Machine Vision and Applications
https://doi.org/10.1007/s00138-019-01006-y

SPECIAL ISSUE PAPER

Spotting words in silent speech videos: a retrieval-based approach

Abhishek Jha[1] · **Vinay P. Namboodiri**[2] · **C. V. Jawahar**[1]

Received: 24 October 2018 / Accepted: 2 January 2019

Abstract
Our goal is to spot words in silent speech videos without explicitly recognizing the spoken words, where the lip motion of the speaker is clearly visible and audio is absent. Existing work in this domain has mainly focused on recognizing a fixed set of words in word-segmented lip videos, which limits the applicability of the learned model due to limited vocabulary and high dependency on the model's recognition performance. Our contribution is twofold: (1) we develop a pipeline for recognition-free retrieval and show its performance against recognition-based retrieval on a large-scale dataset and another set of out-of-vocabulary words. (2) We introduce a query expansion technique using pseudo-relevant feedback and propose a novel re-ranking method based on maximizing the correlation between spatiotemporal landmarks of the query and the top retrieval candidates. Our word spotting method achieves 35% higher mean average precision over recognition-based method on large-scale LRW dataset. We also demonstrate the application of the method by word spotting in a popular speech video ("*The great dictator*" by Charlie Chaplin) where we show that the word retrieval can be used to understand what was spoken perhaps in the silent movies. Finally, we compare our model against ASR in a noisy environment and analyze the effect of the performance of underlying lip-reader and input video quality on the proposed word spotting pipeline.

Keywords Keyword spotting · Lip-reading · Visual speech recognition · Recognition-free retrieval

1 Introduction

Parsing information from videos has been explored in various ways in computer vision. Recent advances in deep learning have facilitated many such tasks. One such parsing requirement is of reading lips from videos. This has applications in surveillance or aiding improvements in speech recognition in noisy outdoor settings. Solving this problem has been attempted using methods based on recurrent neural networks (RNN) [34] and spatiotemporal deep convolutional networks [36]. However, for practical applications, recognizing lip motion into words is still in its nascent stages, with state-of-the-art models [46] being limited to a constrained vocabulary. In this paper, we adopt a recognition-free "word spotting" approach that does not suffer from the vocabulary limitations. Unlike text documents, where the performance in character recognition [56], word recognition [17] and spotting research [49] has seen a great boost in the post-deep learning era, this approach has been rarely pursued for lip-reading task.

Training a lip-reader requires careful word-level annotation, which is expensive even for a small vocabulary set. Although progress in speech recognition [55] has resulted in better audio-to-text prediction and can be used for annotation, such methods are often prone to changes in accent and presence of noise in the audio channel. Lip-reader's performance is also susceptible to similar sounding words [46]. In recognition-based retrieval, we use a lip-reader to predict the word spoken in a video clip. Evidently, if the word is wrongly predicted due to variations in visual appearance, it would never appear in the top results. In contrast, for recognition-free retrieval, the "word spotting," i.e., matching of words is based on the feature representation of the target word without explicitly predicting the word itself. It intrinsically compares the features of the target word with the query word. Hence, even if the target word is misclassified it appears in the top results.

✉ Abhishek Jha
abhishek.jha@research.iiit.ac.in

Vinay P. Namboodiri
vinaypn@iitk.ac.in

C. V. Jawahar
jawahar@iiit.ac.in

[1] CVIT, IIIT Hyderabad, Hyderabad, India

[2] Department of Computer science, IIT Kanpur, Kanpur, India

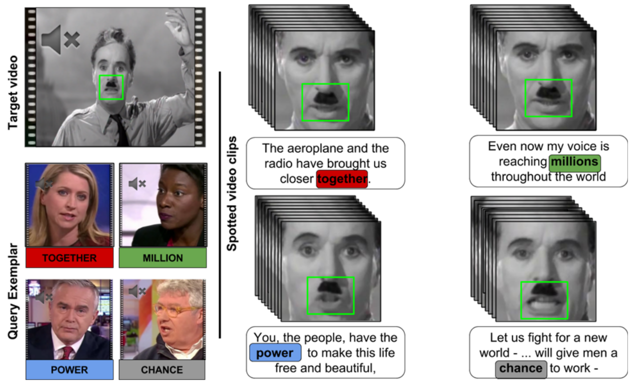

Fig. 1 Example of word spotting in black and white Charlie Chaplin silent video: (left) target is the silent video and queries are the exemplars spoken by different people;(right) retrieved video clip segments where the words "together," "million," "power" and "chance" are present

We are motivated by the fact that for handwritten documents word spotting has shown better performance for retrieving target words in different handwriting styles than word recognition [42]. Likewise, we show that recognition-free retrieval can also be useful for spotting words when target words come from a different source than the data used for training a lip-reader, like archaic black and white documents films in Fig. 1. We further investigate the applicability of recognition-free pipeline for out-of-vocabulary word spotting, for a different domain of data with respect to what has been used for training the lip-reading model. Figure 1 shows few sample results of our pipeline for spotting different query words in the black and white video clip in four spoken sentences.

We further show that the word spotting performance can be improved by using a novel re-ranking method for top-k retrieval candidates. We also adapt the standard pseudo-relevance feedback query expansion method for lip-reading task. Our pipeline takes silent speech videos as input and retrieves a queried word that is provided again as a video clip from the target input dataset. The target video is first densely segmented into "word proposal clips," where these clips may or may not contain any word. Any "word proposal clip" is considered a spotted word if the similarity measure between the query and the target "word proposal clip" is greater than a particular threshold.

We show improvement in word spotting on a standard large-scale lip video dataset Lip-reading in the Wild (LRW) [14], and another standard dataset GRID corpus [16] for showing domain invariance. We also assess our pipeline's performance in a popular speech video by Charlie Chaplin: "*The great dictator*". Finally, we extend our work [35] with additional experimental evaluations and comparisons. We compare lip word spotting with word spotting using ASR to analyze the robustness of our pipeline against noise in the channel. We perform qualitative analysis of the retrieval pipeline and its label assignment stage. We also show the tolerance of our method against lip-reader's performance and quality of input video.

2 Related work

Research in visual speech recognition has been pursued for at least two decades [4,7,31] with earlier approaches focusing mainly on handcrafted features and HMM-based speech recognizers [5,32,40]. Some of these approaches have been thoroughly reviewed in [29,57]. Wand et al. [52] showed word-level lip-reading using an LSTM [34] stacked over a two-layered neural network on GRID corpus dataset [16]. Similarly, research in automatic speech recognition (ASR) traditionally used HMM models along with hand engineered stages for modeling acoustic speech. Before deep learning era, recurrent neural networks [44] and deep belief networks [43] have also been used in ASR.

DeepSpeech [28] extended the work of the first end-to-end audio-based ASR which uses joint RNN-CTC model [27] for large datasets. Using spatiotemporal convolutions along with the joint RNN-CTC model, Assael et.al. [2] introduced the first end-to-end lip-reading model. It uses a connectionist temporal classification (CTC) [25], providing one of the best results on GRID corpus [16]. Attention-based model "Listen, Attend and Spell" was introduced by Chan et.al. [8] in 2016. Subsequent attention-based models outperformed the CTC-based models[12]. Extending the LAS model to visual ASR, Chung et. al [14] presented lip-reading model which surpasses human level performance. They use multiple lip-reading models that fuses the temporal sequence at different layers of underlying VGG-M model [9] to classify the input video clip into 500 words.

