SELECTING RELEVANT VISUAL FEATURES FOR SPEECHREADING

V. Estellers, M. Gurban, J.P. Thiran

Ecole Polytechnique Fédérale de Lausanne, Signal Processing Laboratory 5, Lausanne, Switzerland

ABSTRACT

A quantitative measure of relevance is proposed for the task of constructing visual feature sets which are at the same time relevant and compact. A feature's relevance is given by the amount of information that it contains about the problem, while compactness is achieved by preventing the replication of information between features in the set. To achieve these goals, we use mutual information both for assessing relevance and measuring the redundancy between features. Our application is speechreading, that is, speech recognition performed on the video of the speaker. This is justified by the fact that the performance of audio speech recognition can be improved by augmenting the audio features with visual ones, especially when there is noise in the audio channel. We report significant improvements compared to the most commonly used method of dimensionality reduction for speechreading, linear discriminant analysis.

Index Terms— Feature extraction, image processing, speech recognition.

1. INTRODUCTION

Extracting information from images is difficult, especially when there is high variability in the color, texture and shape of the objects being analyzed. Illumination can also add to this variability. Ideal visual features would capture much of the required information, with little of the variability. Ideally, a quantitative measure of the relevance of the features should be used, a measure adapted to the recognition problem. The relevance of the features reflects their usefulness for the problem that we are trying to solve.

We are analyzing methods of extracting relevant information for speech recognition from the visual modality. Visual speech recognition, or speechreading, can be used to enhance the quality of audio speech recognition, especially in the presence of noise [1]. However, the video has a higher dimensionality and at the same time less relevant information compared to the audio. Two main approaches have been used to reduce the dimensionality of the video, before presenting it to a classifier. The first one has its roots in image compression, and uses image transforms, like the discrete cosine transform (DCT). The second is based on the shape of the mouth, using either contours or active shape models. We adopt the first approach, since it has been shown that the DCT can outperform shape models when the region of interest is properly centered [2].

Our approach to feature extraction is based on using mutual information (MI) as a measure of the relevance of individual features. The MI is computed between each feature and the class labels, which in our case should be the related to speech. We perform an extensive analysis on the type of class labels that should be used, from very small but numerous classes, to a reduced set of large classes.

But since we aim to obtain in the end a compact feature set, we want to avoid situations where features from the set contain the same information, that is, they are redundant. In the end, the relevance of the features will be maximized, while their redundancy is minimized.

The structure of the article is as follows. First, we present the feature selection methods that we will compare, mentioning where similar methods have been applied to audio-visual speech recognition. Then, we present the experimental setup, with our recognition system and the database used. Finally, we present our results and compare them with previous work.

Our contribution is a method of selecting visual features and thus reducing the dimensionality of the visual feature vector for audio-visual speech recognition. The novelty of the presented work consists in the way a redundancy penalty was introduced in the measure used to select features in the particular context of AVSR. Our method is based on maximizing the MI between the features and the class labels, while also minimizing their redundancy with respect to the same class labels. In our best knowledge, this approach was not applied before to the same problem. Although similar methods exist, typically they just maximize MI without penalizing for redundancy.

This article continues and expands our previous work presented in [3]. We expand the study by using larger feature vectors, up to a dimensionality of 192, while our previous tests were performed up to a dimensionality of 50. We also analyze the effect on performance of the type of class labels used, which in our case can be short-time sub-phonetic units, phonemes or even whole words. This analysis is the second contribution brought by our paper to the problem of visual feature extraction for speechreading.

This work is supported by the Swiss National Science Foundation through the IM2 NCCR.

2. FEATURE SELECTION WITH MUTUAL INFORMATION

Feature selection and extraction are important problems in the classification field. A good overview of dimensionality reduction methods in the context of classification can be found in [4]. Our focus here are methods where the quality of features is evaluated using MI. The reason for using MI is that it can find both linear and nonlinear dependencies in the data, contrary to other measures. Another justification comes from Fano's inequality [5], which gives the probability of error p_e when trying to estimate one random variable from another. In our particular case, as we are trying to estimate the class variable C from the features, this equation can be written as:

$$p_e \geq \frac{H(C|F) - 1}{logN} = \frac{H(C) - I(C;F) - 1}{logN} \tag{1}$$

where N is the number of classes, F is the feature set and H is the entropy. The equation gives a lower bound for the probability of error, but does not guarantee that this lower bound will be reached by the classifier. However, "bad" features are guaranteed to lead to a poor classification result, since they would lead to a high lower bound on the error probability.

