Toward Artificial Synesthesia: Linking Images and Sounds via Words

Han Xiao, Thomas Stibor
Department of Informatics
Technical University of Munich
Garching, D-85748
{xiaoh,stibor}@in.tum.de

Abstract

We tackle a new challenge of modeling a perceptual experience in which a stimulus in one modality gives rise to an experience in a different sensory modality, termed synesthesia. To meet the challenge, we propose a probabilistic framework based on graphical models that enables to link visual modalities and auditory modalities via natural language text. An online prototype system is developed for allowing human judgement to evaluate the model's performance. Experimental results indicate usefulness and applicability of the framework.

1 Introduction

A picture of a golden beach might stimulate human's hearing, probably, by imagining the sound of waves crashing against the shore. On the other hand, the sound of a baaing sheep might illustrate a green hillside in front of your eyes. In neurology, this kind of experience is termed *synesthesia*. That is, a perceptual experience in which a stimulus in one modality gives rise to an experience in a different sensory modality. Without a doubt, the creative process of humans (e.g. painting and composing) is to a large extent attributed to their synesthesia experiences. While cross-sensory links such as sound and vision are quite common to humans, machines do not possess the same ability naturally. Nevertheless, synesthetic perception is never a mysterious term for machines as it is for psychologists and neuroscientists. Images and sounds represent distinct modalities, yet both modalities capture the same underlying concepts as they were used to describe the same objects. In this paper, we are aiming to associate images and sounds using a multi-modality model.

Before contemplating the problem of multi-modal modeling, we illustrate the links between images and sounds in Figure 1. Loosely speaking, there are two types of links between images and sounds, namely *explicit linking* and *implicit linking*. Explicit linking happens ubiquitously. For instance, for those who have been to the sea, it is easy to associate the sound of waves with the picture of blue sea. On the other hand, implicit linking is more sophisticated. Assume you are aware that J.S. Bach was a grand violinist and you know how a violin sounds like. Yet, you have never heard Bach playing violin personally. Now by showing you a portrait of Bach, the sound of violin might involuntarily ring in your ears. The major difference between these two links is: the correspondence between image and sound is observed in first case, whereas image and sound are not directly associated in the second case, they are linked together by another intermediate but obscure modality.

As natural language is based on visual and auditive stimuli, we believe text is a reasonable and effective intermediate modality to bridge the gap between images and sounds. Seen from the perspective of machine learning, implicit linking via text stimulates particular interests for the following reasons.

- Implicit linking makes full use of the data resources. In particular, an implicit linking model can be trained on three separate data sets, i.e. images/text, sounds/text and text. These three type of data can be easily acquired from the web. By contrast, an explicit linking model needs aligned images/sounds data for training. That is, one has to collect a set of images, each of which corresponds to a collection of associated sounds. Unfortunately, a high-quality images/sounds data set is scarce and expensive. Hence the explicit linking model is limited realizable due to the lack of corresponding information between two modalities. On the other hand, the implicit

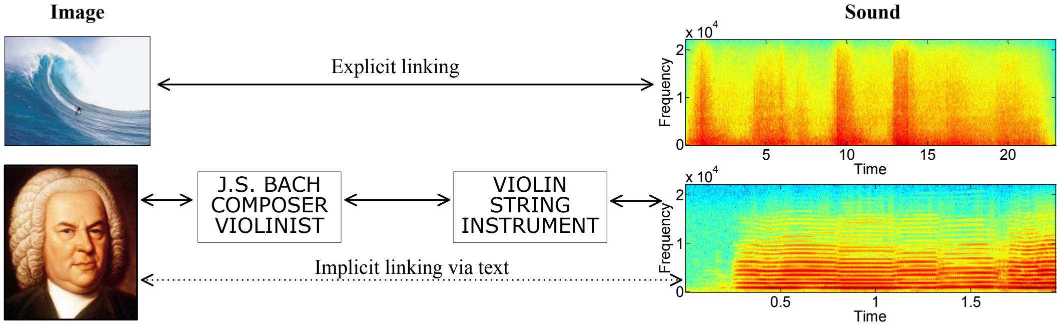


Figure 1: The upper part of the figure illustrates explicit linking, where the image and sound are linked together directly. Sounds are visualized as spectrogram using the short-time Fourier transform. The lower part consists of two entries from our data set: a captioned portrait of J.S Bach (left side) and a captioned sound snippet of violin (right side). The captions are shown in uppercase. Note, that our data set does not contain the correspondence between images and sounds.

linking provides an approach to model two modalities in an indirect manner by leveraging an intermediate modality.