Lip-reading involves modeling temporal sequences of lip video clips into phonemes [48] or characters; hence, better sequence learning models using deep networks proved to be pivotal in lip-reading research. Earlier attempts at sequence-to-sequence learning have been for machine translation applications [10,50] using RNN encoder-decoder, but they lacked long-term dependencies between input and output sequences. Attention mechanism [3] by Bahdanau et. al. overcame this shortcoming by using an attention vector in the bottleneck layer between encoder and decoder to focus on components of input sequence based on the output sequence. Chung et al. [13] have proposed Watch Listen Attend and Spell (WLAS) architecture that leverages attention model [3] for doing character level prediction of input lip videos. They provide the best results on Lip-reading in the Wild (LRW) dataset and GRID corpus [16]. They, however, use a much larger Lip-reading Sentences (LRS) dataset that is not widely available [13] for pretraining, hence making it a data-intensive model that is not accessible. In a recent work, Stafylakis and Tzimiropoulos [46] trained a model entirely on LRW dataset to give state-of-the-art result for word-level prediction. This model consists of three parts: a spatiotemporal convolutional front-end, followed by a Resnet-34 [30], and a bidirectional LSTM [26] at the end. Since this model has been trained to classify lip videos into one of 500 word classes, it does not address out-of-vocabulary words. Our pipeline employs recognition architectures based on [14] and [46] as feature extractors to show how recognition-free leverages these features spaces for improved retrieval performance.

Initial work in word spotting appeared in speech recognition community, majority relying on HMMs [24,45]. Kernel machines and large margin classifiers introduced by Keshet et al. [37] in discriminative supervised setting resulted in an improvement over the previous methods. Post-deep learning, RNNs with CTC objective functions gave a major improvement over the HMMs [19] for modeling temporal audio speech signals. Unlike audio speech, visual speech is a spatiotemporal signal. Hence, our choice of feature extractors contains VGG-M [9] and Resnet-34 [30] modules for modeling facial features, and uses LSTM and temporal convolution for modeling temporal information.

Word spotting is a well-defined problem in document analysis[33] and retrieval [23]: hand writing recognition [20,22,42,49], word image retrieval [39], scene-text [53], etc. In speech domain, Keshet et al. [37] improve word spotting in audio speech by learning phrases using discriminative supervised learning.

Although a large corpus of work exists for word spotting for documents, images and audio speech, the visual speech domain has been largely ignored. Liu et al. [41] employ fusion of HMM classification scores on the handcrafted feature of the individual modalities to spot words in multimedia. The work that is closest to our approach is by Wu et al. [54]. In their approach, the authors use geometric and appearance-based features to build their word spotting pipeline and they rely on the knowledge of optimal handcrafted feature. Another recent work on keyword spotting on lip videos by Stafylakis and Tzimiropoulos [47] proposed a two stream network: a 3D resnet followed by a RNN, and an grapheme(or sequences of letters)-to-phoneme encoder–decoder architecture to learn an embedding from the lip space and keyword's graphemes to the same phoneme space. Their model requires an external supervisory signal in the form of phoneme ground truth from CMU dictionary to train the common embedding, which may not be present for low resource languages.

In our work, though we also adopt a recognition-free retrieval approach, we do so using recognition-based features and show that the recognition-free approach improves on the recognition-based approach. We further also improve the base recognition-free pipeline by using query expansion and re-ranking extensions. We benchmark our work on standard datasets.

3 Proposed method

In this section, we will discuss the individual components of our proposed word spotting pipeline and move along to develop a holistic overview of the method.

3.1 Recognition-free retrieval

Recognition-based retrieval relies on recognizing words in lip videos by completely depending on the lip-reading model. During testing a video clip containing a word is classified as one of the word in the vocabulary it is trained on. Moreover, modeling a lip-reader with open vocabulary is an active area of research.

Retrieving a word from a set of candidate silent videos without directly recognizing each candidate words being spoken is recognition-free retrieval or word spotting. This opens up an opportunity to use a sub-performing lip-reader with incorrect word recognition.

In a recognition-free setup, the user formulates a query and a rank list is computed based on its distance from all the clips in the target corpus (retrieval set), such that most similar candidate is given the highest rank. Since word spotting systems rely heavily on computing similarity, the quality of the feature representation is more important than the classification of input clips.

Word spotting based on the modality of query is of two types: query by string (QbS) where the input query is a string and the retrieval is a video, and query by exemplar (QbE), where query is a video and retrieval is also a video. In this work, our query will be through exemplar.

3.2 Preprocessing

We use the recognition models as described in [13] and [46] as feature extractors. These models take inputs as a fixed length input of spatial dimension 225×225 and 112×112, respectively, with a sequence length of 29 frames. The feature extractors are trained on LRW [14] dataset which consists of fixed length video clips of 29 frames and 1.16 s duration, with actual word at the center. Hence, it is required to preprocess the input videos (other than that of LRW) before feeding them to the feature extractors. As shown in Fig. 2, the preprocessing step proceeds by just sampling the input video at 25 frames per second, then converting the sampled frames to gray scale. Since words can be of different length we circular pad gray-scaled sequence of frames on both the sides such that the actual content is at the center of the sequence.

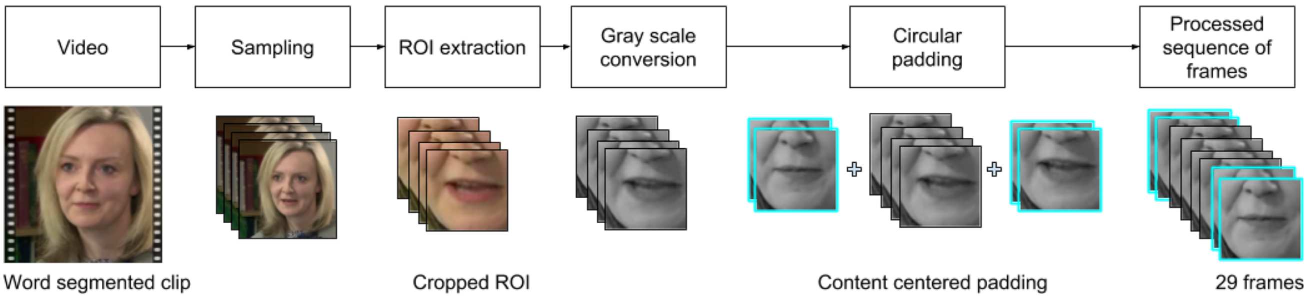

Fig. 2 Preprocessing: the pipeline which takes a variable length word clip and converts it into a fixed length sequence of frames

Circular padding of length 2 for a sequence: {1, 2, 3, 4, 5} on both sides gives {4, 5, 1, 2, 3, 4, 5, 1, 2}.

3.3 Video features

Our first feature extractor only uses the visual stream of the WLAS architecture and hence called Watch, Attend and Spell (WAS) model [13]. Chung et al. [13] train WLAS model on LRS dataset [13] and fine-tune it on LRW dataset [14]. As LRS dataset [13] is not yet publicly available, we trained our WAS model entirely on LRW dataset. WAS contains two modules: a VGG-M convolution module and an attention-based sequence-to-sequence LSTM module, followed by 28 neurons with softmax nonlinearity. Our output sequence for a lip video clip is maximum 20 character long, 28 dimensional(D) (A to Z, *eos*, *padding*) ground truth (GT) word label. Using early stopping, we achieve a word accuracy of 53%.

We also employ another network "N3" as described by Stafylakis and Tzimiropoulos [46] for feature extraction. This network is composed of three modules: A layer of 3D convolutions followed by three dense layers (fully connected layers), and finally a temporal convolution layer. The final layer has 500 neurons with softmax nonlinearity. The classification accuracy of this model is 69.7%. We will address this model as CMT in this paper.

In both the feature extractors, the choice of features is the softmax scores or the probabilities of a lip videos belonging to different words in the vocabulary, instead of sparsely belonging to only one word. We also experimented with the output of the last dense layer as feature representation for the input video and found softmax scores to be empirically better.

