images and from images to text. Similarly, Xu et al. [239] used a coordinated space of videos and their descriptions for cross-modal retrieval. Jiang and Li [97] and Cao et al. [33] use cross-modal hashing to perform multimodal translation from images to sentences and back, while Hodosh et al. [86] use a multimodal KCCA space for image-sentence retrieval. Instead of aligning images and sentences globally in a common space, Karpathy et al. [103] propose a multimodal similarity metric that internally aligns image fragments (visual objects) together with sentence fragments (dependency tree relations).

Retrieval approaches in semantic space tend to perform better than their unimodal counterparts as they are retrieving examples in a more meaningful space that reflects both modalities and that is often optimized for retrieval. Furthermore, they allow for bi-directional translation, which is not straightforward with unimodal methods. However, they require manual construction or learning of such a semantic space, which often relies on the existence of large training dictionaries (datasets of paired samples).

**Combination-based models** take the retrieval based approaches one step further. Instead of just retrieving examples from the dictionary, they combine them in a meaningful way to construct a better translation. Combination based media description approaches are motivated by the fact that sentence descriptions of images share a common and simple structure that could be exploited. Most often the rules for combinations are hand crafted or based on heuristics.

Kuznetsova et al. [119] first retrieve phrases that describe visually similar images and then combine them to generate novel descriptions of the query image by using Integer Linear Programming with a number of hand crafted rules. Gupta et al. [77] first find $k$ images most similar to the source image, and then use the phrases extracted from their captions to generate a target sentence. Lebret et al. [124] use a CNN-based image representation to infer phrases that describe it. The predicted phrases are then combined using a trigram constrained language model.

A big problem facing example-based approaches for translation is that the model is the entire dictionary — making the model large and inference slow (although, optimizations such as hashing alleviate this problem). Another issue facing example-based translation is that it is unrealistic to expect that a single comprehensive and accurate translation relevant to the source example will always exist in the dictionary — unless the task is simple or the dictionary is very

large. This is partly addressed by combination models that are able to construct more complex structures. However, they are only able to perform translation in one direction, while semantic space retrieval-based models are able to perform it both ways.

### 4.2 Generative approaches

Generative approaches to multimodal translation construct models that can perform multimodal translation given a unimodal source instance. It is a challenging problem as it requires the ability to both understand the source modality and to generate the target sequence or signal. As discussed in the following section, this also makes such methods much more difficult to evaluate, due to large space of possible correct answers.

In this survey we focus on the generation of three modalities: language, vision, and sound. Language generation has been explored for a long time [177], with a lot of recent attention for tasks such as image and video description [20]. Speech and sound generation has also seen a lot of work with a number of historical [91] and modern approaches [161], [164]. Photo-realistic image generation has been less explored, and is still in early stages [137], [178], however, there have been a number of attempts at generating abstract scenes [261], computer graphics [47], and talking heads [5].

We identify three broad categories of generative models: *grammar-based*, *encoder-decoder*, and *continuous generation* models. Grammar based models simplify the task by restricting the target domain by using a grammar, e.g., by generating restricted sentences based on a ⟨subject, object, verb⟩ template. Encoder-decoder models first encode the source modality to a latent representation which is then used by a decoder to generate the target modality. Continuous generation models generate the target modality continuously based on a stream of source modality inputs and are most suited for translating between temporal sequences — such as text-to-speech.

**Grammar-based models** rely on a pre-defined grammar for generating a particular modality. They start by detecting high level concepts from the source modality, such as objects in images and actions from videos. These detections are then incorporated together with a generation procedure based on a pre-defined grammar to result in a target modality.

Kojima et al. [112] proposed a system to describe human behavior in a video using the detected position of the person's head and hands and rule based natural language generation that incorporates a hierarchy of concepts and actions. Barbu et al. [15] proposed a video description model that generates sentences of the form: *who* did *what* to *whom* and *where* and *how* they did it. The system was based on handcrafted object and event classifiers and used a restricted grammar suitable for the task. Guadarrama et al. [76] predict ⟨subject, verb, object⟩ triplets describing a video using semantic hierarchies that use more general words in case of uncertainty. Together with a language model their approach allows for translation of verbs and nouns not seen in the dictionary.

To describe images, Yao et al. [243] propose to use an and-or graph-based model together with domain-specific lexicalized grammar rules, targeted visual representation scheme, and a hierarchical knowledge ontology. Li et al. [126] first detect objects, visual attributes, and spatial relationships between objects. They then use an $n$-gram language model on the visually extracted phrases to generate ⟨subject, preposition, object⟩ style sentences. Mitchell et al. [147] use a more sophisticated tree-based language model to generate syntactic trees instead of filling in templates, leading to more diverse descriptions. A majority of approaches represent the whole image jointly as a bag of visual objects without capturing their spatial and semantic relationships. To address this, Elliott et al. [53] propose to explicitly model proximity relationships of objects for image description generation.

Some grammar-based approaches rely on graphical models to generate the target modality. An example includes BabyTalk [117], which given an image generates ⟨object, preposition, object⟩ triplets, that are used together with a conditional random field to construct the sentences. Yang et al. [241] predict a set of ⟨noun, verb, scene, preposition⟩ candidates using visual features extracted from an image and combine them into a sentence using a statistical language model and hidden Markov model style inference. A similar approach has been proposed by Thomason et al. [213], where a factor graph model is used for video description of the form ⟨subject, verb, object, place⟩. The factor model exploits language statistics to deal with noisy visual representations. Going the other way Zitnick et al. [261] propose to use conditional random fields to generate abstract visual scenes based on language triplets extracted from sentences.

An advantage of grammar-based methods is that they are more likely to generate syntactically (in case of language) or logically correct target instances as they use predefined templates and restricted grammars. However, this limits them to producing formulaic rather than creative translations. Furthermore, grammar-based methods rely on complex pipelines for concept detection, with each concept requiring a separate model and a separate training dataset.

**Encoder-decoder models** based on end-to-end trained neural networks are currently some of the most popular techniques for multimodal translation. The main idea behind the model is to first encode a source modality into a vectorial representation and then to use a decoder module to generate the target modality, all this in a single pass pipeline. Although, first used for machine translation [101], [208], such models have been successfully used for image captioning [139], [223], and video description [181], [222]. While encoder-decoder models have been mostly used to generate text, they can also generate images [137], [178], and speech and sound [161], [164].

The first step of the encoder-decoder model is to encode the source object, this is done in modality specific way. Popular models to encode acoustic signals include RNNs [36] and DBNs [82]. Most of the work on encoding words sentences uses distributional semantics [146] and variants of RNNs [13]. Images are most often encoded using convolutional neural networks (CNN) [114], [193]. Although there are methods for learning video representations [59], [192], hand-crafted features are still used [181], [213]. While it is possible to use unimodal representations to encode the source modality, it has been shown that using a coordinated

space (see Section 3.2) leads to better results [110], [166].

Decoding is most often performed by an RNN or an LSTM using the encoded representation as the initial hidden state [56], [137], [223], [223]. A number of extensions have been proposed to traditional LSTM models to aid in the task of translation. A guide vector could be used to tightly couple the solutions in the image input [95]. Venugopalan et al. [222] demonstrate that it is beneficial to pre-train a decoder LSTM for image captioning before fine-tuning it to video description. Rohrbach et al. [181] explore the use of various LSTM architectures (single layer, multilayer, factored) and a number of training and regularization techniques for the task of video description.

A problem facing translation generation using an RNN is that the model has to generate a description from a single vectorial representation of the image, sentence, or video. This becomes especially difficult when generating long sequences as these models tend to *forget* the initial input. This has been partly addressed by including the encoded information during every step of the decoder [95]. Attention models (see Section 5.2) have also been proposed to allow the decoder to better focus on certain parts of an image [238], sentence [13], or video [244] during generation.

Generative attention-based RNNs have also been used for the task of generating images from sentences [137], while the results are still far from photo-realistic they show a lot of promise. More recently, a large amount of progress has been made in generating images using generative adversarial networks [74], which have been used as an alternative to RNNs for image generation from text [178].

While neural network based encoder-decoder systems have been very successful they still face a number of issues. Devlin et al. [51] suggest that it is possible that the network is *memorizing* the training data rather than learning how to understand the visual scene and generate it, based on the observation that $k$-nearest neighbor models perform similarly to those based on generation. Furthermore, such models often require large quantities of data for training.