This shows that a feature set with a high MI with the class labels is desirable. However, computing MI from data is not trivial. The estimation of probability density functions is required, which can not be accurately done in high dimensions. This is why most feature selection algorithms that use MI actually use two or three-dimensional measures, never more. This means that at most two features are used together with the class label to compute the joint probability density.

The simplest approximation is to use the maximum MI between each individual feature and the class labels. If $F = \{Y_1, Y_2 \dots Y_n\}$ is the initial set of features, and $\{\pi_1, \pi_2 \dots \pi_m\}$ is a permutation on a subset of dimension m of the set of feature indices $\{1 \dots n\}$, then the set of selected features can be written as $S = \{Y_{\pi_1}, Y_{\pi_2} \dots Y_{\pi_m}\} \subset F$. To select features by their individual maximum MI with the class means choosing at step $k+1$ the feature [4, 6]:

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} I(Y_i; C) \tag{2}$$

where $S_k = S_{k-1} \cup \{Y_{\pi_k}\}$ is the set of features selected at step k. This is equivalent with assuming that the MI that we want to maximize, $I(S;C)$, can be approximated with the sum of individually computed MI values $I(Y_k;C)$, with $Y_k \in S$.

However, this does not take into account any redundancy that may be present in the features. At the extreme, if two features have identical values and a high MI with the labels, they will both be chosen, even if the second feature does not bring any new information. So, in order to keep the set of relevant features small, redundancy should be penalized.

Redundancy between features can also be expressed in information-theoretic terms. Indeed, the redundancy between features Y_i and Y_j is measured by their MI, $I(Y_i;Y_j)$. However, as the set of selected features grows, we need to compute the redundancy of the candidate feature with the whole set of previously selected features, that is $I(Y_k; S_{k-1})$. This again requires high-dimensional probability density functions. The same approximation as for Equation (2) can be applied, that is, $I(Y_k; S_{k-1})$ is the sum of individual MI values $I(Y_k;Y_i)$ with $Y_i \in S_{k-1}$. An algorithm that does just that is the MIFS algorithm [7]:

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i;C) - \beta \sum_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j}) \right] \tag{3}$$

Here the redundancy is approximated not with the sum, but with a proportion β of the sum, which the authors recommend setting to between 0.5 and 1.

A similar approach is to penalize the average redundancy [8]:

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i;C) - \frac{1}{|S_k|} \sum_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j}) \right] \tag{4}$$

where $|S_k|$ is the size of set S_k. In the end, none of these methods has a good theoretical justification, since the high-dimensional MI values simply can not be approximated with lower-dimensional ones.

A better justification can be given for the information-theoretic methods based on the conditional mutual information, (CMI) as a measure [9], $I(X;C|Y) = I(X,Y;C) - I(Y;C)$. This shows how much the random variable X increases the information we have about C when Y is given. The selection criterion is the following:

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} \left[\min_{Y_{\pi_j} \in S_k} I(Y_i; C | Y_{\pi_j}) \right] \tag{5}$$

The idea is here to find the feature in the already chosen set to which the candidate feature adds the least information, that is, a candidate feature would bring at least that much information to the set. Using the definition for the three-way MI $I(X;Y;C) = I(Y;C) - I(Y;C|X)$ [5] we can rewrite the formula as:

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i;C) - \max_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j}; C) \right] \tag{6}$$

which shows that, in fact, the algorithm can be interpreted as choosing the candidate feature which adds the most over its most-redundant counterpart from the set.

