Lipreading Using Spatiotemporal Histogram of Oriented Gradients

Karel Paleček
The Institute of Information Technology and Electronics
Technical University of Liberec
Czech Republic
Email: karel.palecek@tul.cz

***Abstract*—We propose a visual speech parametrization based on histogram of oriented gradients (HOG) for the task of lipreading from frontal face videos. Inspired by the success of spatiotemporal local binary patterns, the features are designed to capture dynamic information contained in the input video sequence by combining HOG descriptors extracted from three orthogonal planes that span x, y and t axes. We integrate our features into a system based on hidden Markov model (HMM) and show that by utilizing robust and properly tuned parametrization this traditional scheme can outperform recent sophisticated embedding approaches to lipreading. We perform experiments on three different datasets, two of which are publicly available. In order to conduct an unbiased feature comparison, the process of model learning including hyperparameter tuning is as automatized as possible. To this end, we rely heavily on cross validation.**

I. Introduction

It has been repeatedly shown that in humans, understanding speech is a multi-modal process. Probably the most famous example of this fact is the well known McGurk effect. It illustrates how the apparent movement of speaker's lips might influence the actual acoustic perception. However, the visual cues alone generally do not carry enough information for reliable speech understanding. Therefore, much of the work in the area of automatic lip-reading only considers simpler scenarios such as recognition with small vocabulary or targets some specific sub-problem, e. g. audio-visual fusion [1]. An overview of recent advances in lipreading is presented in [2].

In last several years, lipreading systems based on graph embedding and manifold learning algorithms have become quite popular. Rather than manually crafting ideal features, such systems exploit sophisticated modeling techniques in order to project the high dimensional input to a more discriminative subspace better suited for classification. For example, in [3] Zhou et al. treated the input sequences as graphs, where each node represents single frame and edges denote adjacency both in time and across speakers. They used their framework for embedding and length-normalizing input videos in order to improve classification with SVM. Another example is [4], where Pei et al. fused several features via multidimensional scaling (MDS) with utilization of random forest for efficient computation of affinity matrix and classified the embedded sequences by a manifold alignment algorithm.

However, the main disadvantage of similar approaches is their inapplicability to recognition based on sub-word units, e. g. continuous speech recognition. The projection algorithms behave essentially as static classifiers, meaning that the whole utterance must first be normalized to a specified length before it can be classified as a single feature vector. This makes them closely tied to the target application, e. g. isolated phrase or digit recognition. Thus, although interesting from an algorithmic point of view, utilization of these methods in real world applications remains an open question.

In this work, we follow the traditional scheme with the well established hidden Markov model (HMM) as a classification algorithm and instead of modeling techniques we focus on good feature design. Our parametrization relies on histogram of oriented gradients (HOG) that was originally introduced in Dalal's and Triggs' seminal paper [5] as a robust and discriminative descriptor for automatic pedestrian detection. Later, it was also applied for lipreading, see e. g. [4], [6].

Normally the HOG descriptor is extracted for each input image or video frame individually and thus cannot capture the dynamics of speech, which is essential for speech recognition. One of the first modifications of HOG-like features for lipreading designed to capture speech dynamics was presented in [7]. Pachoud et al. viewed the input video as a three dimensional structure, which they then subsequently divided into partially overlapping regions called macro-cuboids. On each such region, they computed generalized 3D SIFT descriptor and classified resulting sequences by aligning the input to trained data at multiple scales. Another dynamic modification proposed by the original authors of HOG was applied for lipreading by Rekik et al. [8]. Here, the HOG descriptor was extracted from x and y images of optical flow and concatenated into single vector, thus capturing the change between two consecutive frames.