**Continuous generation models** are intended for sequence translation and produce outputs at every timestep in an online manner. These models are useful when translating from a sequence to a sequence such as text to speech, speech to text, and video to text. A number of different techniques have been proposed for such modeling — graphical models, continuous encoder-decoder approaches, and various other regression or classification techniques. The extra difficulty that needs to be tackled by these models is the requirement of temporal consistency between modalities.

A lot of early work on sequence to sequence translation used graphical or latent variable models. Deena and Galata [49] proposed to use a shared Gaussian process latent variable model for audio-based visual speech synthesis. The model creates a shared latent space between audio and visual features that can be used to generate one space from the other, while enforcing temporal consistency of visual speech at different timesteps. Hidden Markov models (HMM) have also been used for visual speech generation [212] and text-to-speech [253] tasks. They have also been extended to use cluster adaptive training to allow for training on multiple speakers, languages, and emotions allowing for more control when generating speech signal [252] or visual speech parameters [5].

Encoder-decoder models have recently become popular for sequence to sequence modeling. Owens et al. [164] used an LSTM to generate sounds resulting from drumsticks based on video. While their model is capable of generating sounds by predicting a cochleogram from CNN visual features, they found that retrieving a closest audio sample based on the predicted cochleogram led to best results. Directly modeling the raw audio signal for speech and music generation has been proposed by van den Oord et al. [161]. The authors propose using hierarchical fully convolutional neural networks, which show a large improvement over previous state-of-the-art for the task of speech synthesis. RNNs have also been used for speech to text translation (speech recognition) [75]. More recently encoder-decoder based continuous approach was shown to be good at predicting letters from a speech signal represented as a filter bank spectra [36] — allowing for more accurate recognition of rare and out of vocabulary words. Collobert et al. [44] demonstrate how to use a raw audio signal directly for speech recognition, eliminating the need for audio features.

A lot of earlier work used graphical models for multimodal translation between continuous signals. However, these methods are being replaced by neural network encoder-decoder based techniques. Especially as they have recently been shown to be able to represent and generate complex visual and acoustic signals.

### 4.3 Model evaluation and discussion

A major challenge facing multimodal translation methods is that they are very difficult to evaluate. While some tasks such as speech recognition have a single correct translation, tasks such as speech synthesis and media description do not. Sometimes, as in language translation, multiple answers are correct and deciding which translation is better is often subjective. Fortunately, there are a number of approximate automatic metrics that aid in model evaluation.

Often the ideal way to evaluate a subjective task is through human judgment. That is by having a group of people evaluating each translation. This can be done on a Likert scale where each translation is evaluated on a certain dimension: naturalness and mean opinion score for speech synthesis [161], [252], realism for visual speech synthesis [5], [212], and grammatical and semantic correctness, relevance, order, and detail for media description [40], [117], [147], [222]. Another option is to perform preference studies where two (or more) translations are presented to the participant for preference comparison [212], [252]. However, while user studies will result in evaluation closest to human judgments they are time consuming and costly. Furthermore, they require care when constructing and conducting them to avoid fluency, age, gender and culture biases.

While human studies are a gold standard for evaluation, a number of automatic alternatives have been proposed for the task of media description: BLEU [167], ROUGE [129], Meteor [50], and CIDEr [219]. However, the use of them has faced a lot of criticism and have been shown to only weakly correspond to human judgements [54], [90].

Hodosh et al. [86] propose using retrieval as a proxy for image captioning evaluation, as a better way to reflect human judgments. Instead of generating captions, a retrieval

Table 4: Summary of our taxonomy for the multimodal alignment challenge. For each sub-class of our taxonomy, we include reference citations and modalities aligned.

| ALIGNMENT | MODALITIES | REFERENCE |
|---|---|---|
| **Explicit** | | |
| Unsupervised | Video + Text | [136], [210], [211] |
| | Video + Audio | [160], [215], [259] |
| Supervised | Video + Text | [24], [260] |
| | Image + Text | [113], [138], [168] |
| **Implicit** | | |
| Graphical models | Audio/Text + Text | [194], [224] |
| Neural networks | Image + Text | [102], [236], [238] |
| | Video + Text | [244], [249] |

based system ranks the available captions based on their fit to the image, and is then evaluated by assessing if the correct captions are given a high rank. As a number of caption generation models are generative they can be used directly to assess the likelihood of a caption given an image and are being adapted by image captioning community [103], [110]. Such retrieval based evaluation metrics have also been adopted by the video captioning community [182].

Visual question-answering (VQA) [135] task was proposed partly due to the issues facing evaluation of image captioning. VQA is a task where given an image and a question about its content the system has to answer it. Evaluating such systems is easier due to the presence of a *correct* answer, turning the task into a multimodal fusion (see Section 6) rather than a translation one. Image co-reference task [113], [138] was proposed to address this ambiguity as well, by framing the task as that of multi-modal alignment (see Section 5).

We believe that addressing the evaluation issue will be crucial for further success of multimodal translation systems. This will allow not only for better comparison between approaches, but also for better objectives to optimize.

## 5 ALIGNMENT

We define multimodal alignment as finding relationships and correspondences between sub-components of instances from two or more modalities. For example, given an image and a caption we want to find the areas of the image corresponding to the caption's words or phrases [102]. Another example is, given a movie, aligning it to the script or the book chapters it was based on [260]. Ability to do this is particularly important for multimedia retrieval, as it enables us to search video content based on text, e.g. finding scenes in a movie where a particular character appears, or finding images that contain blue chairs.

We categorize multimodal alignment into two types – *implicit* and *explicit*. In explicit alignment, we are explicitly interested in aligning sub-components between modalities, e.g., aligning recipe steps with the corresponding instructional video [136]. Implicit alignment is used as an intermediate (often latent) step for another task, e.g., image retrieval based on text description can include an alignment step between words and image regions [103]. An overview of such approaches can be seen in Table 4 and is presented in more detail in the following sections.

### 5.1 Explicit alignment

We categorize papers as performing explicit alignment if their main modeling objective is alignment between sub-components of instances from two or more modalities. A very important part of explicit alignment is the similarity metric. Most approaches rely on measuring similarity between sub-components in different modalities as a basic building block. These similarities can be defined manually or learned from data. We identify two types of algorithms that tackle explicit alignment — *unsupervised* and (weakly) *supervised*. The first type operates with no direct alignment labels (i.e., labeled correspondences) between instances from the different modalities. The second type has access to such (sometimes weak) labels.

**Unsupervised** multimodal alignment tackles modality alignment without requiring any direct alignment labels. Most of the approaches are inspired from early work on alignment for statistical machine translation [29] and genome sequences [116], [151]. To make the task easier the approaches assume certain constrains on alignment, such as temporal ordering of sequence or an existence of a similarity metric between the modalities.

Dynamic time warping (DTW) [116], [151] is a dynamic programming approach that has been extensively used to align multi-view time series. DTW measures the similarity between two sequences and finds an optimal match between them by time warping (inserting frames). It requires the timesteps in the two sequences to be comparable and requires a similarity measure between them. DTW can be used directly for multimodal alignment by hand-crafting similarity metrics between modalities; for example Anguera et al. [8] use a manually defined similarity between graphemes and phonemes; and Tapaswi et al. [210] define a similarity between visual scenes and sentences based on appearance of same characters [210] to align TV shows and plot synopses. DTW-like dynamic programming approaches have also been used for multimodal alignment of text to speech [80] and video [211].

As the original DTW formulation requires a pre-defined similarity metric between modalities, it was extended using canonical correlation analysis (CCA) to map the modalities to a coordinated space. This allows for both aligning (through DTW) and learning the mapping (through CCA) between different modality streams jointly and in an unsupervised manner [187], [258], [259]. While CCA based DTW models are able to find multimodal data alignment under a linear transformation, they are not able to model non-linear relationships. This has been addressed by the deep canonical time warping approach [215], which can be seen as a generalization of deep CCA and DTW.

Various graphical models have also been popular for multimodal sequence alignment in an unsupervised manner. Early work by Yu and Ballard [247] used a generative graphical model to align visual objects in images with spoken words. A similar approach was taken by Cour et al. [46] to align movie shots and scenes to the corresponding screenplay. Malmaud et al. [136] used a factored HMM to align recipes to cooking videos, while Noulas et al. [160] used a dynamic Bayesian network to align speakers to videos. Naim et al. [153] matched sentences with corre-

sponding video frames using a hierarchical HMM model to align sentences with frames and a modified IBM [29] algorithm for word and object alignment [16]. This model was then extended to use latent conditional random fields for alignments [152] and to incorporate verb alignment to actions in addition to nouns and objects [203].

