Spotting Visual Keywords from Temporal Sliding Windows

Yue Yao, Tianyu Wang, Heming Du, Liang Zheng, Tom Gedeon
yue.yao@anu.edu.au,tianyu.wang2@anu.edu.au,du_heming@hotmail.com,liang.zheng@anu.edu.au,tom@cs.anu.edu.au
The Australian National University
Canberra, ACT

ABSTRACT

Visual Keyword Spotting (KWS), as a newly proposed task deriving from visual speech recognition, has plenty of room for improvements. This paper details our Visual Keyword Spotting system used in the first Mandarin Audio-Visual Speech Recognition Challenge (MAVSR 2019). With the assumption that the vocabularies of target dataset are a subset of the vocabulary of the training set, we proposed a simple and scalable classification based strategy that achieves 19.0% mean average precision (mAP) on this challenge. Our method is based on the idea of using sliding windows to bridge between the word-level dataset and the sentence-level dataset, showing that a strong word level classifier can be directly used in building sentence embedding, thereby making it possible to build a KWS system.

KEYWORDS

Visual keyword spotting, lip reading, video classification

ACM Reference Format:
Yue Yao, Tianyu Wang, Heming Du, Liang Zheng, Tom Gedeon. 2019. Spotting Visual Keywords from Temporal Sliding Windows. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340555.3356101

1 INTRODUCTION

Lip reading is a rising area in computer vision, with the emergence of multiple types of datasets [1, 4–6, 11]. Usually, existing lip reading datasets can be divided into word-level or sentence-level datasets. Word-level lip-reading datasets often store short video segments. In each segment, a person is pronouncing exactly one word from a pre-defined vocabulary.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3356101

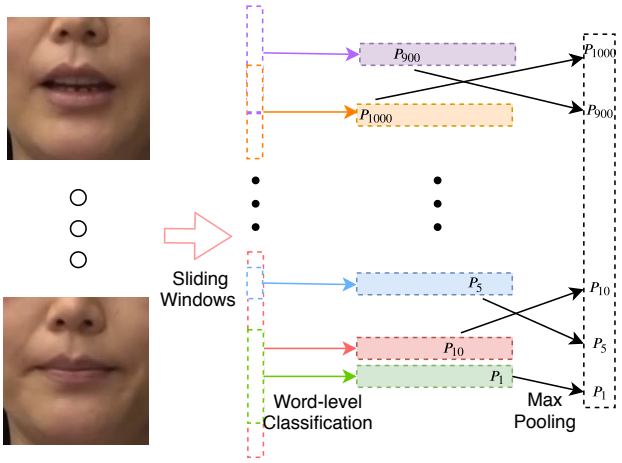

Figure 1: System Overview: sliding windows are used to split long videos into small pieces, then word level classification is performed on each sliding window, finally a simple element-wise max pooling is used to integrate information across sliding windows.

One outstanding benchmark is the LRW dataset proposed in 2016 containing 500 different English words [4]. With a substantial diversity in speech conditions, the LRW dataset started our work on lip reading in real world practice. After this, the LRW-1000 dataset employed in the MAVSR 2019 challenge is the first Madarin word level dataset which has 1000 classes.

But from our intuition, a practical system should be able to work with sentence level data but not only word level data. In a real world scenario, we usually do not have exact word splits beforehand. Furthermore, for a long time, the gap between word level lip reading and sentence level lip reading has been huge. Due to ambiguity and the semantic coherence property of lip reading, applying a word level system naively to sentence level data does not produce satisfactory results. To the best of our knowledge, the only usual connection between a word level system and a sentence level system is that they could share the same visual front-end which takes in consecutive raw video frames and output a frame embedding [1]. As a result, it is a bit wasteful if the purpose of

the existence of word-level dataset is only to train powerful feature extractors.

The KWS challenge in MAVSR 2019 gives us a chance to rethink the relationship between a word level system and a sentence level system. In this challenge, it only provides a word level training set but needs us to perform KWS inference on sentence level data. The limitations of the training set forces us to build system with small input size. Naturally, the training data cannot be directly used for sentence level data.

The segment proposals used in temporal action localization give us inspiration [3, 10]. In temporal action localization, the key to the success of segment proposals usage is that the classification loss can give strong guidance to the update of the system [3]. Such strong guidance comes from the classification accuracy for action recognition, which has reached 79.0% top 1 accuracy on Kinetics-400 [7] and 98.0% top 1 accuracy on UCF101 dataset [2]. This allows us to form an analogy to lip reading. In the LRW dataset, the state of the art system has already reached 83% accuracy for the 500-class classification task and it is 38.3% for the LRW-1000 dataset. As a result, that shows we have the potential to use segment proposals as a link between word level and sentence level datasets.