3.4 Overall pipeline

In this section, we propose a pipeline for spotting words in silent lip videos. In order to demonstrate generic nature of our pipeline, we first train our two different feature extractors on LRW dataset. We project the query set, consisting of preprocessed annotated video clips, and retrieval set video

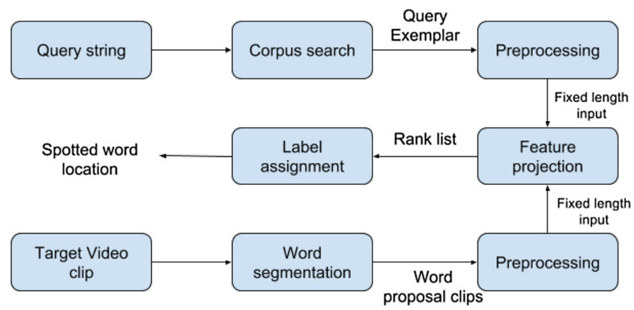

Fig. 3 Overall pipeline: first a string is searched in an annotated corpus to formulate an exemplar which is then preprocessed, and projected into feature space. Target video is then segmented into word clips, either using given time-stamp or dense segmentation, preprocessed and projected in the same feature space. A ranking is computed based on the cosine similarity between query exemplar and the word proposal clips. Label is transferred based on majority voting, as discussed later in Sect. 3.4

clips which do not have any labels into the feature space. The label of the query is assigned to a particular candidate clip in the retrieval set, only if the mean similarity score of that candidate with all the same label queries is greater than a threshold; otherwise, it is assumed the candidate word proposal clip does not contain a full word. In Fig. 3, we show our overall pipeline.

More precisely, if q_i^c is the feature representation of ith query belonging to label c and r_j is feature representation of the jth word proposal clip, the similarity score between the two is given by n_{ij}^c in Eq. 1.

$$n_{ij}^c = \frac{(q_i^c)^T \cdot r_j}{\|q_i^c\| \cdot \|r_j\|} \tag{1}$$

The average similarity between all the queries q^c belonging to label c and the candidate r_j is given by s_j^c in the below Eq. 2.

$$s_j^c = \frac{\sum_{|q^c|} n_{ij}^c}{|q^c|} \tag{2}$$

Finally, the label assignment for the candidate r_j is c if the mean similarity score between all the queries belonging to label c, i.e., s_j^c, is greater than ρ. Otherwise, we consider the word proposal clip is either noise or does not contain the whole word, as represented by ϕ.

$$\text{label}_{r_j} = \begin{cases} c & \text{if } s_j^c > \rho \\ \phi & \text{otherwise} \end{cases} \tag{3}$$

Hence, these word proposal clips are spotted as word c using the queries q_i^c in the target video. We can further use enhancements over this pipeline to improve the retrieval performance, which we will discuss in the next section.

4 Enhancements

In this section, we discuss a query expansion technique to search videos with semantic relevance to the given query, followed by re-ranking method to improve ordering of top-k results.

4.1 Query expansion and re-ranking

Query expansion, in image retrieval [1], has been widely used to improve retrieval performance by increasing the recall and obtain additional documents which might get missed with the original query. Similar to documents, we first feed a *seed* query to our retrieval system which gives us a ranked list of all the candidates from the retrieval set. From this set, top-k candidates are selected to construct a new query based on the weighted sum of the query and top-k candidates feature vectors as the pseudo-relevance feedback to improve the retrieval results.

Re-ranking is used to improve the ranking of top retrieval results for a given query. Some of the prominent re-ranking method [18,51] relies on geometrical consistency between query and its top retrieval candidates. Fergus et al. [18] uses RANSAC [21] to re-rank top results from Google Image search engine. Unlike images, lip videos are temporal in nature with each word consisting of a specific set of phonemes. To adapt such a method for lip videos, we extract spatiotemporal features. Out of total 68 facial landmarks [38], we first compute the distance between all the 20 landmarks associated with lip and the lip-central landmark (landmark no. 63), as shown by "red" color landmark in Fig. 4a. Both landmarks no. 63 and 67, being in the center, are clearly visible for different head poses and hence can be chosen for computing distances. However, on an average, the motion of the upper lip is lesser than the lower lip for most of the word utterances, makes landmark 63 more stable and a better choice.

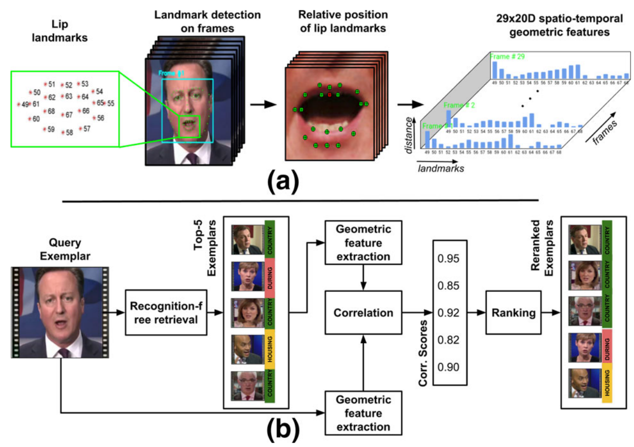

Fig. 4 Re-ranking using geometric cues of lip video: **a** shows method of extracting spatiotemporal feature using lip landmarks of each frame of the video clip; **b** shows re-ranking of top-5 retrieved candidates based on the correlation between spatiotemporal features of top-5 candidates and that of the query

This geometric feature extraction results in a 20D spatial feature for each frame, or $20 \times 29D$ spatiotemporal feature for the video clip. We then re-rank our candidate using their temporal lip landmark correlation with the query lip video, as shown in Fig. 4b. Using recognition-free retrieval, top-k candidates are selected for a given query. Spatiotemporal features for both top-k candidates and query are extracted. The correlation of landmark of the lip region of these top-k candidates with the query is computed; the re-ranking is done in the order of decreasing correlation.

5 Experiments

5.1 Datasets

Lip-reading in the Wild (LRW) [14] has 500 words classes with 1000 clips for training, 50 for testing and 50 for validation for each of the words, which has been curated from BBC news videos. Each word clip is of length 1.16 s duration containing 29 frames. We use the LRW to train both feature extractors. The proposed retrieval pipeline only uses the test set for querying and validation set for retrieval, since training set has been used to train feature extractors.

GRID corpus [16] contains 1000 phrases, spoken by each of 33 speakers. Each phrase has a fixed syntax containing 6 words: $command(4) + color(4) + preposition(4) + letter(25) + digit(10) + adverb(4)$; an example of which is "put red at G 9 now." We use speakers 10–19, similar to [52], in our experiment. For showing domain invariance, we randomly sample 1000 phrases from these speakers to create our query set. Similarly, we sample another 1000 phrases from the same speakers to create our retrieval set. All the

Fig. 5 Random frames from LRW dataset (top row), GRID corpus (middle row) and Charlie Chaplin "*The great dictator*" speech video (bottom row)

speech videos are word segmented and preprocessed before feeding to feature extractors.

For qualitative results, we show lip-reading on **Charlie Chaplin's** famous "The great dictator" speech video. We only use the video, without audio cues for our experiment. The video is segmented into sentence level video clips using the time-stamps provided by YouTube subtitles, which also gives the ground truth annotations. The retrieval corpus is made by densely segmenting these sentence videos into word proposal clips. Randomly selected frames from these three datasets are shown in Fig. 5.

5.2 Implementation

For WAS, we use the pretrained VGG-M model from Chung and Zisserman [15], and only train attention sequence-to-sequence LSTM module, while freezing the weights of VGG-M module. We use the LRW training set for training our model, with validation set used for parameter tuning. The network has been trained with batch size 64, cross-entropy loss and SGD optimizer. Initial learning rate was set to 0.1 with a decay of 0.01% every two iterations. No data augmentation was used.

For training CMT, we follow the similar procedure as mentioned in Stafylakis and Tzimiropoulos [46] to train our model end to end. Again, the batch size of 64 was taken with cross-entropy loss and SGD optimizer was used. Initial learning rate was set to $3e^{-3}$ with exponential decay in learning rate when the validation loss does not decrease for 2 epochs. We also perform data augmentation with random cropping of 4 pixels around the lip region of interest (ROI), and horizontally flipping all frames of randomly chosen input clips. For both the networks, WAS and CMT, early stopping was employed if validation accuracy failed to improve over 3 consecutive epochs. We implement both the networks in Keras deep learning library [11].