Both DTW and graphical model approaches for alignment allow for restrictions on alignment, e.g. temporal consistency, no large jumps in time, and monotonicity. While DTW extensions allow for learning both the similarity metric and alignment jointly, graphical model based approaches require expert knowledge for construction [46], [247].

**Supervised** alignment methods rely on labeled aligned instances. They are used to train similarity measures that are used for aligning modalities.

A number of supervised sequence alignment techniques take inspiration from unsupervised ones. Bojanowski et al. [23], [24] proposed a method similar to canonical time warping, but have also extended it to take advantage of existing (weak) supervisory alignment data for model training. Plummer et al. [168] used CCA to find a coordinated space between image regions and phrases for alignment. Gebru et al. [68] trained a Gaussian mixture model and performed semi-supervised clustering together with an unsupervised latent-variable graphical model to align speakers in an audio channel with their locations in a video. Kong et al. [113] trained a Markov random field to align objects in 3D scenes to nouns and pronouns in text descriptions.

Deep learning based approaches are becoming popular for explicit alignment (specifically for measuring similarity) due to very recent availability of aligned datasets in the language and vision communities [138], [168]. Zhu et al. [260] aligned books with their corresponding movies/scripts by training a CNN to measure similarities between scenes and text. Mao et al. [138] used an LSTM language model and a CNN visual one to evaluate the quality of a match between a referring expression and an object in an image. Yu et al. [250] extended this model to include relative appearance and context information that allows to better disambiguate between objects of the same type. Finally, Hu et al. [88] used an LSTM based scoring function to find similarities between image regions and their descriptions.

### 5.2 Implicit alignment

In contrast to explicit alignment, implicit alignment is used as an intermediate (often latent) step for another task. This allows for better performance in a number of tasks including speech recognition, machine translation, media description, and visual question-answering. Such models do not explicitly align data and do not rely on supervised alignment examples, but learn how to latently align the data during model training. We identify two types of implicit alignment models: earlier work based on graphical models, and more modern neural network methods.

**Graphical models** have seen some early work used to better align words between languages for machine translation [224] and alignment of speech phonemes with their transcriptions [194]. However, they require manual construction of a mapping between the modalities, for example a generative phone model that maps phonemes to acoustic features [194]. Constructing such models requires training data or human expertise to define them manually.

**Neural networks** Translation (Section 4) is an example of a modeling task that can often be improved if alignment is performed as a latent intermediate step. As we mentioned before, neural networks are popular ways to address this translation problem, using either an encoder-decoder model or through cross-modal retrieval. When translation is performed without implicit alignment, it ends up putting a lot of weight on the encoder module to be able to properly summarize the whole image, sentence or a video with a single vectorial representation.

A very popular way to address this is through *attention* [13], which allows the decoder to focus on sub-components of the source instance. This is in contrast with encoding all source sub-components together, as is performed in a conventional encoder-decoder model. An attention module will tell the decoder to look more at targeted sub-components of the source to be translated — areas of an image [238], words of a sentence [13], segments of an audio sequence [36], [41], frames and regions in a video [244], [249], and even parts of an instruction [145]. For example, in image captioning instead of encoding an entire image using a CNN, an attention mechanism will allow the decoder (typically an RNN) to focus on particular parts of the image when generating each successive word [238]. The attention module which learns what part of the image to focus on is typically a shallow neural network and is trained end-to-end together with a target task (e.g., translation).

Attention models have also been successfully applied to question answering tasks, as they allow for aligning the words in a question with sub-components of an information source such as a piece of text [236], an image [65], or a video sequence [254]. This allows for better accuracy and leads to better model interpretability [3]. In particular, different types of attention models have been proposed to address this problem, including hierarchical [133], stacked [242], and episodic memory attention [236].

Another neural alternative for aligning images with captions for cross-modal retrieval was proposed by Karpathy et al. [102], [103]. Their proposed model aligns sentence fragments to image regions by using a dot product similarity measure between image region and word representations. While it does not use attention, it extracts a latent alignment between modalities through a similarity measure that is learned indirectly by training a retrieval model.

### 5.3 Discussion

Multimodal alignment faces a number of difficulties: 1) there are few datasets with explicitly annotated alignments; 2) it is difficult to design similarity metrics between modalities; 3) there may exist multiple possible alignments and not all elements in one modality have correspondences in another. Earlier work on multimodal alignment focused on aligning multimodal sequences in an unsupervised manner using graphical models and dynamic programming techniques. It relied on hand-defined measures of similarity between the modalities or learnt them in an unsupervised manner. With recent availability of labeled training data supervised learning of similarities between modalities has become possible.

However, unsupervised techniques of learning to jointly align and translate or fuse data have also become popular.

## 6 Fusion

Multimodal fusion is one of the original topics in multimodal machine learning, with previous surveys emphasizing early, late and hybrid fusion approaches [52], [255]. In technical terms, multimodal fusion is the concept of integrating information from multiple modalities with the goal of predicting an outcome measure: a class (e.g., happy vs. sad) through classification, or a continuous value (e.g., positivity of sentiment) through regression. It is one of the most researched aspects of multimodal machine learning with work dating to 25 years ago [251].

The interest in multimodal fusion arises from three main benefits it can provide. First, having access to multiple modalities that observe the same phenomenon may allow for more robust predictions. This has been especially explored and exploited by the AVSR community [170]. Second, having access to multiple modalities might allow us to capture complementary information — something that is not visible in individual modalities on their own. Third, a multimodal system can still operate when one of the modalities is missing, for example recognizing emotions from the visual signal when the person is not speaking [52].

Multimodal fusion has a very broad range of applications, including audio-visual speech recognition (AVSR) [170], multimodal emotion recognition [200], medical image analysis [93], and multimedia event detection [122]. There are a number of reviews on the subject [11], [170], [196], [255]. Most of them concentrate on multimodal fusion for a particular task, such as multimedia analysis, information retrieval or emotion recognition. In contrast, we concentrate on the machine learning approaches themselves and the technical challenges associated with these approaches.

While some prior work used the term multimodal fusion to describe all multimodal algorithms, we classify approaches as fusion when the multimodal integration is performed at the later prediction stages, with the goal of predicting outcome measures. Recently, the line between multimodal representation and fusion has been blurred for models such as deep neural networks where representation learning interacts with classification or regression objectives.

We classify multimodal fusion into two main categories: *model-agnostic* approaches (Section 6.1) that are not directly dependent on a specific machine learning method; and *model-based* (Section 6.2) approaches that explicitly address fusion in their construction — such as kernel-based approaches, graphical models, and neural networks. An overview of such approaches can be seen in Table 5.

Table 5: A summary of our taxonomy of multimodal fusion approaches. Out — output type (class — classification or reg — regression), Temp — is temporal modeling possible.

| Fusion type | Out | Temp | Task | Reference |
|---|---|---|---|---|
| **Model-agnostic** | | | | |
| Early | class | no | Emotion rec. | [35] |
| Late | reg | yes | Emotion rec. | [175] |
| Hybrid | class | no | Multimedia event detection | [122] |
| **Model-based** | | | | |
| Kernel-based | class | no | Object class. | [32], [69] |
| | class | no | Emotion rec. | [94], [189] |
| Graphical models | class | yes | AVSR | [78] |
| | reg | yes | Emotion rec. | [14] |
| | class | no | Media class. | [97] |
| Neural networks | class | yes | Emotion rec. | [100], [232] |
| | class | no | AVSR | [157] |
| | reg | yes | Emotion rec. | [39] |

### 6.1 Model-agnostic approaches

Historically, the vast majority of multimodal fusion has been done using model-agnostic approaches [52]. Such approaches can be split into *early* (i.e., feature-based), *late* (i.e., decision-based) and *hybrid* fusion [11]. Early fusion integrates features immediately after they are extracted (often by simply concatenating their representations). Late fusion on the other hand performs integration after each of the modalities has made a decision (e.g., classification or regression). Finally, hybrid fusion combines outputs from early fusion and individual unimodal predictors. An advantage of model agnostic approaches is that they can be implemented using almost any unimodal classifiers or regressors.

Early fusion could be seen as an early attempt by multimodal researchers to perform multimodal representation learning — as it can learn to exploit the correlation and interactions between low level features of each modality. It also only requires the training of a single model, making the training pipeline easier compared to late and hybrid fusion.