- Implicit linking is more likely to capture the subtle association between images and sounds. For instance, we feed pairwise pictures and sounds of a violin to train an explicit linking model. Given a new sound snippet of a violin, the model is unlikely to link it with the portrait of J.S.Bach. On contrary, an implicit linking model can propagate the relatedness crossing three modalities: from sound to text (violin's sound → "violin"), from text to text ("violin"→ "violinist"), and finally from text to image ("violinist" → Bach's portrait). Therefore, an implicit linking model can achieve a more comprehensive synesthesia than an explicit model.
- Natural language relies on the process of semiosis to relate a sign with a particular meaning. Jointly modeling visual and auditory information enables us to gain insight into the language itself, for instance by studying the following problems: how are words or morphemes related to sensory information? How is the syntactic system concatenating words into different phrases and sentences under different scenes? By exploring the implicit linking, we might use the methodology of machine learning research to answer the above questions originated in the field of cognitive science and neurolinguistics.

We intend to link images and sounds in an implicit manner. In particular, using natural language text as an intermediate representation for both visual and auditory modality and bringing them together. Our motivation is that the natural language reveals the underlying concepts in both visual and auditory modality, meanwhile encompasses the semantic relations of polysemy and synonymy, which suggests a bridge between images and sounds. A complete matching process follows three steps: translating the original modality into text, analyzing the text, and translating the text to target modality. The problem we focus on can be described in two ways. First, one might attempt to predict sounds given an image, where sounds should be either directly (approximately matching with visual objects) or indirectly (a reasonable synesthesia stimulated by the scene) related with the image. We refer to this task as *image composition*. Secondly, one might attempt to predict images that either directly or indirectly relate to the given sound, which is denoted as *sound illustration* in this paper.

There are several practical applications that derive from image composition and sound illustration. For example, a digital photo management software with an image composition plugin can automatically link suitable sound effects for every picture in the album, which will greatly enrich the user experiences. The art museum can also exploit image composition to attract visitors by giving them environmental sounds of what they currently see. Moreover, image composition and sound illustration can also be used to provide an assisted multimedia context for people with disabilities like blindness and deafness.

We explore and exploit probabilistic topic models, such as latent Dirichlet allocation (LDA) [3] to model the implicit links. Probabilistic topic models find a low dimensional representation of data under the assumption that each datum can exhibit multiple "topics". This idea has been successfully adapted and imported to many computer vision problems [1, 2, 5, 8]. In this paper, we develop a probabilistic framework that exploits LDA and correspondence-LDA models (Corr-LDA) [2] to perform image composition and sound illustration simultaneously. For the sake of clarity the paper is structured as follows: Previous works on multi-modal modeling are briefly reviewed in Sect. 2. In Sect. 3, we describe the input representation of images and sounds as well as the preprocessing step. Sect. 4 formulates the image composition and sound illustration tasks in a probabilistic framework, and introduces our approach of jointly modeling images, sounds and text. Experimental results are illustrated in Sect.5. Sect. 6 concludes.

2 Related Work

A number of papers have considered probabilistic models for multi-modal data, especially for modeling images and text. As linking an image with associated text is extremely useful in image annotation, multi-media information retrieval and object recognition, manifold models are proposed. The co-occurrence model allows to compute in a straightforward manner the probability of associating words with image grids [10]. Inspired by the techniques in machine translation [4], one can consider images and text as two different languages. Thus, linking images and words can be viewed as a process of translating from visual vocabulary to textual vocabulary [5, 12]. Leveraging on the bags-of-words representation of images and text, many approaches originated in the field of text modeling such as: Hofmann's hierarchical aspect model [11], translation model [4] and latent Dirichlet allocation (LDA) model [3]. These models were extended for predicting words from images [1]. LDA was further extended to correspondence-LDA (Corr-LDA) to model the generative process of image regions and words in the same latent space [2]. Additionally, a supervised extension was proposed to perform classification [23].

In another line of research, modeling text and audio focused on music classification of genre, emotion, and instrumentation for text-based music information retrieval [7, 14, 22]. These approaches classify music and "tag" them with class labels (e.g., "pop", "jazz", "blues") from a limited textual vocabulary. More recently, several approaches have been developed to annotate music with a larger and more diverse vocabulary of tags [6, 17, 20].

Our work can be viewed as a combination of multi-modal modeling, information retrieval and natural language processing. The contribution of this paper is threefold. First, to the best of our knowledge the idea of artificial synesthesia and cross-sensory implicit linking have not been well explored in the field of machine learning. Second, we leverage an intermediate modality, that is text, to bridge the gap between images and sounds, which differs from ordinary approaches based on explicit linking. Third, we represent images, text and sounds in a generic probabilistic framework, which provides a clean, solid and extensible infrastructure.