Word spotting on LRW dataset has been shown considering LRW test set as query set and LRW validation set as retrieval set. Here, we want to assign label to the query video clips, considering we know the GT label for retrieval set. Both the query and retrieval set are first preprocessed, as discussed in Sect. 3.2. Since all the video clips are 29 frames long, circular padding is not required during preprocessing. After feature extraction, the query is searched in the retrieval set; the candidate with highest cosine similarity is ranked highest. To transfer word label from retrieval set the query, we take the majority vote of top-5 candidates in the retrieval set.

During query expansion, we first search a *seed* query in the retrieval set to get top-5 candidates. The "New query" is the weighted sum of the top-5 candidates with weights 0.1 each and *seed* query with weight 0.5, as shown in Fig. 7. This query is then used to retrieve a new set of candidates which becomes our final retrieval for the *seed* query.

For each query video coming from LRW test set, we retrieve top-10 candidates from LRW validation set using recognition-free retrieval. For Re-ranking, we then extract spatiotemporal feature for both query video and its top-10 retrieval candidates using DLib [38] and OpenCV [6] libraries. Correlation between spatiotemporal features of query and candidates was computed and was used to re-rank the top-10 candidates. This method proves to be effective in refining the search results for our retrieval pipeline.

For showing word spotting in Charlie Chaplin video, as shown in Fig. 6, the sentence videos are densely segmented into fixed length (29 frames) word proposal clips by taking stride of 3 frames. We spot the words in retrieval corpus consisting of these clips. Since the segmentation is dense there will be very few word proposal clips which will entirely cover actual words spoken in the video. As discussed in Sect. 3.4, we calculate the average similarity score between all the query exemplars coming from LRW validation set belonging to a particular word label and a word proposal clip from Charlie Chaplin video. If the average similarity is more than a threshold (ρ), we assign the word label to the word proposal clip. We empirically selected the value of $\rho = 0.3$ for this experiment.

5.3 Baselines

We compare our pipeline with recognition-based retrieval. WAS [13], in the original paper, was first pretrained on LRS dataset, and later fine-tuned on LRW dataset, gives a word accuracy of 76.2%. Our WAS model trained solely on LRW dataset gives the word accuracy of 53%. The recognition-based baseline of our WAS is given in Table 1, column 1. Another lip-reader CMT gives the word accuracy of 69.7%. The recognition-based baseline is given in Table 1 column 3.

For GRID corpus, we do not fine-tune our LRW trained base feature extractors on GRID corpus. The recognition-

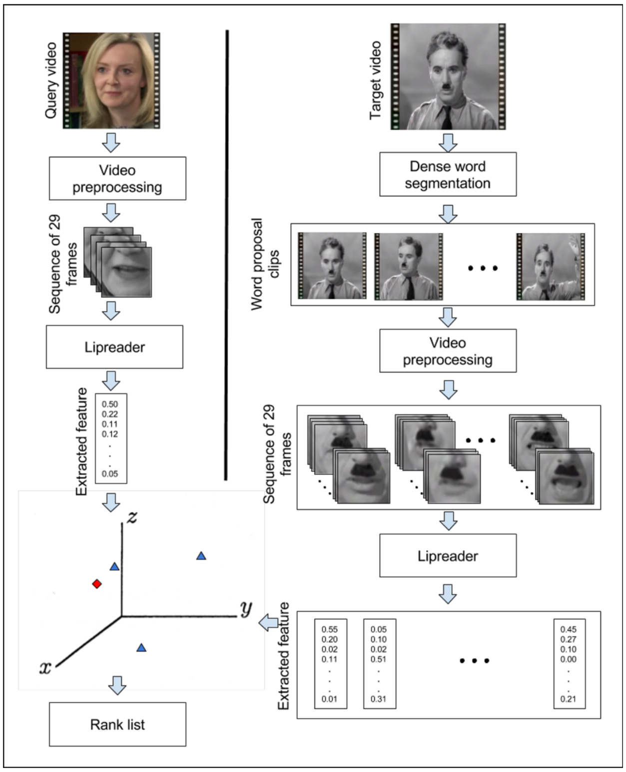

Fig. 6 Word spotting in Charlie Chaplin video: (left) a query exemplar with known annotation is preprocessed into fixed length input and fed to the feature extractor; (right) the Charlie Chaplin video is first densely segmented into word proposal clips and fed to the feature extractor. All the word proposal clips and query exemplar are projected into feature space, and ranking is computed based on cosine similarity

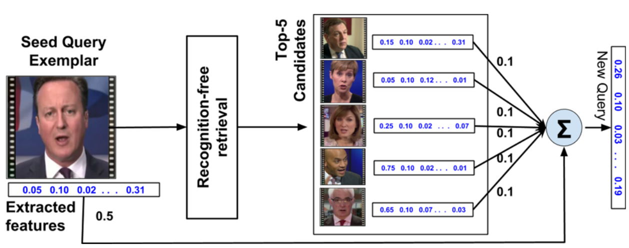

Fig. 7 Formulation of new query: the weighted sum of the feature representation of *seed* query and its top-5 retrieved candidates becomes the new query

based baseline for the domain invariance out-of-vocabulary retrieval is shown in Table 3, columns 1 and 3.

5.4 Evaluation metric

For search-based applications, the most important performance factor is: how many good results are in the top search results. Hence, Precision@K, which measures the precision at fixed lower levels of retrieval results, makes sense as an important performance metric. It considers the number of desirable results out of the top-k retrieval results without taking into account the overall rank ordering of the search results.

Recall@K is another important evaluation metric that we show, which is the number of desired results retrieved among top-k search results, with respect to the total number of available positive results.

While Precision@K and Recall@K give specific insights into the performance of the retrieval system, both measure performance for a fixed number of retrievals (K) and are insensitive to the overall rank ordering of the search results. We therefore also report the mean average precision (mAP) for our retrieval system. mAP provides a measure of the quality of retrieval across different recall levels. mAP has been shown to have especially good discrimination and stability, and is one of the most standard evaluation measures for word spotting.

Table 1 Retrieval performance for LRW dataset: Left two columns show recognition-based (RB) baseline and recognition-free (RF) performances for WAS features; right two columns show the similar results for CMT features. Across columns (first row) mAP is mean average precision, (second row) P@10 is precision at 10, (third row) R@10 is recall at 10, and (last row) % imp.in mAP is percentage mAP improvement of recognition-free retrieval over baseline

	WAS		CMT	
	RB (BL)	RF (ours)	RB (BL)	RF (ours)
mAP	0.2317	**0.3149**	0.3807	**0.5698**
P@10	0.2928	0.4566	0.3253	0.6519
R@10	0.0586	0.0913	0.0651	0.1304
% imp.in mAP	–	**35.90**	–	**49.67**

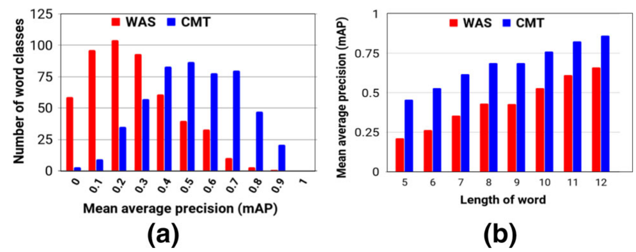

Fig. 8 **a** Number of words below a certain mAP for WAS and CMT-based pipeline: *y*-axis is the number of words, and *x*-axis is the mAP; **b** variation of mean average precision (mAP) with the length of the word for CMT- and WAS-based pipeline: *y*-axis is average mAP, and *x*-axis is word length in LRW vocabulary

6 Results

6.1 Comparison with baseline methods

Recognition-free retrieval or word spotting on LRW dataset when the base lip-reader is WAS gives an absolute improvement of 35.9% over the recognition-based baseline of mAP