In this paper, we propose a dynamization of HOG descriptor that is inspired by the spatiotemporal local binary patterns (LBP) introduced by Zhao et al. in [9]. LBP describes the texture in terms of a histogram of binary numbers that are formed by comparing each pixel of the image to its close neighborhood. Zhao et al. extended the static LBP by considering the neighborhood not only in the spatial domain, but also in time axis. As the descriptor was extracted from three orthogonal planes (TOP), i. e. xy, xt and yt, they named the parametrization as LBPTOP. Here, we modify the static HOG descriptor in a similar fashion, i. e. by computing gradients in

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

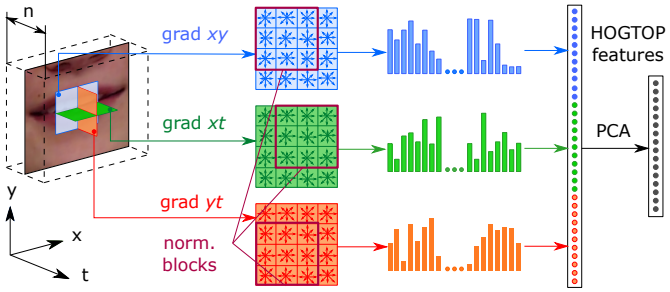

Fig. 1. Extraction of spatiotemporal histogram of oriented gradients.

three directions and combining the result into visual speech parametrization.

II. Spatiotemporal Histogram of Oriented Gradients

Extraction of the HOG descriptor from a single frame proceeds as follows. The input image is divided into regular grid of small rectangular cells with a typical size of 8×8 pixels. For each pixel (x, y) of each cell, a local gradient (g_x, g_y) is calculated e. g. by symmetrical difference or by using a Sobel filter. Weighted by their magnitude, the gradient orientations are then accumulated into a local orientation histogram $\boldsymbol{h}$ with b bins. In order to achieve robustness to local contrast variations, histograms from $m \times n$ blocks of neighboring cells are concatenated into a single vector $\boldsymbol{v}$ and normalized by a multiplicative factor $a = (\|\boldsymbol{v}\|_k^k + \epsilon^k)^{-1/q}$, where $\epsilon^k \ll \|\boldsymbol{v}\|_k^k$ prevents division by zero for regions without a texture and commonly $k = q = 2$. The image descriptor is formed by concatenating all normalized vectors $\boldsymbol{v}$ into a single hypervector $\boldsymbol{x}$. Note that due to local contrast normalization, each particular histogram $\boldsymbol{h}$ enters up to mn different blocks and therefore the resulting image descriptor $\boldsymbol{x}$ is highly redundant. In case of 64×64 image, 8 orientation bins and 3×3 cells per block, the dimension of $\boldsymbol{x}$ is 2592, whereas without local contrast normalization it is only 512.

In order to capture speech dynamics for our target application, instead of only calculating spatial image gradient $\boldsymbol{g}_{xy}$, we also extract gradients from xt and yt planes where the time axis t spans the preceding and following frames. Each of the three gradient images is processed statically, i. e. by forming orientation histograms with subsequent local contrast normalization. The descriptors $\boldsymbol{x}_{xy}$, $\boldsymbol{x}_{xt}$ and $\boldsymbol{x}_{yt}$ corresponding to their respective planes are concatenated into a single vector and due to high redundancy and dimensionality decorrelated and reduced by principal component analysis (PCA). We capture dynamics longer than just two neighboring frames by approximating the derivative in t axis by a convolution with non-causal difference of Gaussian kernel of length 7. Contrary to the original work by Dalal and Triggs, we also apply this kernel for x and y directions. Analogously to [9], we denote the resulting features as Histogram of Oriented Gradient from Three Orthogonal Planes (HOGTOP). The whole extraction procedure is depicted in Fig. 1.

III. System Overview

The visual speech features are calculated on the 64×64 pixel region of interest (ROI), whose extraction consists of three steps. First, an approximate position of the face is estimated using the well known Viola-Jones algorithm. Second, precise facial shape represented by 93 facial landmarks is found by utilizing the Explicit Shape Regression method (ESR) [10]. The ESR represents a face alignment technique that takes a discriminative rather than optimization-based approach to both learning and fitting. However, due to fixed number of iterations and the lack of objective function the final landmark positions are slightly different in each frame, which introduces an inter-frame jitter. We reduce it by running the detector from 10 random perturbations of the Viola-Jones detection and taking the median of the fit shapes. The ROI is then considered to be a square area barely covering the mouth and its closest surroundings. In order to achieve scale invariance we define its size relative to the normalized mean facial shape. The coordinates of the ROI in the input image are then found by computing Euclidean transformation between the normalized shape and the detected one via least squares minimization. To further reduce the inter-frame landmark jitter and stabilize the ROI extraction, we average the fitting results over three neighboring frames in time. Example ROIs obtained from Kinect video and depth streams are shown in Fig. 2.