In the end, the goal of all these algorithms is to maximize the joint MI between the S and C, which could be expanded like this (chain rule [5]):

$$\begin{aligned} I(S;C) &= I(Y_{\pi_1}, Y_{\pi_2}, \ldots, Y_{\pi_m}; C) \\ &= \sum_{k=1}^{m} I(Y_{\pi_k}; C | Y_{\pi_1}, \ldots, Y_{\pi_{k-1}}) \qquad (7) \\ &= \sum_{k=1}^{m} \left[I(Y_{\pi_k}; C) - I(Y_{\pi_k}; C; Y_{\pi_1}, \ldots, Y_{\pi_{k-1}}) \right] \\ &= \sum_{k=1}^{m} \left[I(Y_{\pi_k}; C) - I(Y_{\pi_k}; C; S_{k-1}) \right] \end{aligned}$$

An iterative algorithm could maximize the terms of this sum one by one.

$$Y_{\pi_k} = \underset{Y_i \in F \setminus S_k}{\arg\max} \left[I(Y_i; C) - I(Y_i; C; S_{k-1}) \right] \qquad (8)$$

Since Y_{π_k} is the particular Y_i that maximizes the k^{th} term of the sum, all previously mentioned criteria (Eq. 3, 4, 5) can be interpreted as approximations of this general optimization. They all maximize the difference between $I(Y_i; C)$ and an approximation of the redundancy $I(Y_i; C; S_{k-1})$ between Y_i, S_{k-1} and the class labels C. However, nothing can be said about which of these approximation is actually better.

For speechreading, only the simplest method has been previously used. In [10], the authors select the features used for visual speech recognition based on either the MI between features and class labels, or the joint MI between two features and the class label. Neither measure contains a penalty for redundancy. The base visual features used here are discrete cosine transform (DCT) coefficients. Their findings show that the coefficients in the odd columns of the DCT have a much higher relevance, because of the symmetry of the mouth, as confirmed in [11]. The authors consider both phones and subphonetic states as classes of interest, finding that subphonetic classes lead to a significant improvement in performance. This confirms the findings in [12] whose authors use linear discriminant analysis as a means for dimensionality reduction. Their results also show a decrease in performance with coarser classes.

In [13], MI is used to select features from a principal components analysis (PCA) of the mouth region, leading to "mutual information eigenlips". Here too there is no penalty for redundancy, and by contrast with the previous approaches, only coarse word class labels are used.

Our approach differs from the previously mentioned ones in the fact that we include a penalty measure for selecting DCT features for speechreading, which leads to improved results compared to both the LDA, which is the commonly used transform for dimensionality reduction in speechreading, and to the maximum MI approach mentioned above. We also perform an extensive analysis of the influence of class labels on the recognition result, using three types of classes: subphonetic units, phonemes and words.

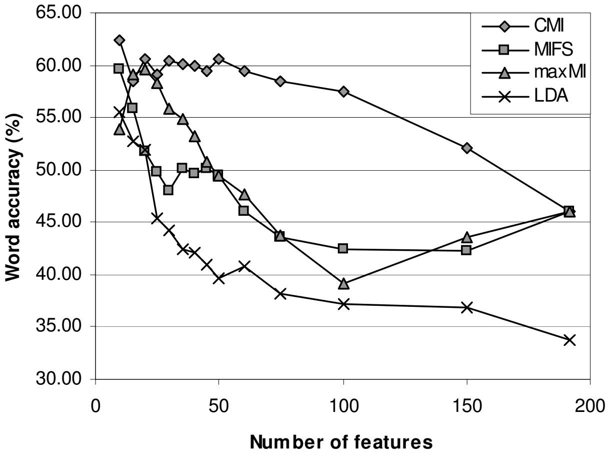

Fig. 1. Recognition results with three MI selection algorithms, compared to the LDA, for dimensionality values ranging from 10 to 192. The classes used here are phonemes.

3. THE EXPERIMENTAL SETUP

We perform speechreading experiments on the CUAVE database [2]. We use the static portion of the "individuals" section of the database, consisting of 36 speakers repeating the digits from "zero" to "nine" five times. Our experiments are speaker independent, using leave-one-out validation, that is 35 speakers are used for training and one for testing. The final reported result is an average of the 36 runs.

Our speechreading system consists of tri-phone hidden Markov models (HMMs) for each phoneme in the database. The phone labeling is obtained by forced alignment using the audio.

The audio features used are mel-frequency cepstral coefficients (MFCCs) with first and second temporal derivatives, and cepstral mean normalization. The visual features are selected using different MI selection algorithms from a pool of DCT coefficients on the region of interest (ROI), which consists of a 128x128 image of the speaker's mouth, normalized for size, centered and rotated. The DCT coefficients are the most important ones taken in a zig-zag order, as in the MPEG/JPEG standard, together with first and second temporal derivatives, and with their means removed. As in [11], the even columns of the DCT are removed.