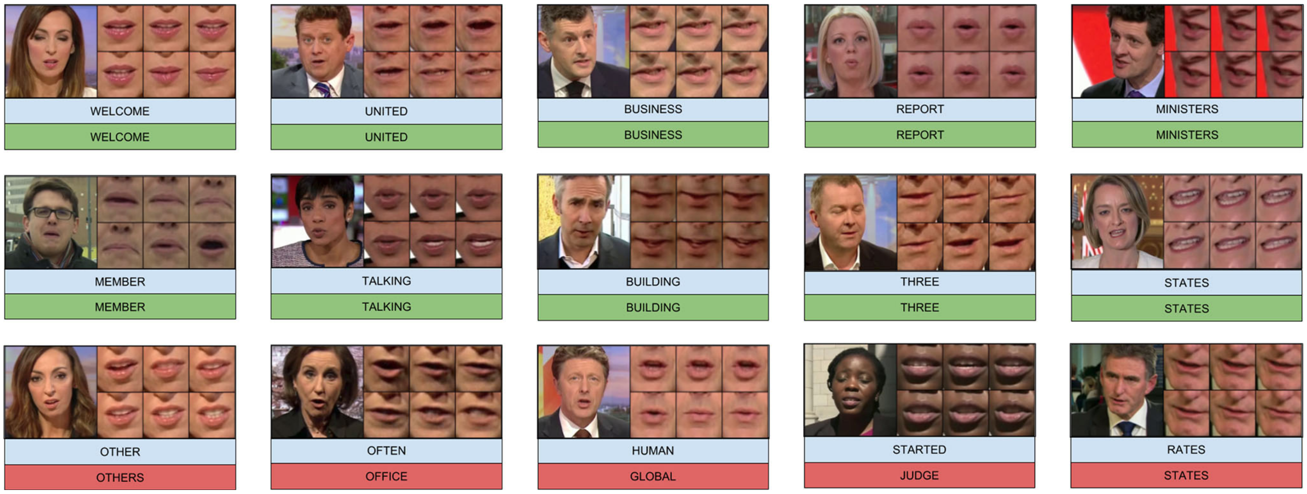

Fig. 9 Qualitative results on LRW dataset: each image depicts the central frame of the query video clip (left) and a sequence of lip ROIs of 6 consecutive frames around central frame, shown in raster order (right); (middle) blue boxes are the ground truths; (bottom) green boxes are correct predictions, while the red ones are incorrect predictions. Label is propagated to a query based on the majority label present in the top-5 retrieval candidates (color figure online)

0.23; Table 1, column 2. Similarly, for recognition-free retrieval using CMT lip-reader there is an improvement of 49.67% over the recognition-based baseline of mAP 0.38; Table 1, column 4. For recognition-free retrieval using WAS (in red) and CMT (in blue) feature extractor, Fig. 8a shows the number of words below a certain mAP value. The variation of average mAP with the length of the words in the LRW vocabulary is shown in Fig. 8b. It can be seen that the average mAP value increases with the increase in word length. The qualitative results for word spotting on LRW dataset using CMT features can be seen in Fig. 9.

Query expansion on LRW dataset using two lip-readers: WAS and CMT give a mAP of 0.3146 and 0.5722, respectively; Table 2, columns 2 and 5. Although the mAP results are comparative to the recognition-free method, we see an overall increase in Recall@10. Also, re-ranking using spatiotemporal cues improves the retrieval performance for WAS and CMT, giving a mAP of 0.3179 and 0.5709, respectively, Table 2, columns 3 and 6.

Charlie Chaplin "*The great dictator*" speech video, contains 39 words from LRW vocabulary. It has a total of 54 spoken sentences, out of which 33 sentences actually contain LRW vocabulary words. Hence, the query set contains 50 exemplars, from LRW validation set, belonging to each of these 39 common vocabulary words. Using our CMT-based recognition-free pipeline, we were able to correctly spot instances of 13 instances of the common vocabulary words in 11 sentences, whereas on using recognition-based pipeline, only 6 instances of common vocabulary words in 6 sentences are correctly predicted. The qualitative results can be seen in Fig. 10, where we spot the sentences which contain the query words.

Table 2 Different recognition-free performance for LRW dataset: Left three columns are recognition-free (RF), query expansion (QExp) and re-ranking (ReR) performances for WAS features; right three columns show similar results for CMT features. Across columns (first row) mAP is mean average precision, (second row) P@10 is precision at 10, and (last row) R@10 is recall at 10

	WAS			CMT		
	RF	QExp	ReR	RF	QExp	ReR
mAP	0.3149	0.3146	**0.3179**	0.5698	0.5722	**0.5709**
P@10	0.4566	0.4591	0.4566	0.6519	0.6572	0.6519
R@10	0.0913	**0.0918**	0.0913	0.1304	**0.1314**	0.1304

6.2 Domain invariance

Domain invariance provides us the robustness of the pipeline for target data distribution different from the one it is trained on. GRID corpus contains 51 words with only 1 common word available in LRW dataset vocabulary. Hence, this experiment also shows out-of-vocabulary retrieval performance of the proposed pipeline.

On GRID corpus, the recognition-based baseline is 0.033 (mAP) for WAS features and 0.06 (mAP) for CMT features, while the recognition-free performance is 0.068 (mAP)for WAS and 0.177 (mAP) for CMT, Table 3, column 2. This signifies the utility of recognition-free retrieval for out-of-vocabulary words when the underlying lip-reader is constrained by vocabulary size.

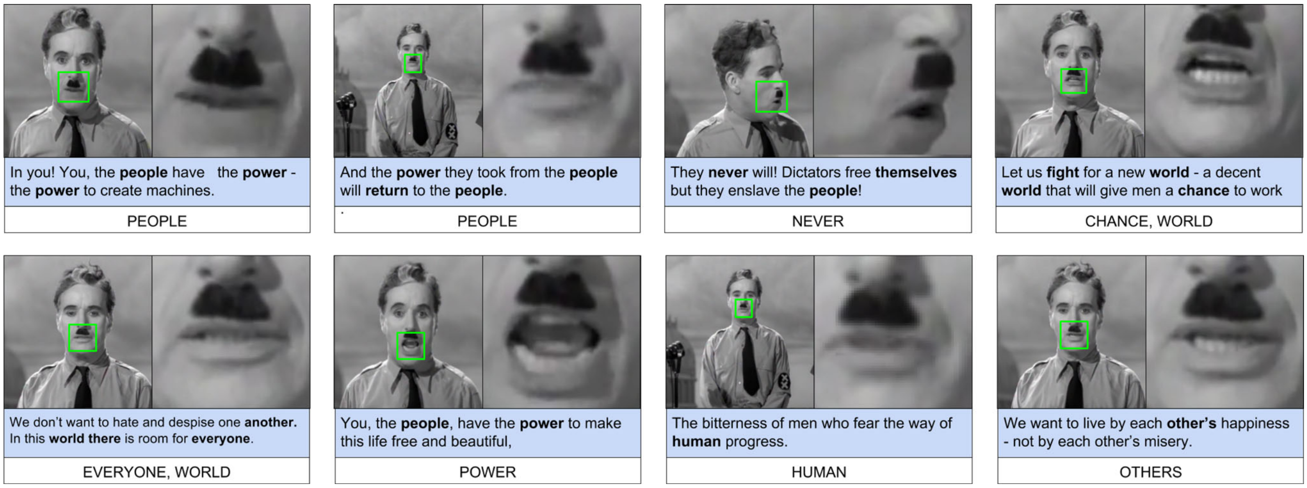

Fig. 10 Qualitative results on Charlie Chaplin "*The great dictator video*": each image is one of the frames in the sentence clips extracted from the speech video. The top text box in blue color contains the subtitles with **bold** text showing the common LRW vocabulary word present in the subtitle. The bottom text box shows the correctly spotted word (color figure online)

Table 3 Domain invariance results on Grid corpus dataset (for both WAS and CMT): Left column has recognition-based (RB) baseline performance and right has our recognition-free (RF) performance where (first row) mAP is mean average precision, (second row) P@10 is precision at 10, (third row) R@10 is recall at 10, and (last row) % imp.in mAP is the percentage mAP improvement of our proposed method over baseline

	WAS		CMT	
	RB (BL)	RF(ours)	RB (BL)	RF(ours)
mAP	0.033	**0.068**	0.060	**0.177**
P@10	0.034	0.219	0.224	0.322
R@10	0.002	0.016	0.019	0.020
% imp.in mAP	–	**106**	–	**195**