All the visual speech features considered in the experiments are extracted densely for each frame (ROI) of the input video and subjected to several steps of post-processing. First, the feature vectors are reduced to several tens of coefficients. Then, sequences of $2K + 1$ parametrizations centered around the current frame are concatenated into a single hypervector and reduced by linear discriminant analysis (LDA). We obtain the class labels for LDA by force-aligning the training utterances on a phoneme level. Finally, mean feature vector is subtracted from the whole utterance and features eventually coupled with their Δ coefficients (first order difference from previous frame). The resulting sequence is then fed into the classifier. All of the hyperparameters such as the optimal number of feature coefficients, K or inclusion of Δ features, are cross-validated, see section V for details.

For classification we apply the hidden Markov model (HMM) with Gaussian mixture emission probability as implemented in the HTK 3.4.1 toolkit. Since in the experiments we perform recognition of isolated words and phrases, we build separate HMM for each utterance in the vocabulary. The vocabulary sizes along with other information for each of the three datasets are presented in Table I.

IV. Data

We evaluate the proposed features on three different datasets, two of which are publicly available.

TULAVD is our own dataset recorded at the Technical University of Liberec. It contains data from 54 speakers, of which 23 are female and 31 male with age ranging from 20 to 70 years. Each speaker uttered 50 isolated words and 100 sentences in Czech language, which were automatically

TABLE I
DATA SPLIT AND EVALUATION PROTOCOLS USED IN THE EXPERIMENTS.

dataset	vocabulary	# speakers	split	protocol
TULAVD	50 words	54	36:9:9	6×CV
OuluVS	10 phrases	20	19:1	20×LOOCV
CUAVE	10 digits	36	24:6:6	6×CV

selected according to phonetic balance. Audiovisual utterances were captured by two Logitech C920 FullHD webcams and Microsoft Kinect, which also offers depth stream that is fully synchronized with the video. Only the Kinect RGBD data with resolution of 640×480 pixels at 30 fps from the isolated words part of the dataset is used in this work. In order to achieve sufficient resolution and depth precision, the speakers were positioned approximately 80 cm from the sensor. Moreover, we minimize the noise in the depth stream by linearly interpolating all missing values (zeros) using neighboring values.

OuluVS [9] is a popular publicly available dataset containing 20 speakers (17 male, 3 female), each of which utters 10 different short phrases five times. Examples of such phrases are for instance "Hello!" or "How are you?". The videos were recorded at 25 fps with resolution of 720×576 pixels in an interlaced mode. Even though OuluVS ships with four different kinds of pre-extracted ROIs, we use our own extraction procedure as described in section III.

CUAVE [11] represents another widely used publicly available dataset. Each of the 36 speakers (17 male, 19 female) utters digits zero through nine in English five times in four different ways: zero to nine, nine to zero with head moving, zero to nine from both profile views, and randomly with head moving. The first three types are pronounced separately, the last are spoken as a phone number, i. e. connected digits. Resolution of the video is 720×480 pixels at 29.97 fps. In this work, we only focus on the first part, i. e. isolated digits with static head pose.

V. EXPERIMENTS

We follow two slightly different speaker independent (SI) evaluation protocols based on k-fold cross validation (KFCV) in order to fairly compare the considered features. In cases of TULAVD and CUAVE, one of the $k-1$ training blocks is separated to serve as validation data for tuning the model hyperparameters. Thus, all of the models ESR detector, feature selection, or HMM, are only trained on the $k-2$ training blocks for each split of each dataset and tuned on the remaining validation block. The score reported in the experiments is the average word accuracy (WAcc) achieved on the corresponding test blocks. The advantage of such procedure is the robustness versus overfitting, since the test set cannot provide any feedback on how well the model will perform on unseen data. Also, as it is almost fully automatic (apart from having to specify the hyperparameter search space), it minimizes the risk that the researcher will favor one model by tuning it more carefully than the others. Due to low number of speakers and in order to preserve compatibility with existing work, in case of

