Learning Like a Toddler: Watching Television Series to Learn Vocabulary from Images and Audio

Emre Yılmaz
ESAT-PSI
KU Leuven, Belgium
eyilmaz@esat.kuleuven.be

Konstantinos Rematas
ESAT-PSI, iMinds
KU Leuven, Belgium
krematas@esat.kuleuven.be

Tinne Tuytelaars
ESAT-PSI, iMinds
KU Leuven, Belgium
tinne.tuytelaars@esat.kuleuven.be

Hugo Van hamme
ESAT-PSI
KU Leuven, Belgium
hugo.vanhamme@esat.kuleuven.be

ABSTRACT

This paper presents the initial findings of our efforts to build an unsupervised multimodal vocabulary learning scheme in a realistic scenario. For this purpose, a new multimodal dataset, called Musti3D, has been created. The Musti3D database contains episodes from an animation series for toddlers. Annotated with audiovisual information, this database is used for the investigation of a non-negative matrix factorization (NMF)-based audiovisual learning technique. The performance of the technique, i.e. correctly matching the audio and visual representations of the objects, has been evaluated by gradually reducing the level of supervision starting from the ground truth transcriptions. Moreover, we have performed experiments using different visual representations and time spans for combining the audiovisual information. The preliminary results show the feasibility of the proposed audiovisual learning framework.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning-*knowledge acquisition, language acquisition*; I.2.m [Artificial Intelligence]: Miscellaneous

General Terms: Algorithms, Languages, Experimentation

Keywords: Audiovisual learning, non-negative matrix factorization, multimodal dataset, discriminative patches, histogram of acoustic co-occurrences

1. INTRODUCTION

In their first few years of life, children learn to name the objects they are confronted with in their environment. To mimic that process with a machine requires learning from at least two modalities, in this case the auditory and the visual, to create a crossmodal model that links the learned object representations in both domains and so establishes the grounding process.

In this work, the task of learning the relation between visual and auditory description of words is formulated as an *unsupervised* learning problem. Unsupervised learning is an important technique to build truly artificially intelligent systems: systems that learn from observing their environment.

While there is a body of work that investigates the unsupervised multimodal analysis of images and text [2, 3, 7, 16], the combination of continuous speech and tags [13] or images and continuous speech [4, 11] are far less studied. In earlier work, visual objects are presented in isolation or on a clean background and the audio is often recorded on purpose for the specific task of teaching a robot the semantics of the visual objects [5, 10, 15].

Both in vision as well as in speech processing, it has been shown that adding a limited amount of weak supervision significantly improves the pattern discovery capabilities. We investigate whether the same benefits of weak supervision can be obtained in an unsupervised setting, replacing the tags by multimodal information, using a non-negative matrix factorization approach for learning from multiple asynchronous input streams [14]. The key assumption is that co-occurrences over different modalities provide a strong cue for semantics or relevance. In particular, we relate properties of objects in images and audio. The described learning task is shown in Figure 1.

Our main contributions can then be summarized as follows: i) we have created and annotated a new multimodal dataset. Our main research question investigates whether it is possible to learn visual and audio representations for objects from watching a television series for toddlers, where the audio describes the visual scene, but at the same time exhibits a large degree of complementarity. ii) we gradually reduce the level of supervision, starting from ground truth transcriptions for both domains to only using segmentation info (bounding boxes and audio word delineations) without grounding. iii) we compare two different visual representations in this context: bag-of-visual-words, which capture relatively low-level image characteristics, and discriminative patches, which can be seen as an example of more mid-level features.

The remainder of this paper is organized as follows. Section 2 details the audio and video features used in the proposed learning scheme. The NMF-based audiovisual learning framework is discussed in Section 3. The experimental setup is explained and the results are presented in Section 4. Section 5 discusses the results and Section 6 concludes the paper.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655036.