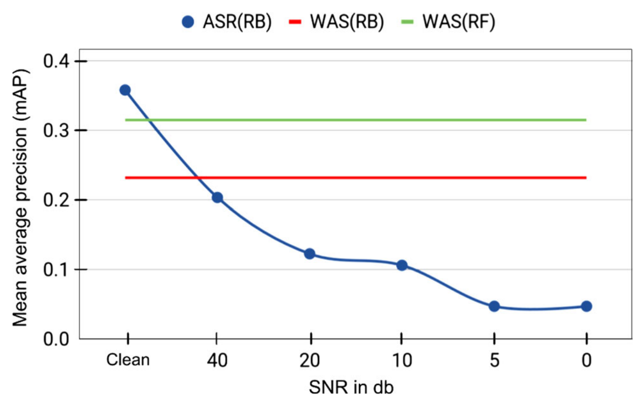

Fig. 11 ASR performance with respect to lip word spotting: (blue) shows mAP of ASR prediction for different SNR on LRW test set, while (red) shows the mAP of WAS-based recognition-based (RB) retrieval on corresponding lip videos. Similarly, (green) denoted the WAS recognition-free (RB) performance (color figure online)

6.3 ASR versus lip word spotting

One of the major applications of word spotting in lip domain can be spotting keywords in a noisy environment. This is particularly useful in present-day scenario as voice assisted technology is emerging as a new way of human–computer interface (HCI). For this technology to work effectively in an ambient noise environment, like inside cars, public transit, industries, streets, etc., one could exploit visual cues by observing the lip motion. Here, we compare the performance of automatic speech recognition (ASR) system in the noisy environment, against the lip word spotting pipeline. We use a pretrained ASR system called DeepSpeech [28], and generate predictions for all the samples in the LRW test set. We first extract the audio from the samples and generate predictions, we repeat the experiment while introducing the same amount of white noise in each test sample, simulating the ambient noise. The ASR mAP scores can be seen in Fig. 11, in blue curve, while the mAP for our lip recognition-free word spotting pipeline is in green line with recognition-based counterpart in red line. We find that initially the ASR performs better than lip word spotting, when the introduced noise is minimum, but decreases drastically in the presence of noise. It should be noted that, the introduced noise in this scenario is ambient, while during cross talk and crowd situation the noise can be intelligible speech, thereby further degrading the ASR performance.

7 Analysis

7.1 Qualitative analysis of the retrieval pipeline

We performed a qualitative analysis of the label assignment stage, to complement the results shown in Table 1 and Fig. 8.

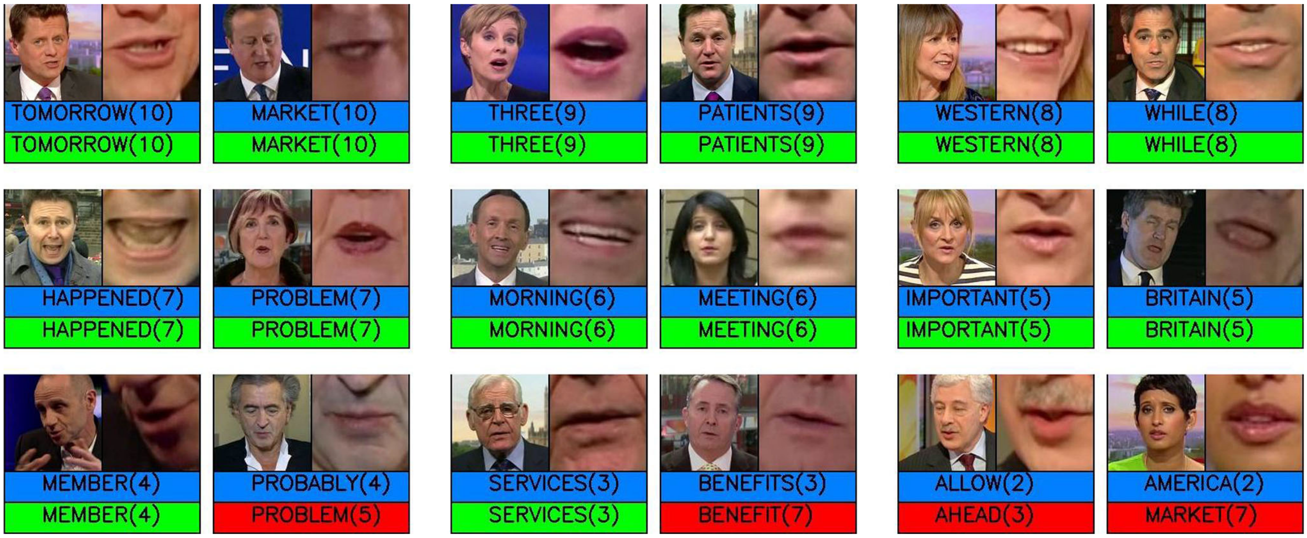

Fig. 12 Qualitative performance of the pipeline: for each image, (top) depicts a randomly sampled frame from the query video clip and its lip ROI; (middle) blue boxes are the ground truths with (.) denoting number of the top-10 candidates in the retrieval set having the same label as the ground truth(GT) of the input query exemplar; (bottom) green boxes are correct predictions, while the red ones are incorrect predictions, with (.) denoting number of the top-10 candidates in the retrieval set having the predicted label. Label is propagated to a query based on the majority label present in the top-10 retrieval candidates (color figure online)

For each video clip in LRW test set as a query sample, we retrieved top-10 candidates from the retrieval set whose ground truth was known, using recognition-free retrieval. Based on the majority voting, we assigned a label to the query. We also calculated the number of samples in top-10 retrieval candidates belonging to the true class of the query. This allowed us to investigate the failure cases, as shown in Fig. 12. We found numerous cases where the failure in the prediction is due to prediction of a visemically similar word or same root word as that of the ground truth of the query (see for example Fig. 12 (bottom row). Hence, the number of retrieval candidate considered for label assignment is an important hyperparameter. The precision curve and recall curve for different value of k can be seen in Fig. 13.

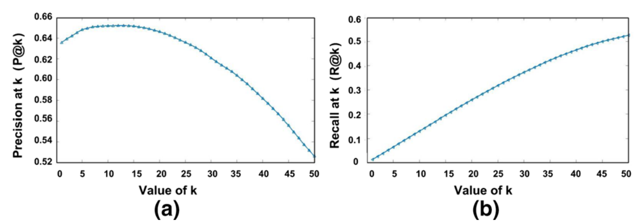

Fig. 13 Precision and Recall at k on LRW test set: **a** shows precision at k ($P@k$), on y-axis, for different values of k, on x-axis. Similarly, **b** shows recall at k ($R@k$), on y-axis, for different values of k, on x-axis

7.2 Dependence on quality of lip-reader

Our word spotting pipeline uses recognition-based networks: CMT and WAS as the feature extractor. Hence, the performance of the proposed pipeline is intrinsically dependent on the quality of underlying lip-reading architecture. Therefore, we investigate this dependence on feature extractor with different lip-reader quality. Moreover, a detailed analysis of our pipeline with different quality of feature extractors can give us the performance trend of the word spotting pipeline. This trend enables us to predict the performance of the proposed pipeline if a better lip-reader is used for feature extraction.

To obtain this relationship between the performance of the recognition-based and the recognition-free pipeline, we require different feature extractors with varying lip-reading performance. Hence, we train our CMT lip-reader [46] till different epochs to get lip-reading models with validation accuracy ranging from 10 to 70% word accuracy on LRW dataset [14]. This acts as a proxy for using lip-reader with different quality as a feature extractor in our word spotting pipeline. The graph showing word spotting performance of recognition-based and the recognition-free pipeline for different quality of CMT feature extractor can be seen in Fig. 11.