In contrast, late fusion uses unimodal decision values and fuses them using a fusion mechanism such as averaging [188], voting schemes [149], weighting based on channel noise [170] and signal variance [55], or a learned model [71], [175]. It allows for the use of different models for each modality as different predictors can model each individual modality better, allowing for more flexibility. Furthermore, it makes it easier to make predictions when one or more of the modalities is missing and even allows for training when no parallel data is available. However, late fusion ignores the low level interaction between the modalities.

Hybrid fusion attempts to exploit the advantages of both of the above described methods in a common framework. It has been used successfully for multimodal speaker identification [234] and multimedia event detection (MED) [122].

### 6.2 Model-based approaches

While model-agnostic approaches are easy to implement using unimodal machine learning methods, they end up using techniques that are not designed for multimodal data. In this section we describe three categories of approaches that are designed to perform multimodal fusion: kernel-based methods, graphical models, and neural networks.

**Multiple kernel learning (MKL)** methods are an extension to kernel support vector machines (SVM) that allow for the use of different kernels for different modalities/views of the data [73]. As kernels can be seen as similarity functions between data points, modality-specific kernels in MKL allows for better fusion of heterogeneous data.

MKL approaches have been an especially popular method for fusing visual descriptors for object detection [32], [69] and only recently have been overtaken by deep

learning methods for the task [114]. They have also seen use for multimodal affect recognition [38], [94], [189], multimodal sentiment analysis [169], and multimedia event detection (MED) [245]. Furthermore, McFee and Lanckriet [142] proposed to use MKL to perform musical artist similarity ranking from acoustic, semantic and social view data. Finally, Liu et al. [130] used MKL for multimodal fusion in Alzheimer's disease classification. Their broad applicability demonstrates the strength of such approaches in various domains and across different modalities.

Besides flexibility in kernel selection, an advantage of MKL is the fact that the loss function is convex, allowing for model training using standard optimization packages and global optimum solutions [73]. Furthermore, MKL can be used to both perform regression and classification. One of the main disadvantages of MKL is the reliance on training data (support vectors) during test time, leading to slow inference and a large memory footprint.

**Graphical models** are another family of popular methods for multimodal fusion. In this section we overview work done on multimodal fusion using *shallow* graphical models. A description of deep graphical models such as deep belief networks can be found in Section 3.1.

Majority of graphical models can be classified into two main categories: generative — modeling joint probability; or discriminative — modeling conditional probability [209]. Some of the earliest approaches to use graphical models for multimodal fusion include generative models such as coupled [155] and factorial hidden Markov models [70] alongside dynamic Bayesian networks [67]. A more recently-proposed multi-stream HMM method proposes dynamic weighting of modalities for AVSR [78].

Arguably, generative models lost popularity to discriminative ones such as conditional random fields (CRF) [120] which sacrifice the modeling of joint probability for predictive power. A CRF model was used to better segment images by combining visual and textual information of image description [63]. CRF models have been extended to model latent states using hidden conditional random fields [172] and have been applied to multimodal meeting segmentation [180]. Other multimodal uses of latent variable discriminative graphical models include multi-view hidden CRF [202] and latent variable models [201]. More recently Jiang et al. [97] have shown the benefits of multimodal hidden conditional random fields for the task of multimedia classification. While most graphical models are aimed at classification, CRF models have been extended to a continuous version for regression [171] and applied in multimodal settings [14] for audio visual emotion recognition.

The benefit of graphical models is their ability to easily exploit spatial and temporal structure of the data, making them especially popular for temporal modeling tasks, such as AVSR and multimodal affect recognition. They also allow to build in human expert knowledge into the models. and often lead to interpretable models.

**Neural networks** have been used extensively for the task of multimodal fusion [157]. The earliest examples of using neural networks for multi-modal fusion come from work on AVSR [170]. Nowadays they are being used to fuse information for visual and media question answering [66], [135], [237], gesture recognition [156], affect analysis [100], [159], and video description generation [98], [221]. Both shallow [66] and deep [159], [221] neural models have been explored for multimodal fusion.

Neural networks have also been used for fusing temporal multimodal information through the use of RNNs and LSTMs. One of the earlier such applications used a bidirectional LSTM was used to perform audio-visual emotion classification [232]. More recently, Wöllmer et al. [231] used LSTM models for continuous multimodal emotion recognition, demonstrating its advantage over graphical models and SVMs. Similarly, Nicolaou et al. [158] used LSTMs for continuous emotion prediction. Their proposed method used an LSTM to fuse the results from a modality specific (audio and facial expression) LSTMs.

Approaching modality fusion through recurrent neural networks has been used in various image captioning tasks, example models include: neural image captioning [223] where a CNN image representation is decoded using an LSTM language model, gLSTM [95] which incorporates the image data together with sentence decoding at every time step fusing the visual and sentence data in a joint representation. A more recent example is the multi-view LSTM (MV-LSTM) model proposed by Rajagopalan et al. [173]. MV-LSTM model allows for flexible fusion of modalities in the LSTM framework by explicitly modeling the modality-specific and cross-modality interactions over time.

A big advantage of deep neural network approaches in data fusion is their capacity to learn from large amount of data. Secondly, recent neural architectures allow for end-to-end training of both the multimodal representation component and the fusion component. Finally, they show good performance when compared to non neural network based system and are able to learn complex decision boundaries that other approaches struggle with.

The major disadvantage of neural network approaches is their lack of interpretability. It is difficult to tell what the prediction relies on, and which modalities or features play an important role. Furthermore, neural networks require large training datasets to be successful.

### 6.3 Discussion

Multimodal fusion has been a widely researched topic with a large number of approaches proposed to tackle it, including model agnostic methods, graphical models, multiple kernel learning, and various types of neural networks. Each approach has its own strengths and weaknesses, with some more suited for smaller datasets and others performing better in noisy environments. Most recently, neural networks have become a very popular way to tackle multimodal fusion, however graphical models and multiple kernel learning are still being used, especially in tasks with limited training data or where model interpretability is important.

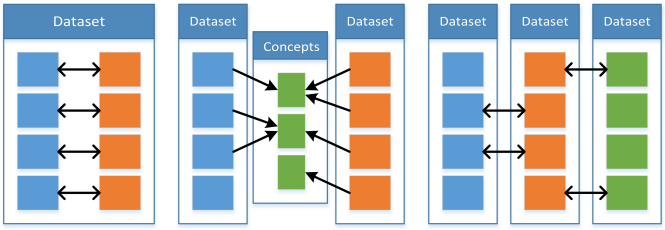
Despite these advances multimodal fusion still faces the following challenges: 1) signals might not be temporally aligned (possibly dense continuous signal and a sparse event); 2) it is difficult to build models that exploit supplementary and not only complementary information; 3) each modality might exhibit different types and different levels of noise at different points in time.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2798607, IEEE Transactions on Pattern Analysis and Machine Intelligence



## 7 Co-learning

The final multimodal challenge in our taxonomy is co-learning — aiding the modeling of a (resource poor) modality by exploiting knowledge from another (resource rich) modality. It is particularly relevant when one of the modalities has limited resources — lack of annotated data, noisy input, and unreliable labels. We call this challenge co-learning as most often the helper modality is used only during model training and is not used during test time. We identify three types of co-learning approaches based on their training resources: parallel, non-parallel, and hybrid. *Parallel-data* approaches require training datasets where the observations from one modality are directly linked to the observations from other modalities. In other words, when the multimodal observations are from the same instances, such as in an audio-visual speech dataset where the video and speech samples are from the same speaker. In contrast, *non-parallel data* approaches do not require direct links between observations from different modalities. These approaches usually achieve co-learning by using overlap in terms of categories. For example, in zero shot learning when the conventional visual object recognition dataset is expanded with a second text-only dataset from Wikipedia to improve the generalization of visual object recognition. In the *hybrid* data setting the modalities are *bridged* through a shared modality or a dataset. An overview of methods in co-learning can be seen in Table 6 and summary of data parallelism in Figure 3.




(a) Parallel (b) Non-parallel (c) Hybrid

Figure 3: Types of data parallelism used in co-learning: *parallel* — modalities are from the same dataset and there is a direct correspondence between instances; *non-parallel* — modalities are from different datasets and do not have overlapping instances, but overlap in general categories or concepts; *hybrid* — the instances or concepts are bridged by a third modality or a dataset.

### 7.1 Parallel data

In parallel data co-learning both modalities share a set of instances — audio recordings with the corresponding videos, images and their sentence descriptions. This allows for two types of algorithms to exploit that data to better model the modalities: co-training and representation learning.