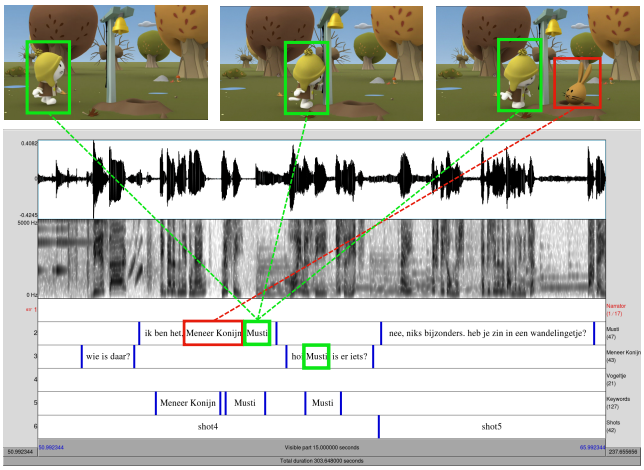

Figure 1: An audiovisual learning example - linking the audio keywords with the visual counterparts.

2. AUDIOVISUAL REPRESENTATIONS

The NMF-based learning framework described in Section 3 can learn object representations from multiple information streams, here the auditory and visual modality and possibly supervisory tags. It requires each *learning example* to be represented as a vector of non-negative values, feature occurrence counts, of a fixed dimension. A learning example accumulates feature counts over a time span in which one expects there would be a relation between the different information streams. In this work, we investigate two choices: the utterance level and the shot level accumulation. A shot is the basic component of a video and commonly defined as an uninterrupted sequence of frames.

2.1 Audio features

The audio information is represented as weighted phone lattice transition probabilities yielding a fixed-length representation of any speech segment with possibly different duration [12]. These features are obtained by labeling the acoustic content of a speech segment with the help of an automatic speech recognizer and accumulating the co-occurrences of each acoustic unit in a histogram.

Using a conventional HMM-based speech recognizer, the network of most probable phone strings (henceforth the phone lattice) is determined. The phone lattice is an acyclic directed graph where the arcs correspond to phones and the nodes mark their start and end times. Every arc is also associated with the acoustic score of the corresponding phone.

The audio features are extracted by transforming this acoustic score into a posterior probability and accumulating the probability of every two consecutive phones ϕ and ψ over the complete phone lattice

$$c(\phi,\psi) = \sum_{\{\alpha:\, h(\alpha)=\phi\}} \sum_{\{\beta:\, h(\beta)=\psi\}} p(\alpha)\, p(\beta)\, \Delta_{\alpha\beta} \qquad (1)$$

in which $h(\alpha)$ and $h(\beta)$ return the phone identity, and $p(\alpha)$ and $p(\beta)$ the posterior probability of the arcs α and β, respectively. If the start node of β is equal to the end node of α, then $\Delta_{\alpha\beta}$ is the inverse of the probability of the common node, otherwise it is equal to zero. This node probability is given by the sum of the posterior probabilities of the incoming (or outgoing) arcs of the corresponding node.

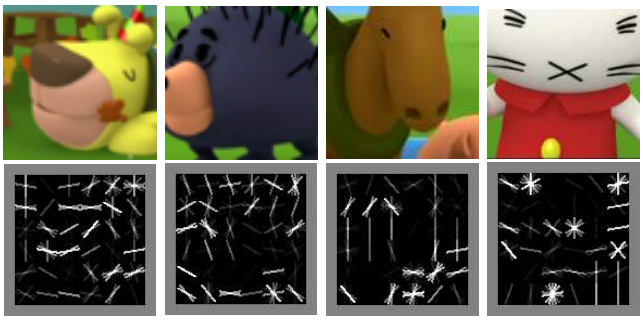

Figure 2: Examples of DPs, together with their (positive) learned weights. Note how the learning focuses on gradients that discriminate the object.

The $c(\phi,\psi)$ values for each speech segment are vectorized and stacked in the columns $\mathbf{V}_a$ of size PxT, in which P is equal to the square of the number of available phones in the alphabet and T is the number of utterances in the database.

2.2 Visual features

2.2.1 *Visual words*

One of the standard approaches to represent images is the Bag of Visual Words (BOW). First, we extract SIFT features from a set of images and we cluster them using k-means; the centers of the resulting clusters are considered as the visual words of a visual vocabulary V_{vis} of size k, with k the predefined number of clusters. Next, for every image we extract dense SIFT and we assign each SIFT descriptor to its closest visual word. By applying sum pooling we generate a histogram representation for the image. Images containing the same object are expected to have overlap in their BOW histograms (similar visual word distribution).