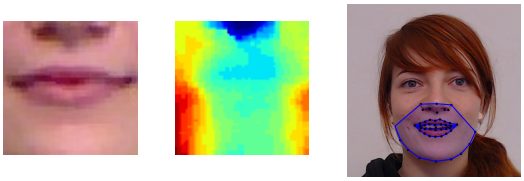

Fig. 2. Example ROI extracted from video and depth data (left and center) and facial landmark configuration for extraction of AAM features (right).

OuluVS we follow the usual speaker-independent leave-one-out cross validation (LOOCV) as in [4], [8], [9]. Overview of data dataset splitting is shown in Tab. I.

We compare our proposed HOGTOP features with five other parametrizations: 2D discrete cosine transform of the ROI (DCT) with energy-based coefficient selection; principal component analysis (PCA) of the ROI (i. e. eigenlips); 46 landmark active appearance model extracted from the lower of the speaker's face (denoted as AAM-r), see fig. 2; spatiotemporal local binary patterns (LBPTOP) [9] extracted densely for each frame of the utterance; and our block-based 3D DCT (DCT3), proposed as a simple baseline for dynamic features. DCT3 similarly to [12] divides the ROI into several overlapping blocks, but extends the approach to incorporate the time axis as well. Each 3D block is approximated by few DCT coefficients, concatenated together and reduced by PCA. As mentioned in section III, all learnable models and their hyperparameters such as energy-based selection of DCT coefficients, PCA dimension, etc., are cross-validated (separately for each dataset) as described earlier in this section.

Table II presents the results achieved on our TULAVD dataset. The experiments were conducted separately on video and depth data, although combination is also possible. The results achieved on OuluVS and CUAVE are reported in table III. There are three scores for each parametrization that differ by post-processing applied: static (i. e. no dynamization), delta (Δ), and dynamization with LDA as mentioned in section III. Note that in the third case, Δ coefficients could also be computed on top of LDA, if doing so was found beneficial by the cross validation. As can be seen, our proposed HOGTOP features outperform the other parametrizations in almost all experiments, often by a large margin. This is mainly due to the efficient exploitation of speech dynamics, which carries essential information for reliable recognition. Contrary to other dynamic parametrizations such as LBPTOP or DCT3, it only focuses on local spatiotemporal changes of the input signal, and thus deals better with variability such as speaker's identity or local contrast variation. However, on CUAVE dataset, HOGTOP performed best only for the static and Δ cases, as there were huge improvements of LBPTOP and DCT3 by LDA dynamization. Reasons for this observation are not clear and will be subjected to further investigation.

State of the art result of 89.7% WAcc for OuluVS dataset was obtained in [4] by a fusion of several features via multidimensional scaling (MDS). However, one disadvantage of such

TABLE II
ACCURACY [%] OF ISOLATED WORD RECOGNITION ON TULAVD.

Param.	video			depth		
	stat.	Δ	LDA	stat.	Δ	LDA
DCT	54.0	68.9	72.5	55.9	66.0	74.4
PCA	51.4	64.4	73.9	55.7	65.3	72.4
AAM-r	58.1	61.8	74.1	59.7	63.0	75.2
LBPTOP	67.4	69.7	74.2	40.9	43.7	64.3
DCT3	61.6	70.8	75.1	62.9	73.0	70.3
HOGTOP	**76.1**	**80.4**	**86.4**	**72.1**	**75.0**	**84.4**

TABLE III
ACCURACY [%] OF PHRASE AND DIGIT RECOGNITION ON OULUVS AND CUAVE, RESPECTIVELY.

Param.	OuluVS			CUAVE		
	stat.	Δ	LDA	stat.	Δ	LDA
DCT	63.0	76.2	79.2	64.7	74.2	81.4
PCA	60.5	73.9	77.9	61.2	69.7	80.1
AAM-r	72.8	76.0	82.1	63.4	64.7	79.0
LBPTOP	62.0	54.2	82.5	69.6	67.6	**91.2**
DCT3	73.3	**82.3**	79.5	71.3	78.6	88.3
HOGTOP	**75.5**	79.6	**85.7**	**80.3**	**81.6**	85.5

TABLE IV
COMPARISON OF OUR WORK TO THE STATE OF THE ART.