**Co-training** is the process of creating more labeled training samples when we have few labeled samples in a multimodal problem [22]. The basic algorithm builds weak classifiers in each modality to bootstrap each other with labels for the unlabeled data. It has been shown to discover more training samples for web-page classification based on the web-page itself and hyper-links leading in the seminal work of Blum and Mitchell [22]. By definition this task requires parallel data as it relies on the overlap of multimodal samples.

Co-training has been used for statistical parsing [185] to build better visual detectors [125] and for audio-visual speech recognition [42]. It has also been extended to deal with disagreement between modalities, by filtering out unreliable samples [43]. While co-training is a powerful method for generating more labeled data, it can also lead to biased training samples resulting in overfitting.

**Transfer learning** is another way to exploit co-learning with parallel data. Multimodal representation learning (Section 3.1) approaches such as multimodal deep Boltzmann machines [206] and multimodal autoencoders [157] transfer information from representation of one modality to that of another. This not only leads to multimodal representations, but also to better unimodal ones, with only one modality being used during test time [157] .

Moon et al. [148] show how to transfer information from a speech recognition neural network (based on audio) to a lip-reading one (based on images), leading to a better visual representation, and a model that can be used for lip-reading without need for audio information during test time. Similarly, Arora and Livescu [10] build better acoustic features using CCA on acoustic and articulatory (location of lips, tongue and jaw) data. They use articulatory data only during CCA construction and use only the resulting acoustic (unimodal) representation during test time.

### 7.2 Non-parallel data

Methods that rely on non-parallel data do not require the modalities to have shared instances, but only shared categories or concepts. Non-parallel co-learning approaches can help when learning representations, allow for better semantic concept understanding and even perform unseen object recognition.

**Transfer learning** is also possible on non-parallel data and allows to learn better representations through transferring information from a representation built using a data rich or clean modality to a data scarce or noisy modality. This type of trasnfer learning is often achieved by using coordinated multimodal representations (see Section 3.2). For example, Frome et al. [64] used text to improve visual representations for image classification by coordinating CNN visual features with word2vec textual ones [146] trained on separate large datasets. Visual representations trained in such a way result in more meaningful errors — mistaking objects for ones of similar category [64]. Mahasseni and Todorovic [134] demonstrated how to regularize a color video based LSTM using an autoencoder LSTM trained on 3D skeleton data by enforcing similarities between their hidden states. Such an approach is able to improve the original LSTM and lead to state-of-the-art performance in action recognition.

**Conceptual grounding** refers to learning semantic meanings or concepts not purely based on language but also on additional modalities such as vision, sound, or even smell [17]. While the majority of concept learning approaches are purely language-based, representations of meaning in humans are not merely a product of our linguistic exposure, but are also *grounded* through our sensorimotor experience and perceptual system [18], [131]. Human semantic knowledge relies heavily on perceptual information [131] and

Table 6: A summary of co-learning taxonomy, based on data parallelism. Parallel data — multiple modalities can see the same instance. Non-parallel data — unimodal instances are independent of each other. Hybrid data — the modalities are *pivoted* through a shared modality or dataset.

| DATA PARALLELISM | TASK | REFERENCE |
|---|---|---|
| **Parallel** | | |
| Co-training | Mixture | [22], [115] |
| Transfer learning | AVSR | [157] |
| | Lip reading | [148] |
| **Non-parallel** | | |
| Transfer learning | Visual classification | [64] |
| | Action recognition | [134] |
| Concept grounding | Metaphor class. | [188] |
| | Word similarity | [107] |
| Zero shot learning | Image class. | [64], [198] |
| | Thought class. | [165] |
| **Hybrid data** | | |
| Bridging | MT and image ret. | [174] |
| | Transliteration | [154] |

many concepts are grounded in the perceptual system and are not purely symbolic [18]. This implies that learning semantic meaning purely from textual information might not be optimal, and motivates the use of visual or acoustic cues to ground our linguistic representations.

Starting from work by Feng and Lapata [62], grounding is usually performed by finding a common latent space between the representations [62], [190] (in case of parallel datasets) or by learning unimodal representations separately and then concatenating them to lead to a multimodal one [30], [105], [179], [188] (in case of non-parallel data). Once a multimodal representation is constructed it can be used on purely linguistic tasks. Shutova et al. [188] and Bruni et al. [30] used grounded representations for better classification of metaphors and literal language. Such representations have also been useful for measuring conceptual similarity and relatedness — identifying how semantically or conceptually related two words are [31], [105], [190] or actions [179]. Furthermore, concepts can be grounded not only using visual signals, but also acoustic ones, leading to better performance especially on words with auditory associations [107], or even olfactory signals [106] for words with smell associations. Finally, there is a lot of overlap between multimodal alignment and conceptual grounding, as aligning visual scenes to their descriptions leads to better textual or visual representations [113], [168], [179], [248].

Conceptual grounding has been found to be an effective way to improve performance on a number of tasks. It also shows that language and vision (or audio) are complementary sources of information and combining them in multimodal models often improves performance. However, one has to be careful as grounding does not always lead to better performance [106], [107], and only makes sense when grounding has relevance for the task — such as grounding using images for visually-related concepts.

**Zero shot learning (ZSL)** refers to recognizing a concept without having explicitly seen any examples of it. For example classifying a cat in an image without ever having seen (labeled) images of cats. This is an important problem to address as in a number of tasks such as visual object classification: it is prohibitively expensive to provide training examples for every imaginable object of interest.

There are two main types of ZSL — unimodal and multimodal. The unimodal ZSL looks at component parts or attributes of the object, such as phonemes to recognize an unheard word or visual attributes such as color, size, and shape to predict an unseen visual class [57]. The multimodal ZSL recognizes the objects in the primary modality through the help of the secondary one — in which the object has been seen. The multimodal version of ZSL is a problem facing non-parallel data by definition as the overlap of seen classes is different between the modalities.

Socher et al. [198] map image features to a conceptual word space and are able to classify seen and unseen concepts. The unseen concepts can be then assigned to a word that is close to the visual representation — this is enabled by the semantic space being trained on a separate dataset that has seen more concepts. Instead of learning a mapping from visual to concept space Frome et al. [64] learn a coordinated multimodal representation between concepts and images that allows for ZSL. Palatucci et al. [165] perform prediction of words people are thinking of based on functional magnetic resonance images, they show how it is possible to predict unseen words through the use of an intermediate semantic space. Lazaridou et al. [123] present a fast mapping method for ZSL by mapping extracted visual feature vectors to text-based vectors through a neural network.

### 7.3 Hybrid data

In the hybrid data setting two non-parallel modalities are bridged by a shared modality or a dataset (see Figure 3c). The most notable example is the Bridge Correlational Neural Network [174], which uses a pivot modality to learn coordinated multimodal representations in presence of non-parallel data. For example, for multilingual image captioning, the image modality would be paired with at least one caption in any language. Such methods have also been used to bridge languages that might not have parallel corpora but have access to a shared pivot language, such as for machine translation [154], [174] and document transliteration [104].

Instead of using a separate modality for bridging, some methods rely on existence of large datasets from a similar or related task to lead to better performance in a task that only contains limited annotated data. Socher and Fei-Fei [197] use the existence of large text corpora in order to guide image segmentation. While Hendricks et al. [81] use separately trained visual model and a language model to lead to a better image and video description system, for which only limited data is available.

### 7.4 Discussion

Multimodal co-learning allows for one modality to influence the training of another, exploiting the complementary information across modalities. It is important to note that co-learning is task independent and could be used to create better fusion, translation, and alignment models. This challenge is exemplified by algorithms such as co-training, multimodal representation learning, conceptual grounding, and zero shot learning (ZSL) and has found many applications in visual classification, action recognition, audio-visual speech recognition, and semantic similarity estimation.

## 8 Conclusion

Multimodal machine learning is a vibrant multi-disciplinary field which aims to build models that can process and relate information from multiple modalities. This paper surveyed recent advances in multimodal machine learning and presented them in a common taxonomy built upon five technical challenges faced by multimodal researchers: representation, translation, alignment, fusion, and co-learning. For each challenge, we presented taxonomic sub-classification that allows to understand the breath of the current multimodal research. Although the focus of this survey paper was primarily on the last decade of multimodal research, it is important to address future challenges with a knowledge of past achievements.

Moving forward, the proposed taxonomy gives researchers a framework to understand current research and identify understudied challenges for future research. We summarized each technical challenge with a discussion of future directions and research problems (see Sections 3.3, 4.3, 5.3, 6.3 and 7.4). We believe that all these aspects of multimodal research are needed if we want to build computers able to perceive, model and generate multimodal signals. One specific area of multimodal machine learning which seems to be under-studied is co-learning, where knowledge from one modality helps with modeling in another modality. This challenge is related to the concept of coordinated representations where each modality keeps its own representation but find a way to exchange and coordinate knowledge. We see these lines of research as promising directions for future research.