But using segment proposals for lip reading directly still faces difficulties. For lip reading, the boundary between a word and the next word can be ambiguous, making it hard to set boundaries by hard rules or a learning system. As a result, at the early stage of this task, we directly use all possible segment proposals, in other words made sliding windows across video frames. In this paper, we trained a word level classification system only and use sliding windows to generate sentence level embedding. For this, we make the following contributions to MAVRC with our proposed system.

- For the first time, we propose a system that is trained on a word level dataset but tested on a sentence level dataset by using sliding windows to set up a link between them.
- A competitive sentence level KWS system on the MAVSR challenge based on merely the word level dataset. Due to the training difficulty of building a lip reading system, our system is simple but will become a comparable baseline for coming KWS system.

2 RELATED WORK

Word-level Visual Speech Recognition is conventionally formulated as a video classification problem [4]. The input is a video that has been temporally trimmed to contain only one specific word and our goal is to correctly classify or identify this word. Tremendous progress has recently been made due to the introduction of large datasets and the developments

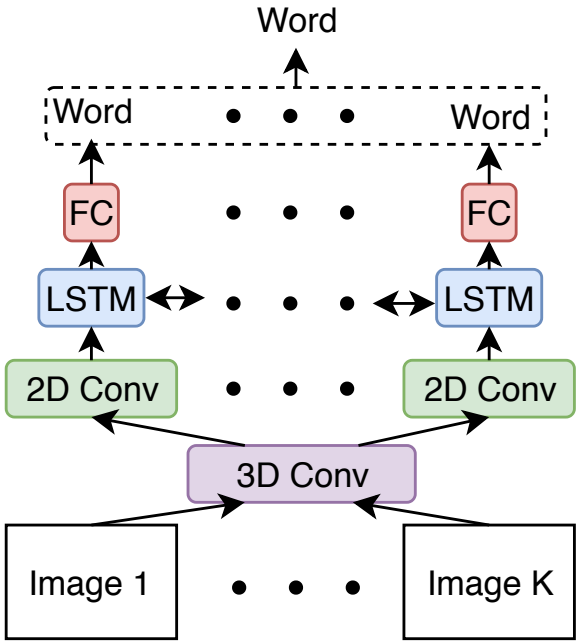

Figure 2: 3D+2D Visual Speech Recognition Structure.

of deep neural network models. However, the trimmed input assumption limits the application of these approaches in real scenarios, where the real world videos are in general untrimmed, and creating word level segmentation is tedious, time consuming and needs human labeling.

Visual Keyword Spotting (KWS) is the problem that we are targeting in this paper. This task requires us to determine whether a text query occurs in a given video with erased audio. Proposed by Themos [9], KWS tasks can be solved in two different ways, one is to directly apply sentence level translation for automatic speech recognition, and the second is to apply a Query-by-Text visual system that takes both words and video as input and perform binary classification to decide the existence of the word. But both these methods require sentence level training data, making it hard to apply to this challenge directly.

3 METHOD

Dataset

We derive our methods from the properties of the LRW-1000 dataset [11]. The dataset vocabulary contains 1000 Chinese words, approximately 718,018 samples in total. Currently, it is the only publicly available large-scale Chinese lip reading recognition dataset. The video sequences in the dataset are all extracted from real world TV programs with complex environmental conditions including differences in illumination, speaker pose, sampling rate, video resolution, etc. For the MAVSR 2019 challenge, we used the LRW-1000 dataset

Algorithm 1 Sliding Windows Algorithm for Building Video Embedding

Input: a sentence level video V in length N and a trained word level classification system NET().
Output: Output a vector in dimension W that represents occurrence probability of each word in this video.

```
P_w ← 0, ∀w ∈ W
for i in range (1, N + 1) do
    for j in range (3, 20) do
        if i + j <= N then O = NET(V[i, j])
            for w in range (1, W + 1) do
                P_w ← max(O_w, P_w)
            end for
        end if
    end for
end for
return P
```

as our training set and two additional datasets are used for validation and testing.