3 Input Representation and Preprocessing

In this section, we briefly introduce the preprocessing step for images and sounds. The goal is to build a visual vocabulary and an auditory vocabulary for representing images and sounds as bags-of-words.

3.1 Image Representation

Following previous work [8], we represent each image as a set of visual words. Here, *visual words* are defined as the centroids of learnt clusters using k-means algorithm. To obtain visual words, we compute the dense SIFT descriptors for each image [13, 16]. Thus, each image is represented as a set of 128 dimensional SIFT descriptors. We then quantize all SIFT descriptors in the collection using k-means algorithm to obtain centroids of learnt clusters, which compose the visual vocabulary for images. Finally, each visual word is assigned a unique integer to serve as its identifier, and the SIFT descriptors are mapped to their corresponding nearest visual word.

3.2 Sound Representation

Each sound snippet is cut into frames, where a frame refers to a sequence of 1024 audio samples. For each frame, we compute the 13 dimensional Mel-Frequency Cepstral Coefficients and 6 groups of widely used statistics (mean and standard deviation) [21]: Energy Entropy, Signal Energy, Zero Crossing Rate, Spectral Rolloff, Spectral Centroid and Spectral Flux. Thus, each sound snippet is represented as a set of 25 dimensional feature vectors. Similar to the preprocessing step of images, all feature vectors in the collection are clustered using k-means algorithm to obtain auditory words. At last, the feature vectors are mapped to their corresponding nearest auditory word.

3.3 Notations

Assuming a training collection $\mathbb{T}$ consists of K annotated images and L tagged sounds $\mathbb{T} = \{\mathbf{I}_1, \ldots, \mathbf{I}_K; \mathbf{S}_1, \ldots, \mathbf{S}_L\}$, we can now unify the notation of images, sounds and their corresponding text as follows:

- An annotated image $\mathbf{I} \in \mathbb{T}$ has a dual representation in terms of visual words and textual words: $\mathbf{I} = \{v_1, \ldots, v_M; w_1, \ldots, w_N\}$. Here $\{v_1, \ldots, v_M\}$ represents the M visual words of $\mathbf{I}$ and $\{w_1, \ldots, w_N\}$ represents N words in the annotations of $\mathbf{I}$.
- A captioned sound snippet $\mathbf{S} \in \mathbb{T}$ has a dual representation in terms of auditory words and textual words: $\mathbf{S} = \{u_1, \ldots, u_M; w_1, \ldots, w_N\}$, Here $\{u_1, \ldots, u_M\}$ represents the M auditory words of $\mathbf{S}$ and $\{w_1, \ldots, w_N\}$ represents N words in the tags of $\mathbf{S}$.

In addition, we define $\mathbf{W}^{\mathrm{i}}$ as the vocabulary of image annotations and $\mathbf{W}^{\mathrm{s}}$ as the vocabulary of sound tags. The complete textual vocabulary is denoted as $\mathbf{W} = \mathbf{W}^{\mathrm{i}} \cup \mathbf{W}^{\mathrm{s}}$.

4 Linking Images and Sounds via Text

An overview of the probabilistic framework for performing the image composition task and sound illustration task is depicted in Figure 2. Following the notations in Section 3.3, these two tasks can be formulated as follows:

Image composition Given an un-annotated image $\mathbf{I}^* \notin \mathbb{T}$, estimate the conditional probability $p(\mathbf{S}|\mathbf{I}^*)$ for every $\mathbf{S} \in \mathbb{T}$. To compose a sound effect, one can pick the sound snippets with the highest probability under $p(\mathbf{S}|\mathbf{I}^*)$ and mix them together.

Sound illustration Given an un-tagged sound $\mathbf{S}^* \notin \mathbb{T}$, estimate the conditional probability $p(\mathbf{I}|\mathbf{S}^*)$ for every $\mathbf{I} \in \mathbb{T}$. The visual scene that best matches the given sound is the image with highest probability under $p(\mathbf{I}|\mathbf{S}^*)$.