We observed that the recognition-free retrieval, shown by red, Fig. 14 (top), always gives better performance than recognition-based retrieval, shown by blue, for different quality of lip-readers used as feature extractors. A second-order polynomial extrapolation , shown by translucent trend line of the different performance curves, shows that our proposed recognition-free word spotting pipeline may perform better than its recognition-based counterpart in case we find a bet-

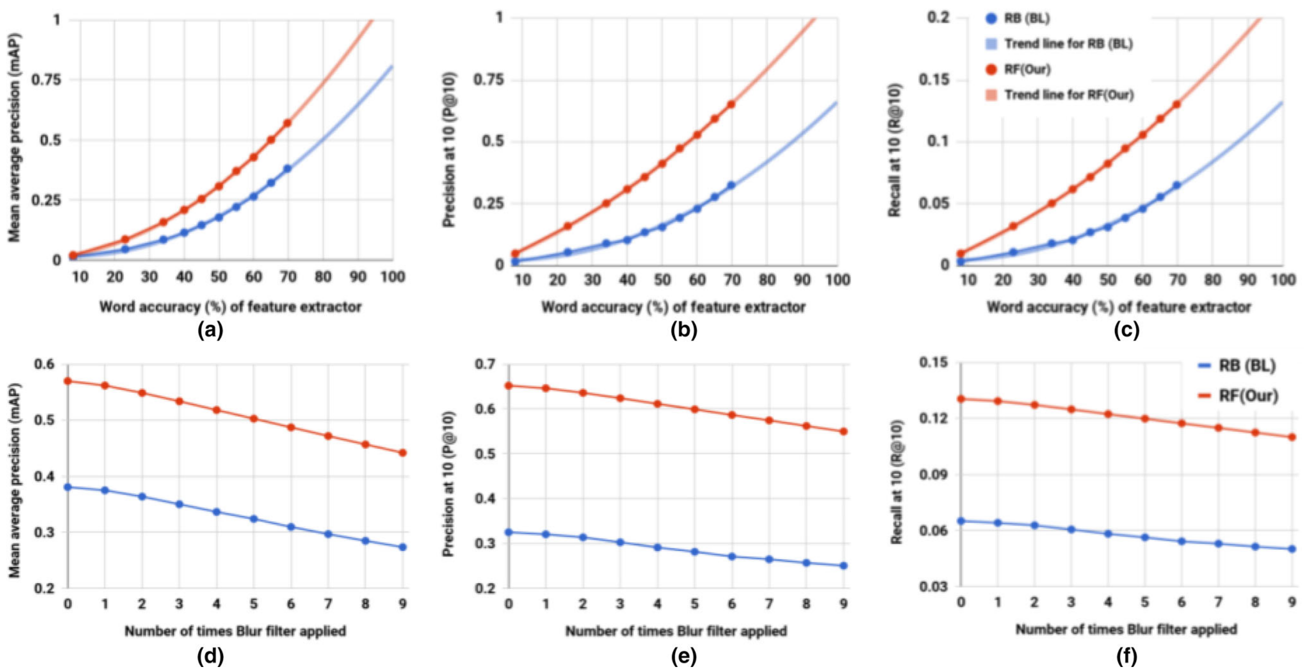

Fig. 14 (Top)Dependence of word spotting pipeline on the quality of feature extractor(lip-reader): **a** shows change in mean average precision(mAP), **b** shows change in precision at 10 ($P@10$), and **c** shows the change in recall at 10 ($R@10$) for different quality of lip-reader (x-axis). Blue line-dot is the empirical data for recognition-based (RB) recognition also the baseline (BL), translucent blue line is its second-order polynomial extrapolation. Similarly, red is for our proposed recognition-free pipeline. (Bottom)Dependence of word spotting pipeline on quality of input video: **d** shows change in mean average precision (mAP), **e** shows change in precision at 10 ($P@10$), and **f** shows change in recall at 10 ($R@10$) for different quality of input video(x-axis), subjected to successive blurring as discussed in Fig. 15 (color figure online)

ter lip-reader. This shows that the proposed pipeline is robust against the quality of lip-reader.

7.3 Dependence on quality of input video

Ubiquity of cellular phone cameras has made it one of the major video capturing devices nowadays. In real-life scenario, videos taken from these devices may not be face centric and region of interest (ROI) can be small or blurred due to lack of focus or motion. Hence, robustness against input video quality is desirable for any word spotting pipeline.

We compare the performance of our proposed recognition-free pipeline against the recognition-based counterpart for different quality of input videos. We blur the frames of input video using 3×3 averaging kernel, as shown in Fig. 15. To degrade the quality of input video, this blurring operation is applied recursively. In each blurring cycle, the blurred video is fed to the word spotting pipeline. We, then, compare the recognition-free retrieval with our baseline recognition-based counterpart for three different evaluation metrics. The results are shown in Fig. 14 (bottom).

We observe that the performance of both recognition-free retrieval and recognition-based retrieval degrades monotonically. Recognition-free performance is steadily better than the recognition-based counterpart for different passes of averaging kernel on input video. Our proposed pipeline outperforms the baseline even when the input video quality gets severely degraded, i.e., after successive 9 blurs as shown in Fig. 14a–c. After 9 blurring cycles the drop in the mAP of recognition-free was 22.49%, while for recognition-based it was 28.11%, implying the proposed pipeline is more robust to the degradation in the quality of input video. Hence, our pipeline may be useful for spotting lips for distant speaker or for surveillance purposes.

7.4 Discussions

Many conclusions can be drawn from the result presented in Sect. 6.1. Recognition-free retrieval performed better than the recognition-based counterpart for spotting words in LRW dataset. From Fig. 8b, we see that the quality of retrieval improves when the length of word increases, as longer the word is more the number of phonemes it contains, and less is the chance of it being similar to other words. Errors in similar sounding words are more likely, as can also be seen in Fig. 9. Failure in word spotting can also occur due to presence of similar root word in the retrieval set, as shown in Fig. 12. Many such words can be tolerated when task-in-hand requires coarser word perception.

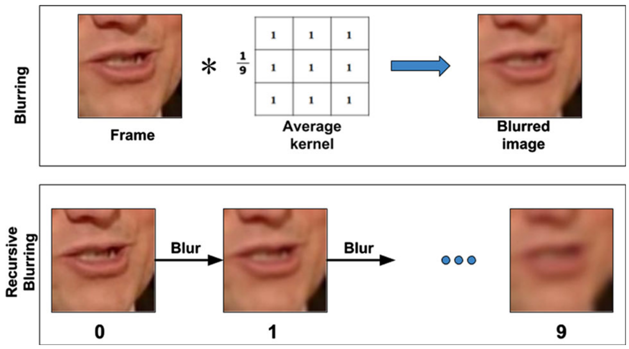

Fig. 15 Blurring operation: (top) shows blurring of the region of interest (ROI) of a frame of the input video clip, (bottom shows successive blurring operation on the ROI of the input video clip; (bottom) The number below the ROI of the frames shows the number of times blurring filter/kernel has been applied on the original input video

The performance of recognition-based retrieval on GRID corpus is inferior to that on LRW dataset, the reason being neither of the two feature extractors in our experiments were fine-tuned on GRID corpus. Still, the recognition-free retrieval showed an improvement over recognition-based. Quality of lip video is also important, as some words in Charlie Chaplin videos were not spotted, due to lower contrast and quality of the lip ROI, as shown in Fig. 10.

In the presence of white noise, ASR performance decreases drastically, while lip word spotting being independent of audio, shows consistent performance. Hence, complimentary use of visual modality can enhance the performance of ASR in noisy environment.

8 Conclusion

We proposed a recognition-free retrieval pipeline and showed its precedence over recognition-based retrieval for the task of word spotting. The base features from WAS and CMT lip-reading models have been used to spot words in LRW dataset with an improvement of about 36% and 50% over the recognition-based counterpart. Pseudo-relevance feedback and re-ranking techniques, using spatiotemporal geometrical cues available in the lip videos, has been incorporated in the pipeline to further improve the retrieval results. We also showed domain invariance of our pipeline through out-of-vocabulary word spotting on GRID corpus dataset with an improvement of 106% and 195% over the baseline using WAS and CMT features, respectively. We presented the practical applicability of our proposed pipeline by spotting words in 11 out of 33 sentences in the "Charlie Chaplin, *The great dictator*" speech video. We showed that in the presence of noise our method performs better than ASR. We analyzed how the selection of number of retrieved candidates can be crucial for retrieval performance. Finally, we empirically showed the robustness of our pipeline against the performance of the underlying lip-reader and the quality of input videos.