OuluVS		CUAVE	
Ref.	Acc [%]	Ref.	Acc [%]
MSHMM	89.9	LBPTOP	91.2
[4]	89.7	HOGTOP	85.5
HOGTOP	85.5	[13]	83.0
[3]	81.3	[12]	77.1

algorithm is the inapplicability to recognition based on sub-word units, e. g. continuous speech recognition. Our system that is focused on good feature design rather than classification algorithm reaches only slightly worse recognition rate, but due to utilization of HMMs it is easily extensible to sub-word units and suitable for incorporation into an audio-based speech decoder. Moreover, with fusion of several parametrizations (PCA, LBPTOP, HOGTOP) via multi-stream synchronous HMM (MSHMM) we obtained accuracy of 89.9 %, a result on par with [4].

Comparison of our work to the state of the art on the CUAVE dataset is rather complicated, as there is no agreed upon evaluation protocol. The closest work in terms of methodology probably are [12], [13]. There were three cases in our experiments that outperformed the 83% word accuracy achieved in [13] with the visemic AAM features, but the results could be influenced by the choice of face alignment algorithm, data split, etc. Table IV summarizes our results and compares them to the results reported in few other works. The MSHMM in the first column denotes the fusion of PCA, LBPTOP and HOGTOP via synchronous multistream HMM.

VI. CONCLUSION

We have presented visual speech features based on spatiotemporal histogram of oriented gradients for automatic lipreading from frontal face videos. The descriptors are extracted from three orthogonal planes, concatenated together and reduced by PCA, thus capturing both texture and dynamic information. We have demonstrated superior performance of our features to other existing parametrizations in experiments on three different datasets. In order to conduct an unbiased comparison, the evaluation protocol was designed to be as independent of the supervisor as possible and most model hyperparameters were cross validated automatically. On both tested publicly available datasets our system achieved state of the art accuracy, showing that provided quality features traditional HMM-based approach can perform on par with sophisticated manifold learning methods. Although not considered in this work, the advantage is then straightforward extensibility to continuous speech recognition and audio-video fusion.

ACKNOWLEDGMENT

This work was supported in part by the Student Grant Scheme (SGS) at Technical University of Liberec.

REFERENCES

[1] J. Chaloupka, J. Nouza, and J. Zdánský, "Audio-visual voice command recognition in noisy conditions," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2008, pp. 25–30.

[2] Z. Zhou, G. Zhao, X. Hong, and M. Pietikinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590 – 605, 2014.

[3] Z. Zhou, G. Zhao, and M. Pietikainen, "Towards a practical lipreading system," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2011, pp. 137–144.

[4] Y. Pei, T. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *IEEE International Conference on Computer Vision, Sydney, Australia, 2013*, 2013, pp. 129–136.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. IEEE Computer Society Conference on*, vol. 1, 2005, pp. 886–893 vol. 1.

[6] A. Savchenko and Y. Khokhlova, "About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems," *Optical Memory and Neural Networks*, vol. 23, no. 1, pp. 34–42, 2014.

[7] S. Pachoud, S. Gong, and A. Cavallaro, "Macro-cuboïd based probabilistic matching for lip-reading digits," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA*, 2008.

[8] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "A new visual speech recognition approach for rgb-d cameras," in *Image Analysis and Recognition*, 2014, pp. 21–28.

[9] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[10] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *in CVPR*, 2012.

[11] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, May 2002, pp. II–2017–II–2020.

[12] P. Lucey and S. Sridharan, "Patch-based representation of visual speech," in *Proceedings of the HCSNet Workshop on Use of Vision in Human-computer Interaction - Volume 56*, Australia, 2006, pp. 79–85.

[13] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans. on Audio, Speech and Language Process.*, vol. 17, no. 3, pp. 423–435, Mar. 2009.