2.2.2 *Discriminative patches*

We additionally consider discriminative patches (DP), a method where each patch captures an aspect of the appearance of specific objects. Following the approach of [1] we want an object specific patch p to be highly activated when the object is present and have low score when other objects or background appears. Therefore we consider a detection task where we train a model w_p for every DP and we use it as a linear classifier to the patches x of a test image:

$$f_p(x) = h(w_p^T x), \qquad (2)$$

with $h(y) = y$ if $y > 0$ and 0 otherwise. Both patches p and x are represented as Histograms of Oriented Gradients (HOG). The weights w_p are learned as an exemplar classifier [9] but instead of a Support Vector Machine we use Linear Discriminant Analysis (LDA) [6]. Assuming that the positive and the negative data come from Gaussian distributions with means μ_p and μ_n and same covariance matrix Σ, the model weights for a patch p are learned:

$$w_p = \Sigma^{-1}(p - \mu_n). \qquad (3)$$

As positive we consider only the HOG features of the patch p, while for estimating μ_n and Σ we use background patches.

The above approach results in one classifier for every sampled patch. However, not all of the patches are suitable for separating the different objects. We select the most discriminative by considering the Mahalanobis distance of a patch to the negative distribution. During testing we scan the image or the bounding box in a sliding window fashion for all

DPs and we keep the score of the maximum activation for each of them (max pooling).

3. NMF-BASED AUDIOVISUAL LEARNING

The links between the audio keywords and video objects are learned by applying non-negative matrix factorization (NMF). The NMF algorithm approximates a non-negative data matrix $\mathbf{V}$ of size MxN as a multiplication of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ of dimensionality MxR and RxN respectively. Here, R is chosen so that $R << M$ and $R << N$ in order to obtain a low-rank approximation, i.e. $\mathbf{V}$ is expressed as an approximate linear combination of basis audiovisual vectors (columns of $\mathbf{W}$) scaled by the weights stored in $\mathbf{H}$.

The factorization is achieved by minimizing the generalized Kullback-Leibler divergence (KLD) between $\mathbf{V}$ and its approximation $\mathbf{WH}$ through multiplicative updates [8]. In the proposed framework, the data matrix $\mathbf{V}$ is comprised of two submatrices $\mathbf{V}_a$ representing the audio features and $\mathbf{V}_v$ representing the video features.

$$\begin{bmatrix} \mathbf{V}_a \\ \mathbf{V}_v \end{bmatrix} \approx \begin{bmatrix} \mathbf{W}_a \\ \mathbf{W}_v \end{bmatrix} \mathbf{H} \tag{4}$$

Defining the audio grounding matrix $\mathbf{V}_{ag}$ as

$$V_{ag}^{ij} = \begin{cases} 1 & \text{if an audio keyword } i \text{ is in utterance } j \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

the unimodal associations are learned by performing the factorization

$$\begin{bmatrix} \mathbf{V}_{ag} \\ \mathbf{V}_a \end{bmatrix} \approx \begin{bmatrix} \mathbf{W}_{ag} \\ \mathbf{W}_a \end{bmatrix} \mathbf{H} \tag{6}$$

where the $\mathbf{W}_{ag}$ is initialized as $\mathbf{W}_{ag} = [\mathbf{I}_K | \mathbf{W}_{gbg}]$. $\mathbf{I}_K$ is a KxK identity matrix where K is the number of audio keywords. $\mathbf{W}_{gbg}$ is a Kx$(R-K)$ matrix with random positive values much smaller than 1.

This setup leads to solutions where each of the K audio keywords is assigned to a single column in $\mathbf{W}_a$. The remaining $(R-K)$ columns, $\mathbf{W}_{gbg}$, are called the *garbage columns* which model other factors like non-keywords or background music. Similarly, the crossmodal associations can be learned by concatenating the audio grounding matrix $\mathbf{V}_{ag}$ with the video features $\mathbf{V}_v$ and following the same procedure as above.