4. RESULTS AND DISCUSSION

We apply several feature selection algorithms on the DCT visual features, aiming to find the one which is best suited for visual speech. Figure 1 shows our results with MIFS, CMI selection, as well as maximum MI with no penalty for redundancy (maxMI). The equations used are 3, 5 and 2. For MIFS, the β parameter's value was fixed at 0.5, as this lead to the

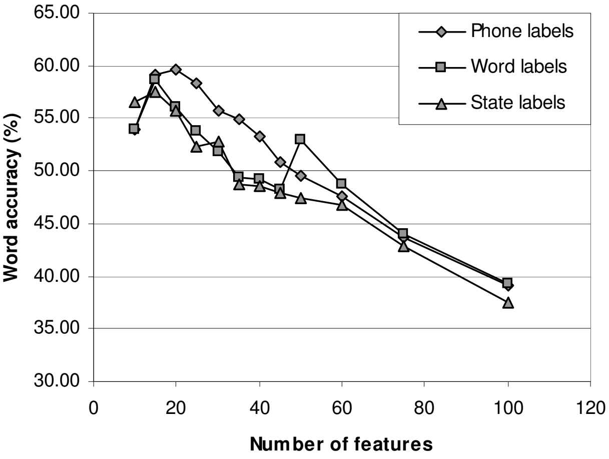

Fig. 2. Recognition results with the maximum MI selection algorithm, using three types of classes.

best results. As can be seen from the figure, the CMI-based algorithm performs best, and this across almost all dimensionality values. The maxMI algorithm also performs well for relatively small feature sets, but its performance decreases for larger ones. We compare our results with the LDA, a common method for visual feature dimensionality reduction, and our results show that there is a clear gain for selection based on MI.

Another area that we investigated was the influence of the type of class labels used in the computation of MI. Figure 2 shows the performance of the maxMI selection algorithm with three types of labels, state, phoneme and word labels. In our case, phoneme labels seem to be the best for the task of AVSR, this probably because the word labels are too coarse, while the state-level classes have a high overlap.

5. CONCLUSION AND FUTURE WORK

We presented a method to reduce the dimensionality of visual feature vectors for speechreading, using MI as both a measure of relevance and of redundancy within the feature set. Our method outperforms the most commonly used method for this task, the LDA. We also show that the phoneme level class labels are better suited for this task, proving that a higher number of classes does not necessarily lead to more discriminative features.

The visual feature selection methods presented here not limited to speechreading, in fact they can be used for any visual classification task, leading to feature sets which are more informative, less redundant and more compact.

6. REFERENCES

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, “Audio-visual automatic speech recognition: an overview,” in *Issues in audio-visual speech processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. 2004, MIT Press.

[2] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, “Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus,” *EURASIP JASP*, vol. 2002(11), pp. 1189–1201.

[3] T. Drugman, M. Gurban, and J.Ph. Thiran, “Relevant feature selection for audio-visual speech recognition,” in *Proceedings of MMSP*, 2007.

[4] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, Kluwer Academic Publishers, 1998.

[5] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, New York, 1991.

[6] R. Reilly and P. Scanlon, “Feature analysis for automatic speechreading,” *Proceedings of the Workshop on Multimedia Signal Processing*, pp. 625–630, 2001.

[7] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5(4), 1994.

[8] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE PAMI*, vol. 27(8), 2005.

[9] F. Fleuret, “Fast binary feature selection with conditional mutual information,” *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[10] P. Scanlon, G. Potamianos, V. Libal, and S.M. Chu, “Mutual information based visual feature selection for lipreading,” *ICSLP*, pp. 2037–2040, 2004.

[11] G. Potamianos and P. Scanlon, “Exploiting lower face symmetry in appearance-based automatic speechreading,” *Proceedings of the International Conference on Audio-Visual Speech Processing*, 2005.

[12] G. Potamianos and H.P. Graf, “Linear discriminant analysis for speechreading,” in *Proceedings of MMSP*, 1998, pp. 221–226.

[13] I. Arsic and J.P. Thiran, “Mutual information eigenlips for audio-visual speech recognition,” *Proceedings of EUSIPCO*, 2006.