Since we can not estimate $p(\mathbf{S}|\mathbf{I}^*)$ and $p(\mathbf{I}|\mathbf{S}^*)$ directly, as there are no explicit correspondences between images and sounds in our data set, the only bridge we can take advantage of is the text in the captioned images and sounds. An intuitive way is to first “translate” the image into natural language text, and then “translate” the text back into sound. Therefore, the conditional probabilities can be approximated as:

$$p(\mathbf{S}|\mathbf{I}^*) \approx \sum_{w \in \mathbf{W}^{\mathrm{i}}} \sum_{w' \in \mathbf{W}^{\mathrm{s}}} p(\mathbf{S}|w') p(w'|w) p(w|\mathbf{I}^*), \tag{1}$$

$$p(\mathbf{I}|\mathbf{S}^*) \approx \sum_{w \in \mathbf{W}^{\mathrm{s}}} \sum_{w' \in \mathbf{W}^{\mathrm{i}}} p(\mathbf{I}|w') p(w'|w) p(w|\mathbf{S}^*). \tag{2}$$

One can observe, that the two approximations have an equivalent representation. As a consequence, we can first focus on the image composition task $p(\mathbf{S}|\mathbf{I}^*)$ and later apply the algorithm to the sound illustration task $p(\mathbf{I}|\mathbf{S}^*)$ straightforwardly. The conditional probability (1) consists of three parts and crosses three modalities, which implies three different models. Fortunately, as images and sounds have been converted to the same representation as shown in Section 3, we can deal with $p(\mathbf{S}|w)$ and $p(w|\mathbf{I})$ using the same model. As we shall see, $p(\mathbf{S}|w)$ and $p(w|\mathbf{I})$ can be derived from Corr-LDA model, and $p(w'|w)$ can be obtained from LDA and a lexical database.

4.1 Modeling Images/Text and Sounds/Text

Our approach is based on the Corr-LDA model proposed in [2]. We modify the Corr-LDA and apply it to model images/text and sounds/text. As images and sounds have been represented in an equivalent form, we hereinafter introduce our approach by taking images/text as example. Formally, fixing the number of topics T, the generative process of an annotated image $\mathbf{I} = \{v_1, \ldots, v_M; w_1, \ldots, w_N\}$ is described as follows:

1. Draw topic proportions $\theta \sim \mathrm{Dirichlet}(\alpha)$
2. For each visual word $v_m, m \in \{1, \ldots, M\}$
   (a) Draw topic assignment $z_m|\theta \sim \mathrm{Multinomial}(\theta)$
   (b) Draw visual word $v_m|z_m \sim \mathrm{Multinomial}(\pi_{z_m})$
3. For each textual word $w_n, n \in \{1, \ldots, N\}$
   (a) Draw discrete indexing variable $y_n \sim \mathrm{Uniform}(1, \ldots, M)$
   (b) Draw textual word $w_n \sim \mathrm{Multinomial}(\beta_{z_{y_n}})$

The graphical model of Corr-LDA is depicted in Figure 3. Two remarks need to be highlighted here. First, we use a multinomial distribution to generate a visual word in step 2.b, whereas the original model used a multivariate gaussian to generate image regions [2]. This slight difference

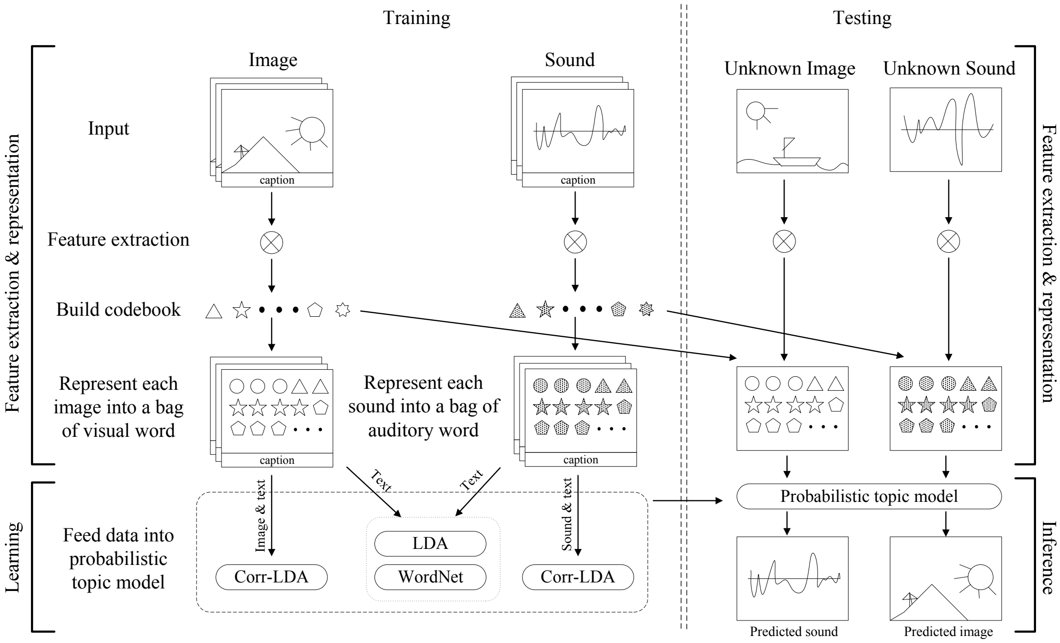
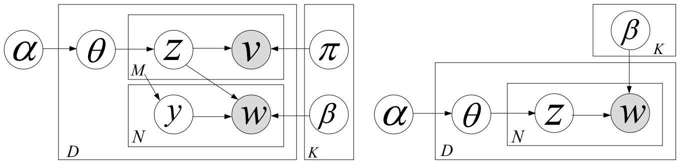