4. EXPERIMENTS

4.1 Dataset

For the audiovisual learning experiments, we have created a new realistic multimodal dataset, called Musti3D, a television series for toddlers with 30 episodes each lasting 5 minutes. We have annotated 22 episodes with various audio and video information such as the orthographic transcription of the utterances, the occurrences of various audio keywords, video objects and the shot boundaries.

We have chosen 12 characters from the series appearing with different frequencies and perform the learning experiments aiming to learn the audiovisual features belonging to these characters. In practice, we have defined 12 video objects and 13 audio keywords (one of the video objects is associated with two audio keywords).

We pick 5 episodes as oracle for evaluation purposes considering the balanced occurrences of the target objects and keywords in the training and test data. The training episodes are used for extracting the reference representations of the target objects and keywords. These reference representations are used to classify the learned audio and visual representations. Due to the unbalanced number of occurrences, we randomly select 10 reference BOW and DP representations for each object and 5 reference HAC representations for each keyword.

4.2 Implementation details

The audio representations are extracted using the annotated keyword boundaries extended with an offset of 0.1 seconds. No additional precaution is taken against the background music. For the HAC feature extraction, phone lattices are generated using a conventional HMM-based automatic speech recognizer. A phonetic set consisting of 40 Dutch phones is used.

For the BOW features, we used dense SIFT at multiple scales and a vocabulary of size 300. To obtain DPs, we sampled approximately 15 patches per bounding box from the training data at multiple scales and we extract features from 6x6 HOG cells. We consider the top 20 DPs per object based on their distance from the negative distribution. For the final image representation, we extract BOW or DP features for each bounding box separately and sum the result.

The final recognition accuracies are obtained after averaging the results of 10 independent trails of 100 NMF iterations. In the experiments using grounding matrices, 3 garbage columns are used resulting in $R_{vg} = 15$ and $R_{ag} = 16$. The learned representations are classified according to the label of the closest reference representation with respect to the generalized KLD.

In the experiments without grounding information, each column of the basis matrices $\mathbf{W}_a$ and $\mathbf{W}_v$ are classified as the label of the closest reference representation. The $\mathbf{W}$ and $\mathbf{H}$ matrices are initialized either in a supervised or unsupervised manner. In the supervised initialization setting, the initial $\mathbf{W}$ and $\mathbf{H}$ are obtained after a single NMF iteration using the visual grounding information. On the other hand, in the unsupervised case, $\mathbf{W}$ and $\mathbf{H}$ are randomly initialized and R is set to 50. The recognition accuracy is obtained as the ratio of the number of unique correct matches to the number of video objects.

4.3 Results

We present the recognition accuracy results in Table 1 listing the experiments that are performed using the video grounding matrix and audio grounding matrix in the upper and middle panel respectively and the experiments without using the grounding information in the lower panel. Firstly, we will focus on the experiments with grounding. The ground truth experiments which are given in the first rows of the upper and middle panels yield that the proposed method is able to find the correct crossmodal match of each video object (audio keyword) with a recognition accuracy of 74% (66%) at utterance-level and 81% (77%) at shot-level.

In the second block of the upper and middle panel, the unimodal learning accuracy of both sources are evaluated. The learned visual representation for shot-level DP features match the reference DP features for 84% of the video objects. Only 24% of the shot-level BOW features match correctly with the reference BOW features. Using utterance-level features slightly improves the accuracy of BOW features to 35%, while it reduces the performance of DP features to

55%. For HAC features, the recognition accuracy is around 71% and 76% for both levels.

In the third blocks, the results of a less supervised crossmodal learning setting, i.e. using the grounding matrices with the crossmodal features (rather than the ground truth crossmodal features), are presented. For visual grounding, both shot- and utterance-level features provide a similar recognition accuracies of 59% and 56% respectively. For audio grounding, the shot-level DP features provide considerably higher performance with 69% compared to the 26% of shot-level BOW features.

The lower panel summarizes the recognition results of the learning experiments without using the grounding information. The upper block shows the results where the basis matrix **W** is initialized in a supervised manner. The shot-level DP features provides 40% recognition accuracy compared to the 29% of the shot-level BOW features. The utterance-level features provide inferior performance for both DP and BOW features. The results in the lower block are obtained with unsupervised initialization of **W**. All features provide around 30 % of recognition accuracy.