Acknowledgements This work is partly supported by Alexa Graduate Fellowship from Amazon.

References

1. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR (2012)
2. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014). arXiv preprint arXiv:1409.0473
4. Basu, S., Oliver, N., Pentland, A.: 3d modeling and tracking of human lip motions. In: ICCV (1998)
5. Bourlard, H.A., Morgan, N.: Connectionist Speech Recognition: A Hybrid Approach, vol. 247. Springer, Berlin (2012)
6. Bradski, G.: The opencv library. Dr. Dobb's J.: Softw. Tools Prof. Progr. **25**(11), 120, 122–125 (2000)
7. Brooke N.M, S.S.: Pca image coding schemes and visual speech intelligibility. In: Proceedings of the Institute of Acoustics, vol. 16 (1994)
8. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: ICASSP, pp. 4960–4964 (2016)
9. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
10. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation (2014). arXiv preprint arXiv:1406.1078
11. Chollet, F., et al.: Keras. https://keras.io (2015)
12. Chorowski, J., Jaitly, N.: Towards better decoding and language model integration in sequence to sequence models (2016). arXiv preprint arXiv:1612.02695
13. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: CVPR (2016)
14. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: ACCV (2016)
15. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV (2016)
16. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
17. Doetsch, P., Kozielski, M., Ney, H.: Fast and robust training of recurrent neural networks for offline handwriting recognition. In: ICFHR (2014)
18. Fergus, R., Perona, P., Zisserman, A.: A visual category filter for google images. In: ECCV (2004)
19. Fernández, S., Graves, A., Schmidhuber, J.: An application of recurrent neural networks to discriminative keyword spotting. In: ICANN (2007)
20. Fischer, A., Keller, A., Frinken, V., Bunke, H.: HMM-based word spotting in handwritten documents using subword models. In: ICMR (2010)
21. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)

22. Frinken, V., Fischer, A., Manmatha, R., Bunke, H.: A novel word spotting method based on recurrent neural networks. IEEE TPAMI **34**(2), 211–224 (2012)
23. Giotis, A.P., Sfikas, G., Gatos, B., Nikou, C.: A survey of document image word spotting techniques. Pattern Recognit. **68**, 310–332 (2017)
24. Gish, H., Ng, K.: A segmental speech model with applications to word spotting. In: ICASSP, vol. 2 (1993)
25. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: ICML (2006)
26. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. In: ICANN (2005)
27. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: ICML, pp. 1764–1772 (2014)
28. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: scaling up end-to-end speech recognition (2014). arXiv preprint arXiv:1412.5567
29. Hassanat, A.B.: Visual words for automatic lip-reading (2014). arXiv preprint arXiv:1409.6689
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
31. Hennecke, M.E.: Audio-visual speech recognition: preprocessing, learning and sensory integration. PhD thesis, Stanford Univ. (1997)
32. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
33. Ho, T.K., Hull, J.J., Srihari, S.N.: A computational model for recognition of multifont word images. Mach. Vis. Appl. **5**(3), 157–168 (1992)
34. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
35. Jha, A., Namboodiri, V., Jawahar, C.V.: Word spotting in silent lip videos. In: WACV (2018)
36. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE TPAMI **35**(1), 221–231 (2013)
37. Keshet, J., Grangier, D., Bengio, S.: Discriminative keyword spotting. Speech Commun. **51**(4), 317–329 (2009)
38. King, D.E.: Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
39. Krishnan, P., Jawahar, C.V.: Bringing semantics in word image retrieval. In: ICDAR (2013)
40. Lee, J.S., Park, C.H.: Robust audio-visual speech recognition based on late integration. IEEE TMM **10**(5), 767–779 (2008)
41. Liu, H., Fan, T., Wu, P.: Audio-visual keyword spotting based on adaptive decision fusion under noisy conditions for human-robot interaction. In: ICRA, pp. 6644–6651 (2014)
42. Manmatha, R., Han, C., Riseman, E.M.: Word spotting: A new approach to indexing handwriting. In: CVPR (1996)
43. Mohamed, A.R., Dahl, G.E., Hinton, G., et al.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**(1), 14–22 (2012)
44. Robinson, T., Hochberg, M., Renals, S.: The use of recurrent neural networks in continuous speech recognition. In: Automatic Speech and Speaker Recognition, pp. 233–258. Springer, Berlin (1996)
45. Rohlicek, J.R., Russell, W., Roukos, S., Gish, H.: Continuous hidden Markov modeling for speaker-independent word spotting. In: ICASSP (1989)
46. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with LSTMs for lipreading (2017). arXiv preprint arXiv:1703.04105
47. Stafylakis, T., Tzimiropoulos, G.: Zero-shot keyword spotting for visual speech recognition in-the-wild (2018). arXiv preprint arXiv:1807.08469
48. Stillittano, S., Girondel, V., Caplier, A.: Lip contour segmentation and tracking compliant with lip-reading application constraints. Mach. Vis. Appl. **24**(1), 1–18 (2013)
49. Sudholt, S., Fink, G.A.: Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In: ICFHR (2016)
50. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS (2014)
51. Tsai, S.S., Chen, D., Takacs, G., Chandrasekhar, V., Vedantham, R., Grzeszczuk, R., Girod, B.: Fast geometric re-ranking for image-based retrieval. In: ICIP (2010)
52. Wand, M., Koutník, J., Schmidhuber, J.: Lipreading with long short-term memory. In: ICASSP (2016)
53. Wang, K., Belongie, S.: Word spotting in the wild. In: ECCV (2010)
54. Wu, P., Liu, H., Li, X., Fan, T., Zhang, X.: A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE TMM **18**(3), 326–338 (2016)
55. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: Achieving human parity in conversational speech recognition (2016). arXiv preprint arXiv:1610.05256
56. Zhang, X.Y., Yin, F., Zhang, Y.M., Liu, C.L., Bengio, Y.: Drawing and recognizing chinese characters with recurrent neural network. IEEE TPAMI 849—862 (2017)
57. Zhou, Z., Zhao, G., Hong, X., Pietikäinen, M.: A review of recent advances in visual speech decoding. Image Vis. Comput. **32**(9), 590–605 (2014)

Abhishek Jha is a master's student in Computer Science at IIIT Hyderabad. He is also a research assistant at Center for Visual Information Technology (CVIT), IIIT Hyderabad. He received his bachelor's degree from IIT(ISM) Dhanbad in 2014. His research focuses on computer vision and machine learning for solving problems at the intersection of multiple modalities like video, audio, and text.

Vinay P. Namboodiri is an Associate Professor at Computer Science and Engineering department at IIT Kanpur. He received his Ph.D. from IIT Bombay in 2008. He was a Postdoctoral fellow in KULeuven, Belgium, and subsequently worked at Bell Labs, Antwerp before joining IIT Kanpur. His research interests include computer vision and machine learning with a focus on deep learning-based research with scarce supervision, vision and language and unsupervised learning. He has around 50 publications in journals and conferences in computer vision. His work has received a best paper award in BMVC 2011 and a Microsoft Techvista award during Ph.D.

C. V. Jawahar is a professor at IIIT Hyderabad, India. He has been with IIIT Hyderabad since 2000. At IIIT Hyderabad, Jawahar leads a group focusing on computer vision, machine learning and multimedia systems. In the recent years, he has been looking into a set of problems that overlap with vision, language and text. He is also interested in large scale multimedia systems with special focus on retrieval. He has more than 50 publications in top tier conferences in computer vision, robotics and document image processing. He has served as a chair for previous editions of ACCV, WACV, IJCAI, ICCV and ICVGIP. Presently, he is an area editor of CVIU and an associate editor of IEEE PAMI.