## References


[1] "TRECVID Multimedia Event Detection 2011 Evaluation," https://www.nist.gov/multimodal-information-group/trecvid-multimedia-event-detection-2011-evaluation, accessed: 2017-01-21.

[2] "YouTube statistics," https://www.youtube.com/yt/press/statistics.html (accessed Sept. 2016), accessed: 2016-09-30.

[3] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the Behavior of Visual Question Answering Models," in *EMNLP*, 2016.

[4] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, 2012.

[5] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *CVPR*, 2013.

[6] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural Module Networks," *CVPR*, 2016.

[7] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[8] X. Anguera, J. Luque, and C. Gracia, "Audio-to-text alignment for speech recognition with very limited resources." in *INTERSPEECH*, 2014.

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual question answering," in *ICCV*, 2015.

[10] R. Arora and K. Livescu, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," *ICASSP*, 2013.

[11] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, 2010.

[12] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation By Jointly Learning To Align and Translate," *ICLR*, 2014.

[14] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional Affect Recognition using Continuous Conditional Random Fields," in *IEEE FG*, 2013.

[15] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang, "Video In Sentences Out," in *Proc. of the Conference on Uncertainty in Artificial Intelligence*, 2012.

[16] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching Words and Pictures," *JMLR*, 2003.

[17] M. Baroni, "Grounding Distributional Semantics in the Visual World Grounding Distributional Semantics in the Visual World," *Language and Linguistics Compass*, 2016.

[18] L. W. Barsalou, "Grounded cognition," *Annual review of psychology*, 2008.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *TPAMI*, 2013.

[20] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures," *JAIR*, 2016.

[21] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh, "VizWiz: Nearly Real-Time Answers to Vvisual Questions," in *UIST*, 2010.

[22] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Computational learning theory*, 1998.

[23] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Weakly supervised action labeling in videos under ordering constraints," in *ECCV*, 2014.

[24] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid, "Weakly-Supervised Alignment of Video With Text," in *ICCV*, 2015.

[25] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *International Conference on Spoken Language*, 1996.

[26] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," *CVPR*, 1997.

[27] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *SIGGRAPH*, 1997.

[28] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data Fusion through Cross-modality Metric Learning using Similarity-Sensitive Hashing," in *CVPR*, 2010.

[29] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, 1993.

[30] E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran, "Distributional Semantics in Technicolor," in *ACL*, 2012.

[31] E. Bruni, N. K. Tran, and M. Baroni, "Multimodal Distributional Semantics," *JAIR*, 2014.

[32] S. S. Bucak, R. Jin, and A. K. Jain, "Multiple Kernel Learning for Visual Object Recognition: A Review," *TPAMI*, 2014.

[33] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep Visual-Semantic Hashing for Cross-Modal Retrieval," in *KDD*, 2016.

[34] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-Announcement," in *Int. Conf. on Methods and Techniques in Behavioral Research*, 2005.

[35] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," *LNCS*, 2008.

[36] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, Attend, and Spell: a Neural Network for Large Vocabulary Conversational Speech Recognition," in *ICASSP*, 2016.

[37] A. Chang, W. Monroe, M. Savva, C. Potts, and C. D. Manning, "Text to 3D Scene Generation with Rich Lexical Grounding," *ACL*, 2015.

[38] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning," *ICMI*, 2014.

[39] S. Chen and Q. Jin, "Multi-modal Dimensional Emotion Recognition Using Recurrent Neural Networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015.

[40] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[41] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NIPS*, 2015.

[42] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell, "Co-Adaptation of audio-visual speech and gesture classifiers," in *ICMI*, 2006.

[43] C. M. Christoudias, R. Urtasun, and T. Darrell, "Multi-view learning in the presence of view disagreement," in *UAI*, 2008.

[44] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv preprint arXiv:1609.03193*, 2016.

[45] P. Cosi, E. Caldognetto, K. Vagges, G. Mian, M. Contolini, C. per Le Ricerche, and C. di Fonetica, "Bimodal recognition experiments with recurrent neural networks," in *ICASSP*, 1994.

[46] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, "Movie/script: Alignment and parsing of video and text transcription," *ECCV*, 2008.

[47] B. Coyne and R. Sproat, "WordsEye: an automatic text-to-scene conversion system," in *SIGGRAPH*, 2001.

[48] F. De la Torre and J. F. Cohn, "Facial Expression Analysis," in *Guide to Visual Analysis of Humans: Looking at People*, 2011.

[49] S. Deena and A. Galata, "Speech-Driven Facial Animation Using a Shared Gaussian Process Latent Variable Model," in *Advances in Visual Computing*, 2009.

[50] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in *EACL*, 2014.

[51] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language Models for Image Captioning: The Quirks and What Works," *ACL*, 2015.

[52] S. K. D'mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems," *ACM Computing Surveys*, 2015.

[53] D. Elliott and F. Keller, "Image description using visual dependency representations," in *EMNLP*, 2013.

[54] ——, "Comparing automatic evaluation measures for image description," in *ACL*, 2014.

[55] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, 2013.

[56] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based Recurrent Neural Networks," in *INTERSPEECH*, 2014.

[57] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.

[58] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," *LNCS*, 2010.

[59] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," *CVPR*, 2016.

[60] F. Feng, R. Li, and X. Wang, "Deep correspondence restricted Boltzmann machine for cross-modal retrieval," *Neurocomputing*, 2015.

[61] F. Feng, X. Wang, and R. Li, "Cross-modal Retrieval with Correspondence Autoencoder," in *ACMMM*, 2014.

[62] Y. Feng and M. Lapata, "Visual Information in Semantic Representation," in *NAACL*, 2010.

[63] S. Fidler, A. Sharma, and R. Urtasun, "A Sentence is Worth a Thousand Pixels Holistic CRF model," in *CVPR*, 2013.

[64] A. Frome, G. Corrado, and J. Shlens, "DeViSE: A deep visual-semantic embedding model," *NIPS*, 2013.

[65] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," in *EMNLP*, 2016.

[66] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering," *NIPS*, 2015.

[67] A. Garg, V. Pavlovic, and J. M. Rehg, "Boosted learning in dynamic bayesian networks for multimodal speaker detection," *Proceedings of the IEEE*, 2003.

[68] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *TPAMI*, 2017.

[69] P. Gehler and S. Nowozin, "On Feature Combination for Multiclass Object Classification," in *ICCV*, 2009.

[70] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," in *NIPS*, 1996.

[71] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," *LNCS*, 2011.

[72] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010.

[73] M. Gönen and E. Alpaydın, "Multiple Kernel Learning Algorithms," *JMLR*, 2011.

[74] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[75] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[76] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," *ICCV*, 2013.

[77] A. Gupta, Y. Verma, and C. V. Jawahar, "Choosing Linguistics over Vision to Describe Images," in *AAAI*, 2012.

[78] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic Modality Weighting for Multi-stream HMMs in Audio-Visual Speech Recognition," in *ICMI*, 2008.

[79] D. R. Hardoon, S. Szedmak, and J. Shawe-taylor, "Canonical correlation analysis; An overview with application to learning methods," Tech. Rep., 2003.

[80] A. Haubold and J. R. Kender, "Alignment of speech to highly imperfect text transcriptions," in *ICME*, 2007.

[81] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, in *CVPR*, 2016.

[82] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, 2012.

[83] G. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmoltz free energy," in *NIPS*, 1993.

[84] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, 2006.

[85] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[86] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *JAIR*, 2013.

[87] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, 1936.

[88] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural Language Object Retrieval," in *CVPR*, 2016.

[89] J. Huang and B. Kingsbury, "Audio-Visual Deep Learning for Noise Robust Speech Recognition," in *ICASSP*, 2013.

[90] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra *et al.*, "Visual storytelling." NAACL, 2016.

[91] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *ICASSP*, 1996.

[92] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015.

[93] A. P. James and B. V. Dasarathy, "Medical image fusion : A survey of the state of the art," *Information Fusion*, vol. 19, 2014.

[94] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multi-task , Multi-Kernel Learning for Estimating Individual Wellbeing," in *Multimodal Machine Learning Workshop in conjunction with NIPS*, 2015.

[95] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the Long-Short Term Memory Model for Image Caption Generation," *ICCV*, 2015.