Visual grounding	utt.-level	shot-level
+GTvideo-GTaudio	74	81
+GTvideo-BOW feat.	35	24
+GTvideo-DP feat.	55	84
+GTvideo-HAC feat.	56	59
Audio grounding	utt.-level	shot-level
+GTaudio-GTvideo	66	77
+GTaudio-HAC feat.	76	71
+GTaudio-BOW feat.	19	26
+GTaudio-DP feat.	35	69
Learning w/o grounding	utt.-level	shot-level
+BOW feat.-HAC feat. (sup.)	17	29
+DP feat.-HAC feat. (sup.)	10	40
+BOW feat.-HAC feat. (unsup.)	31	25
+DP feat.-HAC feat. (unsup.)	28	30

Table 1: Recognition accuracy results in percentages

5. DISCUSSION

First, based on a set of experiments using ground truth transcripts, we have shown that even though audio and video show a high degree of complementarity, co-occurrence information can guide non-negative matrix factorization to a meaningful solution. The appropriate temporal unit for capturing such co-occurrences seems to be the shot-level.

While not yet fully unsupervised, our experiments are encouraging as they show that data from a TV series for toddlers shows sufficient temporal overlap between object names in audio and occurrences of the corresponding objects in the visual stream to learn object semantics in an unsupervised way. This is a leap forward compared to earlier work, where objects are often shown in isolation or on a clean background and the audio is often recorded on purpose for the specific task of teaching a robot the semantics of objects.

We have also compared two different visual representations: bag-of-visual-words, which capture relatively low-level image characteristics, and DPs, which can be seen as an example of mid-level features. While in the current version the DPs are selected based on some supervision, the superior results obtained with these features clearly illustrate that more powerful mid-level representations are critical for learning in an unsupervised manner.

6. CONCLUSION

In this work, an NMF-based audiovisual learning technique has been described and applied to a realistic multimodal dataset containing episodes from a toddler TV series. At several supervision levels, the performance of the technique has been investigated using different video features and time spans for fusing the audio and visual information. From the results, it can be concluded that the learning approach provided promising results on the task of learning the relation between visual and auditory description of words. Future work includes performing feature extraction in an unsupervised manner rather than using the annotated information and deeper investigation of the impact of background music.

7. ACKNOWLEDGMENTS

This work has been supported by the KU Leuven research grant OT/09/028 (VASI).

8. REFERENCES

[1] M. Aubry, B. C. Russell, and J. Sivic. Painting-to-3D model alignment via discriminative visual elements. *ACM Trans. Graph.*, 33(2):1–14, 2014.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, Mar. 2003.

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

[4] J. W. Fisher III, T. Darrell, W. T. F., and P. A. V. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems 13*, pages 772–778. 2001.

[5] K. Gold, M. Doniec, C. Crick, and B. Scassellati. Robotic vocabulary building using extension inference and implicit contrast. *Artificial Intelligence*, 173(1):145 – 166, 2009.

[6] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Proc. ECCV*, pages 459–472, 2012.

[7] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *Proc. ICCV*, pages 2407–2414, Nov 2011.

[8] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[9] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *Proc. ICCV*, pages 89–96, 2011.

[10] T. Nakamura, T. Nagai, and N. Iwahashi. Grounding of word meanings in multimodal concepts using LDA. In *Proc. IROS*, pages 3943–3948, Oct 2009.

[11] D. K. Roy and A. P. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, 2002.

[12] V. Stouten, K. Demuynck, and H. Van Hamme. Discovering phone patterns in spoken utterances by non-negative matrix factorization. *IEEE Signal Processing Letters*, 15:131–134, 2008.

[13] L. ten Bosch, L. Boves, H. Van hamme, and R. K. Moore. A computational model of language acquisition: The emergence of words. *Fundam. Inf.*, 90(3):229–249, Aug. 2009.

[14] H. Van hamme. Integration of asynchronous knowledge sources in a novel speech recognition framework. In *Proc. ISCA ITRW*, June 2008.

[15] C. Yu and D. H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Trans. Appl. Percept.*, 1(1):57–80, 2004.

[16] Y. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.