Word Level Visual Speech Recognition

We adopt the 3D+2D structure from Themos' work [8], which has been shown to be the state of the art method for LRW-1000 [11]. Shown in Fig 2, it contains a 3D convolution spatial-temporal front-end, a 2D residual network in the middle and Bidirectional LSTM back-end. A fully connected layer and softmax layers are placed at each time step of the LSTM output in order to avoid LSTM gradient vanishing. For hyper-parameters, we used Resnet-34 in the middle, the hidden size is 1024 for LSTM and the learning rate is set to 1e-4 at the beginning of training while dropping to 1e-5 after 8 epoches.

Sliding Windows and Sentence Embedding

Given a sequence of the sentence-level video frames, we first enumeration all possible sliding windows. The max frame length of a word in LRW-1000 is 7, with 4.3 being the average frame length. In order to balance the calculation time and performance, we set sliding windows to lengths of 3 to 20. Shown in both Fig 1 and algorithm 1, each sliding window will go through the Word Level Visual Speech Recognition model to get the probability distribution of the 1000 classes. Then we apply the max pooling across all sliding windows to get the final probability distribution of the 1000 classes. The 1000 dimension vector will be the sentence embedding of this video and each element of this vector represents the occurrence probability of a certain word.

4 EXPERIMENT RESULT

The result of the word level visual speech recognition is shown in Table 1. Our system is a little higher than the original baseline due to some hyper-parameter setting. For evaluation targeting KWS, we use two different methods.

Table 1: 3D+2D Strucutre Performance

Method	Top 1	Top 5
Reported [11]	38.19%	63.50%
Reproduced	**38.87%**	**63.86%**

Element-wise Intersection over Union

For a visual key word spotting system, we define Element-wise Intersection over Union (eIoU) as a description of intersection of predicted keyword list X and ground truth keyword list Y of a certain sentence level video. Shown in the formula below, the eIoU will be 100% if two keyword lists are same and will be 0 if two keyword lists have no intersection.

$$eIoU = \frac{X \cap Y}{X \cup Y}$$

The result of eIoU is shown in Table 2, with a comparison of three methods of producing the predicted keyword list given a video embedding. The first row is to set the global boundary on video embedding to decide whether each word exists or not in the predicted keyword list, the second row is to set a separate boundary for each word to decide if it exists or not. The separate boundary is also set by brute enumeration. The third row refer to selecting top K words with highest probability in a sentence embedding, the K is a hyper-parameter and is equal to the video length divided by 13.

Table 2: eIoU Result for Different Methods

Method	eIoU
Global Boundary	13.24%
Separate Boundary	**16.47%**
Top K	16.34%

Mean Average Precision

For our method, we will produce a probability for each word and each video. Therefore, we can also use mean average precision (mAP) to evaluate the overall performance. That is for each word, we draw the Precision-Recall curve and calculate the area under the curve, which is known as average precision (AP). Then, the mean value of APs of all words (mAP), is calculated. When average precision (AP) is used, video rank lists for each word are effectively distinguished. The advantage of using mAP is it can reflect both precision and recall of our algorithm, thereby providing a more comprehensive evaluation.

Table 3: mAP Performance for Words in Different Length

Word Length	mAP
1	9.22%
2	16.04%
3	26.17%
4	41.33%
5	22.81%

We achieved 17.1% mAP on the validation set and 19.0% mAP on the test set, ranked first on MAVSR challenge task 3. The detailed mAP for word length in the validation set is shown in Table 3. We can see that though we get competitive initial global result for visual keyword spotting, it is still unstable for words with different lengths.

5 CONCLUSION

The link between word level lip reading datasets and sentence level lip reading datasets is weak in previous studies. In this paper, we proposed a simple but competitive method to use only a word level training set to generate sentence level lip reading embedding. By using sliding windows, we generate a sentence level video embedding which is capable of performing visual keyword spotting. We achieve competitive performance on MAVSR 2019 challenge.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1130–1139.

[4] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87–103.

[5] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.

[6] Andrzej Czyzewski, Bozena Kostek, Piotr Bratoszewski, Jozef Kotus, and Marcin Szykulski. 2017. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems* 49, 2 (2017), 167–192.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2018. SlowFast Networks for Video Recognition. *arXiv preprint arXiv:1812.03982* (2018).

[8] Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining Residual Networks with LSTMs for Lipreading. *Proc. Interspeech 2017* (2017), 3652–3656.

[9] Themos Stafylakis and Georgios Tzimiropoulos. 2018. Zero-Shot Keyword Spotting for Visual Speech Recognition In-the-wild. In *European Conference on Computer Vision*. Springer, 536–552.

[10] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 5794–5803.

[11] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. 2019. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.