Figure 2: Probabilistic framework for performing the image composition and sound illustration task. The framework is an extension based on the work flow proposed in [8]. Images and sounds are represented in bags-of-words, so that the difference between the two modalities can be omitted. Once we have the algorithm for inferring sounds from an image, we can apply it to infer images from a sound by mirroring the algorithm.

can be attributed to the quantization of feature vectors in our preprocessing step. As a visual word is a discrete indexing variable rather than a high dimensional vector, it makes sense to use the multinomial distribution instead. This modification leads to a variant of the variational inference of Corr-LDA. Second, by simply replacing the visual word v_m by an auditory word u_m, the same generative process can be used to model sound and text. With a trained model in hand, we can



(a) Corr-LDA

(b) LDA

word	prob.
NATURE	.91
THUNDER	.85
BIRD	.85
WIND	.83
FIELD	.80
RECORD	.80
GARDEN	.76

(c) $p_{\text{LDA}}(w|\,\text{rain})$

word	prob.
RAINFALL	.99
SNOW	.92
SEQUENCE	.91
DOWNPOUR	.90
WATER	.89
LIGUID	.89
HAIL	.83

(d) $p_{\text{WordNet}}(w|\,\text{rain})$

Figure 3: (a) A graphical model representation of Corr-LDA. Nodes represent random variables; shaded nodes are observed random variables, unshaded nodes are latent random variables. Edges denote possible dependence between random variables; plates denote replicated structure. Note that the variables y_n are conditioned on M. Corr-LDA is used to model the correspondence between images and text, as well as sounds and text. (b) A graphical model representation of LDA, which is used to model the word relatedness in this paper. (c,d) Exemplary outputs the word relatedness of LDA and WordNet. Seven words with the highest probabilities under $p(w|\text{rain})$ are listed. One can observe, that LDA and WordNet capture the synonyms effectively, yet LDA's result relies more on the quality of the data set. Binding the two models together will make the relatedness measurement more robust.

compute the conditional distributions of interest: $p(\mathbf{I}|w)$ and $p(w|\mathbf{I}^*)$, where $\mathbf{I} \in \mathbb{T}$ and $\mathbf{I}^* \notin \mathbb{T}$. In particular, the distribution over words conditioned on an unseen image is approximated by:

$$p(w|\mathbf{I}^*) \approx \sum_{m=1}^{M} \sum_{z_m} p(z_m|\theta) p(w|z_m, \beta). \tag{3}$$

Moreover, we can rewrite $p(\mathbf{I}|w)$ using Bayes rule as:

$$p(\mathbf{I}|w) = \frac{p(w|\mathbf{I})p(\mathbf{I})}{\sum_{\mathbf{I'}\in\mathbb{T}} p(w|\mathbf{I'})p(\mathbf{I'})}, \tag{4}$$

where $p(\mathbf{I})$ can be computed as follows:

$$p(\mathbf{I}) = p(\theta|\alpha) \prod_{m=1}^{M} p(z_m|\theta)p(v_m|z_m, \pi) \prod_{n=1}^{N} p(y_n|M)p(w_n|z_{y_n}, \beta) \tag{5}$$

By plugging (3) and (5) into (4), we can estimate $p(\mathbf{I}|w)$ for every word and image in the training set. For the sake of completeness, we release a technical note which includes the detailed derivation of the variational inference and the parameter estimation algorithm on the web[1].

4.2 Modeling Text

The remaining problem is to estimate $p(w'|w)$ from the training set, which is actually measuring the semantic relatedness between two words. We make use of the LDA model [3] to solve this problem. To apply LDA on our data set, we build another data set $\mathbb{D}$ containing only captions of all images and sounds in $\mathbb{T}$, where $|\mathbb{D}| = |\mathbb{T}|$. The generative process of a document $D \in \mathbb{D}$ is described as follows:

1. Draw topic proportions $\theta \sim \mathrm{Dirichlet}(\alpha)$
2. For each textual word $w_n, n \in \{1, \ldots, N\}$
   (a) Draw topic assignment $z_n|\theta \sim \mathrm{Multinomial}(\theta)$
   (b) Draw textual word $w_n|z_n \sim \mathrm{Multinomial}(\beta_{z_n})$