[96] Q.-y. Jiang and W.-j. Li, "Deep Cross-Modal Hashing," in *CVPR*, 2017.

[97] X. Jiang, F. Wu, Y. Zhang, S. Tang, W. Lu, and Y. Zhuang, "The classification of multi-modal data with hidden conditional random field," *Pattern Recognition Letters*, 2015.

[98] Q. Jin and J. Liang, "Video Description Generation using Audio and Visual Cues," in *ICMR*, 2016.

[99] B. H. Juang and L. R. Rabiner, "Hidden Markov Models for Speech Recognition," *Technometrics*, 1991.

[100] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulchere, V. Michalski, K. Konda, J. Sebastien, P. Froumenty, Y. Dauphin, N. Boulanger-

Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, 2015.

[101] N. Kalchbrenner and P. Blunsom, "Recurrent Continuous Translation Models," in *EMNLP*, 2013.

[102] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[103] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *NIPS*, 2014.

[104] M. M. Khapra, A. Kumaran, and P. Bhattacharyya, "Everybody loves a rich cousin: An empirical study of transliteration through bridge languages," in *NAACL*, 2010.

[105] D. Kiela and L. Bottou, "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics," *EMNLP*, 2014.

[106] D. Kiela, L. Bulat, and S. Clark, "Grounding Semantics in Olfactory Perception," in *ACL*, 2015.

[107] D. Kiela and S. Clark, "Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception," *EMNLP*, 2015.

[108] Y. Kim, H. Lee, and E. M. Provost, "Deep Learning for Robust Feature Generation in Audiovisual Emotion Recognition," in *ICASSP*, 2013.

[109] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ICLR*, 2015.

[110] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," *TACL*, 2015.

[111] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation," in *CVPR*, 2015.

[112] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *IJCV*, 2002.

[113] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler, "What are you talking about? Text-to-Image Coreference," in *CVPR*, 2014.

[114] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NIPS*, 2012.

[115] M. A. Krogel and T. Scheffer, "Multi-relational learning, text mining, and semi-supervised learning for functional genomics," *Machine Learning*, 2004.

[116] J. B. Kruskal, "An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules," *Society for Industrial and Applied Mathematics Review*, 1983.

[117] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," *TPAMI*, 2013.

[118] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, 2011.

[119] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *ACL*, 2012.

[120] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML*, 2001.

[121] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, 2000.

[122] Z. Z. Lan, L. Bao, S. I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia Tools and Applications*, 2014.

[123] A. Lazaridou, E. Bruni, and M. Baroni, "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world," in *ACL*, 2014.

[124] R. Lebret, P. O. Pinheiro, and R. Collobert, "Phrase-based Image Captioning," *ICML*, 2015.

[125] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using cotraining," in *ICCV*, 2003.

[126] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *CoNLL*. Association for Computational Linguistics, 2011.

[127] Y. Li, S. Wang, Q. Tian, and X. Ding, "A survey of recent advances in visual feature detection," *Neurocomputing*, 2015.

[128] R. W. Lienhart, "Comparison of automatic shot boundary detection algorithms," *Proceedings of SPIE*, 1998.

[129] C.-Y. Lin and E. Hovy, "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics," *NAACL*, 2003.

[130] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal Alzheimer's disease classification," *IEEE Journal of Biomedical and Health Informatics*, 2014.

[131] M. M. Louwerse, "Symbol interdependency in symbolic and embodied cognition," *Topics in Cognitive Science*, 2011.

[132] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[133] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Co-Attention for Visual Question Answering," in *NIPS*, 2016.

[134] B. Mahasseni and S. Todorovic, "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition," in *CVPR*, 2016.

[135] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *ICCV*, 2015.

[136] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, "What's cookin'? interpreting cooking videos using text, speech and vision," *NAACL*, 2015.

[137] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating Images from Captions with Attention," in *ICLR*, 2016.

[138] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and Comprehension of Unambiguous Object Descriptions," in *CVPR*, 2016.

[139] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep Captioning with multimodal recurrent neural networks (m-RNN)," *ICLR*, 2015.

[140] R. Mason and E. Charniak, "Nonparametric Method for Data-driven Image Captioning," in *ACL*, 2014.

[141] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, "Text-to-Visual Speech Synthesis Based on Parameter Generation from HMM," in *ICASSP*, 1998.

[142] B. McFee and G. R. G. Lanckriet, "Learning Multi-modal Similarity," *JMLR*, 2011.

[143] H. McGurk and J. Macdonald, "Hearing lips and seeing voices." *Nature*, 1976.

[144] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *IEEE International Conference on Multimedia and Expo*, 2010.

[145] H. Mei, M. Bansal, and M. R. Walter, "Listen, attend, and walk: Neural mapping of navigational instructions to action sequences," *AAAI*, 2016.

[146] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[147] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, A. Mensch, A. Berg, X. Han, T. Berg, and O. Health, "Midge: Generating Image Descriptions From Computer Vision Detections," in *EACL*, 2012.

[148] S. Moon, S. Kim, and H. Wang, "Multimodal Transfer Deep Learning for Audio-Visual Recognition," *NIPS Workshops*, 2015.

[149] E. Morvant, A. Habrard, and S. Ayache, "Majority vote of diverse classifiers for late fusion," *LNCS*, 2014.

[150] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for Audio-Visual Speech Recognition," in *ICASSP*, 2015.

[151] M. Müller, "Dynamic Time Warping," in *Information Retrieval for Music and Motion*, 2007.

[152] I. Naim, Y. Song, Q. Liu, L. Huang, H. Kautz, J. Luo, and D. Gildea, "Discriminative unsupervised alignment of natural language instructions with corresponding video segments," in *NAACL*, 2015.

[153] I. Naim, Y. C. Song, Q. Liu, H. Kautz, J. Luo, and D. Gildea, "Unsupervised Alignment of Natural Language Instructions with Video Segments," in *AAAI*, 2014.

[154] P. Nakov and H. T. Ng, "Improving statistical machine translation for a resource-poor language using related resource-rich languages," *JAIR*, 2012.

[155] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," *Interspeech*, vol. 2, 2002.

[156] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE TPAMI*, 2016.

[157] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *ICML*, 2011.

[158] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE TAC*, 2011.

[159] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *ICMI*, 2016.

[160] A. Noulas, G. Englebienne, and B. J. Kröse, "Multimodal Speaker diarization," *IEEE TPAMI*, 2012.

[161] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[162] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NIPS*, 2011.

[163] W. Ouyang, X. Chu, and X. Wang, "Multi-source Deep Learning for Human Pose Estimation," in *CVPR*, 2014.

[164] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually Indicated Sounds," in *CVPR*, 2016.

[165] M. Palatucci, G. E. Hinton, D. Pomerleau, and T. M. Mitchell, "Zero-Shot Learning with Semantic Output Codes," in *NIPS*, 2009.

[166] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly Modeling Embedding and Translation to Bridge Video and Language," in *CVPR*, 2016.

[167] K. Papineni, S. Roukos, T. Ward, and W.-j. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *ACL*, 2002.

[168] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," in *ICCV*, 2015.

[169] S. Poria, E. Cambria, and A. Gelbukh, "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis," *EMNLP*, 2015.

[170] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, 2003.

[171] T. Qin, T.-y. Liu, X.-d. Zhang, D.-s. Wang, and H. Li, "Global Ranking Using Continuous Conditional Random Fields," in *NIPS*, 2008.

[172] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields." *IEEE TPAMI*, vol. 29, 2007.

[173] S. S. Rajagopalan, L.-P. Morency, T. Baltrušaitis, and R. Goecke, "Extending Long Short-Term Memory for Multi-View Structured Learning," *ECCV*, 2016.

[174] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran, "Bridge Correlational Neural Networks for Multilingual Multimodal Representation Learning," in *NAACL*, 2015.

[175] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling Latent Discriminative Dynamic of Multi-Dimensional Affective Signals," in *ACII workshops*, 2011.

[176] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACMMM*, 2010.

[177] A. Ratnaparkhi, "Trainable methods for surface natural language generation," in *NAACL*, 2000.

[178] S. Reed, Z. Akata, X. Yan, L. Logeswaran, H. Lee, and B. Schiele, "Generative Adversarial Text to Image Synthesis," in *ICML*, 2016.

[179] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding Action Descriptions in Videos," *TACL*, 2013.

[180] S. Reiter, B. Schuller, and G. Rigoll, "Hidden Conditional Random Fields for Meeting Segmentation," *ICME*, 2007.

[181] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," in *German Conference on Pattern Recognition*, 2015.

[182] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, 2017.

[183] R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann Machines," in *International conference on artificial intelligence and statistics*, 2009.

[184] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, 2007.

[185] A. Sarkar, "Applying Co-Training methods to statistical parsing," in *ACL*, 2001.

[186] B. Schuller, M. F. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 – The First International Audio / Visual Emotion Challenge," in *ACII*, 2011.

[187] S. Shariat and V. Pavlovic, "Isotonic CCA for sequence alignment and activity recognition," in *ICCV*, 2011.

[188] E. Shutova, D. Kelia, and J. Maillard, "Black Holes and White Rabbits : Metaphor Identification with Visual Features," *NAACL*, 2016.

[189] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple Kernel Learning for Emotion Recognition in the Wild," *ICMI*, 2013.

[190] C. Silberer and M. Lapata, "Grounded Models of Semantic Representation," in *EMNLP*, 2012.

[191] ——, "Learning Grounded Meaning Representations with Autoencoders," in *ACL*, 2014.

[192] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *NIPS*, 2014.

[193] ——, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, 2015.

[194] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proceedings of Fonetik*, 2003.

[195] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *NIPS*, 2000.

[196] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, 2005.

[197] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *CVPR*, 2010.

[198] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng, "Zero-shot learning through cross-modal transfer," in *NIPS*, 2013.

[199] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded Compositional Semantics for Finding and Describing Images with Sentences," *TACL*, 2014.

[200] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE TAFFC*, 2012.

[201] Y. Song, L.-P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *CVPR*, 2012.

[202] ——, "Multimodal Human Behavior Analysis: Learning Correlation and Interaction Across Modalities," in *ICMI*, 2012.

[203] Y. C. Song, I. Naim, A. A. Mamun, K. Kulkarni, P. Singla, J. Luo, D. Gildea, and H. Kautz, "Unsupervised Alignment of Actions in Video with Text Descriptions," in *IJCAI*, 2016.

[204] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," *JMLR*, 2014.

[205] N. Srivastava and R. Salakhutdinov, "Learning Representations for Multimodal Data with Deep Belief Nets," in *ICML*, 2012.

[206] N. Srivastava and R. R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," in *NIPS*, 2012.

[207] H. I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, 2014.

[208] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *NIPS*, 2014.

[209] C. Sutton and A. McCallum, "Introduction to Conditional Random Fields for Relational Learning," in *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[210] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Aligning plot synopses to videos for story-based retrieval," *IJMIR*, 2015.

[211] ——, "Book2Movie: Aligning video scenes with book chapters," in *CVPR*, 2015.

[212] S. L. Taylor, M. Mahler, B.-j. Theobald, and I. Matthews, "Dynamic units of visual speech," in *SIGGRAPH*, 2012.

[213] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney, "Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild," in *COLING*, 2014.

[214] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," *arXiv preprint arXiv:1503.01070*, 2015.

[215] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller, "Deep canonical time warping," in *CVPR*, 2016.

[216] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.

[217] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013," in *ACM International Workshop on Audio/Visual Emotion Challenge*, 2013.
[218] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel Recurrent Neural Networks," *ICML*, 2016.
[219] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation Ramakrishna Vedantam," in *CVPR*, 2015.
[220] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-Embeddings of Images and Language," in *ICLR*, 2016.
[221] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving LSTM-based Video Description with Linguistic Knowledge Mined from Text," *EMNLP*, 2016.
[222] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating Videos to Natural Language Using Deep Recurrent Neural Networks," *NAACL*, 2015.
[223] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.
[224] S. Vogel, H. Ney, and C. Tillmann, "HMM-based word alignment in statistical translation," in *Computational Linguistics*, 1996.
[225] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep Multimodal Hashing with Orthogonal Regularization," in *IJCAI*, 2015.
[226] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *arXiv preprint arXiv:1408.2927*, 2014.
[227] L. Wang, Y. Li, and S. Lazebnik, "Learning Deep Structure-Preserving Image-Text Embeddings," in *CVPR*, 2016.
[228] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *ICML*, 2015.
[229] J. Weston, S. Bengio, and N. Usunier, "Web Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings Image Annotation," *ECML*, 2010.
[230] ——, "WSABIE: Scaling up to large vocabulary image annotation," in *IJCAI*, 2011.
[231] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *IMAVIS*, 2013.
[232] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling," *INTERSPEECH*, 2010.
[233] D. Wu and L. Shao, "Multimodal Dynamic Networks for Gesture Recognition," in *ACMMM*, 2014.
[234] Z. Wu, L. Cai, and H. Meng, "Multi-level Fusion of Audio and Visual Features for Speaker Identification," *Advances in Biometrics*, 2005.
[235] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring Inter-feature and Inter-class Relationships with Deep Neural Networks for Video Classification," in *ACMMM*, 2014.
[236] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," *ICML*, 2016.
[237] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," *ECCV*, 2016.
[238] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *ICML*, 2015.
[239] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *AAAI*, 2015.
[240] S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakici, "A Distributed Representation Based Query Expansion Approach for Image Captioning," in *ACL*, 2015.
[241] Y. Yang, C. L. Teo, H. Daume, and Y. Aloimonos, "Corpus-Guided Sentence Generation of Natural Images," in *EMNLP*, 2011.
[242] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," in *CVPR*, 2016.
[243] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu, "I2T: Image parsing to text description," *Proceedings of the IEEE*, 2010.
[244] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *CVPR*, 2015.
[245] Y.-r. Yeh, T.-c. Lin, Y.-y. Chung, and Y.-c. F. Wang, "A Novel Multiple Kernel Learning Framework for Heterogeneous Feature Fusion and Variable Selection," *IEEE Trans. Multimedia*, 2012.
[246] M. H. P. Young, A. Lai, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, 2014.
[247] C. Yu and D. Ballard, "On the Integration of Grounding Language and Lear ning Objects," in *AAAI*, 2004.
[248] H. Yu and J. M. Siskind, "Grounded Language Learning from Video Described with Sentences," in *ACL*, 2013.
[249] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," *CVPR*, 2016.
[250] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling Context in Referring Expressions," in *ECCV*, 2016.
[251] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Communications Magazine*, 1989.
[252] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, S. Krstulovi, and J. Latorre, "Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization," *IEEE Transactions on Audio, Speech & Language Processing*, 2012.
[253] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, 2009.
[254] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun, "Leveraging Video Descriptions to Learn Video Question Answering," in *AAAI*, 2017.
[255] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE TPAMI*, 2009.
[256] D. Zhang and W.-J. Li, "Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization," in *AAAI*, 2014.
[257] H. Zhang, Z. Hu, Y. Deng, M. Sachan, Z. Yan, and E. P. Xing, "Learning Concept Taxonomies from Multi-modal Data," in *ACL*, 2016.
[258] F. Zhou and F. De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *CVPR*, 2012.
[259] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," in *NIPS*, 2009.
[260] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books," in *ICCV*, 2015.
[261] C. L. Zitnick and D. Parikh, "Bringing semantics into focus using visual abstraction," in *CVPR*, 2013.

**Tadas Baltrušaitis** is a scientist in the Microsoft Corporation. His primary research interests lie in the automatic understanding of non-verbal human behaviour, computer vision, and multimodal machine learning. Before joining Microsoft, he post-doctoral associate at the Carnegie Mellon University. He received his Ph.D and Bachelor's degrees in Computer Science. His Ph.D research focused on automatic facial expression analysis in especially difficult real world settings.

**Chaitanya Ahuja** is a doctoral candidate in Language Technologies Institute in the School of Computer Science at Carnegie Mellon University. His interests range in various topics in natural language, computer vision, computational music and machine learning. Before starting with graduate school, Chaitanya completed his Bachelor's at Indian Institute of Technology, Kanpur.

**Louis-Philippe Morency** is an Assistant Professor in the Language Technology Institute at Carnegie Mellon University where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He was formerly research assistant professor in the Computer Sciences Department at University of Southern California and research scientist at USC Institute for Creative Technologies. Prof. Morency received his Ph.D. and Master degrees from MIT Computer Science and Artificial Intelligence Laboratory. His research focuses on building the computational foundations to enable computers with the abilities to analyze, recognize and predict subtle human communicative behaviors during social interactions. He is currently chair of the advisory committee for ACM International Conference on Multimodal Interaction and associate editor at IEEE Transactions on Affective Computing.