The graphical model of LDA is depicted in Figure 3(b). The mixing proportion over topics $\theta_D = p(z|D)$ and the word distribution over topics $\beta = p(w|z)$ are two sets of parameters that need to be estimated from the training set. The LDA model can be trained by three different algorithms: variational Expectation-Maximization [3], Expectation-Propagation [18] and collapsed Gibbs sampling [9]. Given a trained LDA model, the word relatedness between w and w' can be calculated by:

$$p_{\mathrm{LDA}}(w|w') = \frac{1}{\mathcal{C}} \sum_{z_n} p(w|z_n) \frac{n_{w'}}{n_{z_n}} p(w'|z_n), \tag{6}$$

where $n_{w'}$ is the number of w' occurred in $\mathbb{D}$, n_{z_n} is the number of words assigned to topic z_n. $\mathcal{C}$ is a normalization factor to scale the relatedness to $[0, 1]$. Note, however, that the relatedness is calculated on a small data set with limited scope, it might not reflect the ground-truth of semantic similarity. With this issue in mind, we avail ourselves of the WordNet dictionary[2] to smooth $p(w|w')$. We measure the semantic relatedness defined in [15] for every two words. Due to the limit of pages, we omit the details of the algorithm and denote the result from WordNet similarity as $p_{\mathrm{WordNet}}(w|w')$, which is also in the range 0 to 1. An example of relatedness measurements based on LDA and WordNet is depicted in Figure 3(c,d), where we find both algorithms to give reasonable output, yet differ from each other. Therefore, the final relatedness is defined as a mixture of two probabilities:

$$p(w|w') = \sigma\, p_{\mathrm{LDA}}(w|w') + (1 - \sigma) p_{\mathrm{WordNet}}(w|w'), \tag{7}$$

where σ is the smoothing parameter.

In summary, calculating $p(\mathbf{S}|\mathbf{I}^*)$ boils down to two problems. First, to estimate the probabilities $p(w|\mathbf{I}^*), p(\mathbf{S}|w')$ according to (3) and (4) respectively, which we obtain from the Corr-LDA model. Second, to estimate $p(w|w')$ according to (7) which we obtain from the LDA model. Plugging (3), (4) and (7) into (1) and (2), we finally obtain the conditional probabilities of interest.

5 Experimental Results

In this section we will discuss details of the data set used and show experimental results using our approach. Due to the objective difficulties for evaluating synesthesia, the evaluation is mainly performed in a qualitative manner. We will introduce an online system we built for allowing users to identify the predicted sounds/images interactively. Finally, some examples are demonstrated to illustrate different aspects of our approach.

5.1 Data set

For the images/text data set, we downloaded three classes of images from the LabelMe data set [19], namely “street”, “coast”, “forest” and then randomly selected 300 images for each class. The total

[1] `http://home.in.tum.de/~xiaoh/pub/derivation.pdf`

[2] WordNet is a lexical database for the English language (`http://wordnet.princeton.edu/`).

number of images is 900. For each image, the average length of annotations is 7 tokens. The textual vocabulary size of all annotations is 156.

For sounds/text, we downloaded 831 audio snippets from *The Freesound Project*[3], where most of them are natural sounds and synthetic sound effect. The duration of these sound snippets range from 2 seconds to 10 minutes. All sound snippets are converted to 44.1kHz mono WAV format. Each snippet is tagged by the uploader or other online users. The average number of tags per sound is 6 tokens. The textual vocabulary size of sound tags is 1576. An example of an abridged entry from our data set is shown in Figure 1.

We held out 20% of the data for testing purposes and used the remaining 80% to estimate parameters. Our goal is to train first the Corr-LDA model with annotated images, and second the Corr-LDA model with tagged sounds, as well as the LDA model with all annotations and tags. The image composition and sound illustration tasks are performed on un-annotated images and un-tagged sounds, respectively.

5.2 Model Parameters

For computing SIFT descriptors, the size of a patch is set to 16×16, the distance between grid centers is 10. By clustering SIFT descriptors and audio feature vectors, respectively, we obtained 241 visual words and 89 auditory words in total (clusters with less than 5 members are pruned out). The Dirichlet prior α of Corr-LDA and LDA is fixed to 0.1. The number of topic is 40 for both Corr-LDA and LDA. The maximum number of iterations for variational inference and EM algorithm is set to 100. The smoothing parameter σ is set to 0.8.

5.3 Online Evaluation System

Evaluating the performance of the image composition and sound illustration task is difficult for two reasons. First, our data set does not contain the correspondence information between images and sounds. Moreover, a high quality images/sounds data set is scarce and expensive. Therefore, we lack a gold-standard list of associated images or sounds to compare against. Second, the image-sound synesthesia differs from person to person, and as a consequence the judgments from two or three people may not truly reflect the model's performance. Thus, evaluating the image-sound synesthesia in a meaningful manner, requires gathering of exogenous data.

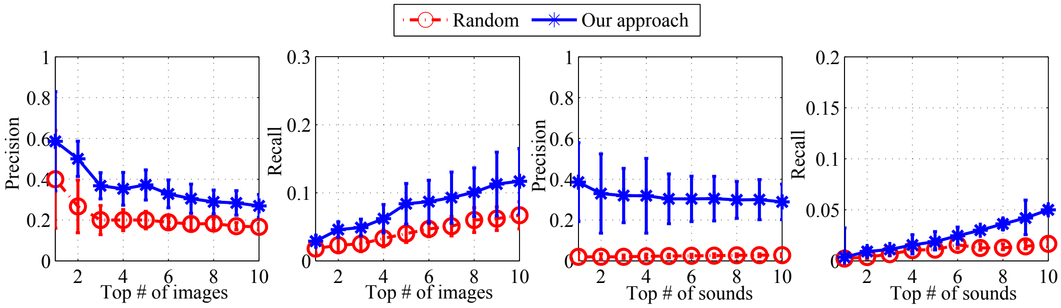
We developed an online evaluation system[4] that allows humans to judge the predicted sounds/images of a randomly given scene. For the image composition task, the webpage will randomly draw an image from the test set as the scene. Meanwhile, ten sound snippets with highest probabilities under $p(\mathbf{S}|\mathbf{I}^*)$ are provided. Users can listen to the snippets and decide whether the sounds are acceptable or not. For the sound illustration task, the webpage will randomly present a sound from the test set as the scene, and provide ten images with highest probabilities under $p(\mathbf{I}|\mathbf{S}^*)$. In both tasks, subjects must identify sounds or images related to the given scene. Occasionally, the system randomly draws images or sounds from the data set as intruders and demonstrates them to the subjects. Decisions from subjects are counted for evaluating the model's performance in terms of precision and recall. We invited 10 people from all walks of life for evaluating the result. As depicted in Fig. 4, our approach achieves more meaningful result of images/sounds association than the random baseline.

5.4 Illustrative Examples

To demonstrate the model's performance, a good prediction and a bad prediction of each task is illustrated in Figure 5. By observing Figure 5(a), one can see that there are 4 out of 5 sounds $(1, 2, 4, 5)$ that are highly related to the left hand picture. Our model successfully made a reasonable synesthesia by highlighting the sounds related with water. In Figure 5(b), 4 out of 5 responsive images $(1, 2, 3, 4$ are images about "car, street") are related to the sound "vehicles passing". Nevertheless, we also notice that our model fails to produce meaningful synesthesia in some cases. Consider for instance the right-hand side of Figure 5(a). Except the "wood stick breaking" sound, one can hardly relate sounds $(2, 3, 4, 5)$ with the scene. At the bottom of Figure 5(b), only the last image is related to the query sound "waterfall flowing".

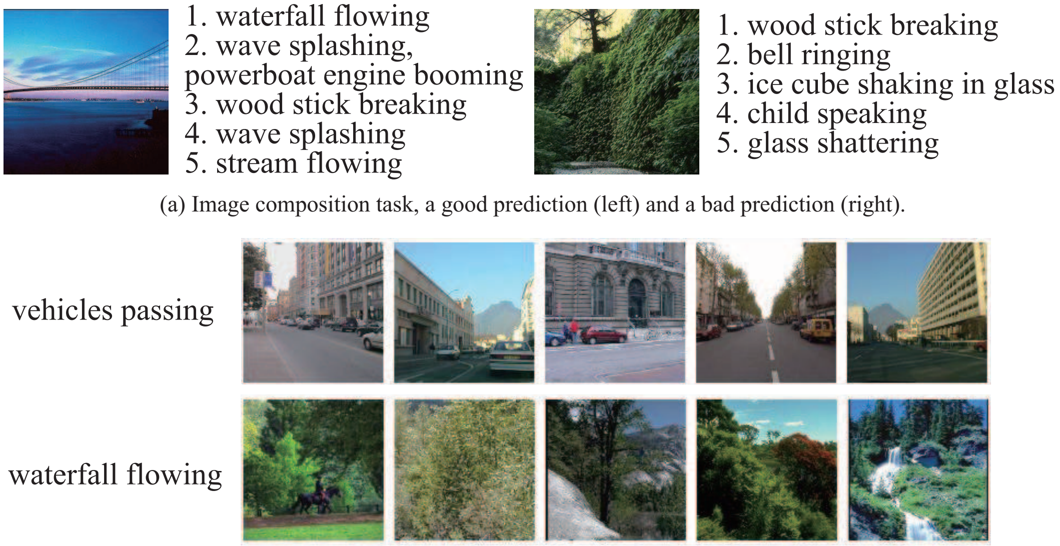
[3] `http://www.freesound.org/`

[4] `http://yulei.appspot.com/`



(a) Image composition task. (b) Sound illustration task.

Figure 4: Given the top-10 predicted images/sounds, the top-N precision and recall of our approach and a random baseline. (a) In image composition task, F-score@10 of our approach and the random baseline is 0.17 and 0.09, respectively. (b) In sound illustration task, F-score@10 of our approach and the random baseline is 0.10 and 0.02, respectively.



1. waterfall flowing
2. wave splashing, powerboat engine booming
3. wood stick breaking
4. wave splashing
5. stream flowing

1. wood stick breaking
2. bell ringing
3. ice cube shaking in glass
4. child speaking
5. glass shattering

(a) Image composition task, a good prediction (left) and a bad prediction (right).

(b) Sound illustration task, a good prediction (top) and a bad prediction (bottom). The images are ranked according to the conditional probabilities from highest probability (left most) to smallest probability (right most).

Figure 5: Example of good prediction and bad prediction of the synesthesia system. (a) The result of the image composition task, where two un-annotated images and five predicted sounds are depicted. (b) The result of the sound illustration task, where the un-tagged sounds and five predicted images are showed. Due to the difficulties to illustrate sounds on paper, we list the predicted sound snippets and manually give them captions.

6 Conclusions and Future Work

We have developed a probabilistic framework to tackle a new challenge called artificial synesthesia. The framework is based on latent Dirichlet models and enables the implicit linking of images and sounds via text. Conducted experiments showed usefulness and applicability on real-world data sets. Furthermore, an online prototype system has been developed to enable humans to evaluate the model's performance.

It has not escaped our notice, that the performance of Corr-LDA is varying on different data sets. In particular, Corr-LDA has difficulties to effectively explore the latent space of images with clutter. We are currently studying other graphical models to address this problem. As our proposed framework is based on probabilistic models, new models can be straightforwardly plugged into our framework.

Our future goal for the image composition task, is to explore a suitable and elegant way of mixing relevant sounds into a single track and compose a lifelike environmental sound effect. For sound illustration, our ideal goal is to automatically paint a single collage by selecting segments from relevant images. Other areas of possible research include using natural language sentences rather than words as a bridge to link visual and auditory modalities.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

[2] D.M. Blei and M.I. Jordan. Modeling annotated data. In *SIGIR*, pages 127–134. ACM, 2003.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[5] P. Duygulu, K. Barnard, J. De Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV*, pages 349–354, 2002.

[6] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. *NIPS*, 20:385–392, 2007.

[7] S. Essid, G. Richard, and B. David. Inferring Efficient Hierarchical Taxonomies for MIR Tasks: Application to Musical Instruments. In *ISMIR*, pages 324–328, 2005.

[8] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005.

[9] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228, 2004.

[10] Y.M. Hironobu, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Manegement*, 1999.

[11] T. Hofmann. Learning and Representing Topic. A Hierarchical Mixture Model for Word Occurrences in Document Databases. In *Proceedings of the Conference for Automated Learning and Discovery*, 1998.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, pages 119–126. ACM, 2003.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006.

[14] T. Li and G. Tzanetakis. Factors in automatic musical genre classification of audio signals. In *ICASSP*, pages 143–146. IEEE, 2004.

[15] D. Lin. An information-theoretic definition of similarity. In *ICML*, pages 296–304. Morgan Kaufmann Publishers Inc., 1998.

[16] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman. SIFT flow: dense correspondence across different scenes. *ECCV*, pages 28–42, 2008.

[17] M. I. Mandel and D.P.W Ellis. Multiple-instance learning for music information retrieval. In *ISMIR*, pages 577–582, 2008.

[18] T.P. Minka and J.D Lafferty. Expectation-propogation for the generative aspect model. In *UAI*, pages 352–359, 2002.

[19] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008.

[20] M. Sordo, C. Laurier, and O. Celma. Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 531–534, 2007.

[21] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, 3rd Edition*. Academic Press, 2006.

[22] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[23] C. Wang, D. Blei, and F.F. Li. Simultaneous image classification and annotation. In *CVPR*, pages 1903–1910. IEEE, 2009.