Learning Sequential Patterns for Lipreading

Eng-Jon Ong
e.ong@surrey.ac.uk
Richard Bowden
r.bowden@surrey.ac.uk

The Centre for Vision, Speech and Signal Processing,
University of Surrey,
Guildford GU27XH, Surrey, UK

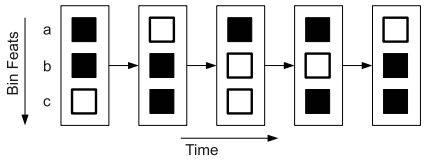
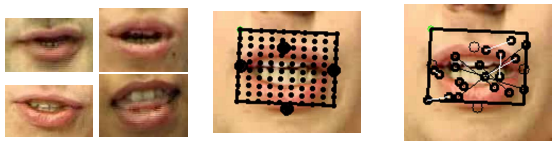


Figure 1: (a)Visualisation of a sequential pattern described. Black squares indicate presence of feature.



(a) (b) (c)

Figure 2: (a) Examples of the mouths for various subjects in the OuluVS database.(b) Visualisation of the grid of points used by for binary comparisons. (c) Visualisation of the 10 binary comparison features.

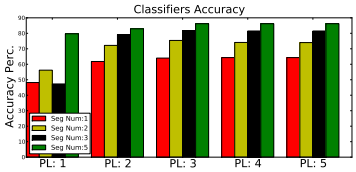


Figure 3: Test data recognition accuracy for classifiers trained with maximum pattern lengths (PL: 1,2,3,5,7) and segment numbers(1,2,3,5).

This paper presents a machine learning approach to Lip Reading and proposes a novel learning technique called sequential pattern boosting that allows us to efficiently search and combine temporal patterns to form strong spatio-temporal classifiers. Attempts at automatic lip reading need to address the demanding challenge that the problem is inherently temporal in nature. It is crucial to model and use spatio-temporal information. To achieve this, we use *sequential patterns*, an ordered sequence of feature subsets. Sequential patterns then form weak classifiers that are combined together into a strong spatio-temporal classifier by means of boosting. A boosted classifier consists of a linear combination of a number (S) of selected weak classifiers, and take the form of: $H(I) = \sum_{i=1}^{S} \alpha_i h_i(I)$. The weak classifiers h_i are selected iteratively based on weights formed during training. In order to determine the optimal weak classifier at each Boosting iteration, the common approach is to exhaustively search the entire set of candidate weak classifiers. However, when dealing with sequential patterns, the number of weak classifiers becomes too large. Specifically, given D features, sequential patterns up to length N and maximum of K items, the number of weak classifiers is the binomial coefficient polynomial: $\binom{D}{K}^N$. In the experiments performed here ($D = 900, K = 3, N = 7$), the total number of weak classifiers is $5x10^{58}$. To address this, we propose a novel Boosting algorithm called Sequential Pattern Boosting (*SP-Boost*). Firstly, we define sequential patterns as follows:

Definition 0.1 *Given a binary feature vector $F = (f_i)_{i=1}^{D}$, let $T \subset \{1,...,D\}$ be a set of integers where $\forall t \in T, f_t = 1$, that is, T represents all the dimensions of F that have the value of 1. We call T as an* itemset. *Let $\mathbf{T} = (T_i)_{i=1}^{|\mathbf{T}|}$ be a sequence of $|\mathbf{T}|$ itemsets. We denote $\mathbf{T}$ as a* sequential pattern. *(Figure 1).*

Attempting to use sequential patterns for classification requires an operator to determine if a sequential pattern is present within a given feature vector sequence:

Definition 0.2 *Let $\mathbf{T}$ and $\mathbf{I}$ be sequential patterns. We say that the sequential pattern $\mathbf{T}$ is present in $\mathbf{I}$ if there exists a sequence $(\beta_i)_{i=1}^{|\mathbf{T}|}$, where $\beta_i < \beta_j$ when $i < j$ and $\forall i = \{1,...,|\mathbf{T}|\}, T_i \subset I_{\beta_i}$. This relationship is denoted with the $\subset_S$ operator, i.e. $\mathbf{T} \subset_S \mathbf{I}$. Conversely, if the sequence $(\beta_i)_{i=1}^{|\mathbf{T}|}$ does not exist, we denote it as $\mathbf{T} \not\subset_S \mathbf{I}$.*

From this, we can then define a sequential pattern weak classifier as follows: Let $\mathbf{T}$ be a given sequential pattern and $\mathbf{I}$ be an itemset sequence derived from some input binary vector sequence F. A sequential pattern weak classifier or *SP weak classifier*, $h^{\mathbf{T}}(\mathbf{I})$, can be constructed as follows:

$$h^{\mathbf{T}}(\mathbf{I}) = \begin{cases} 1, & \text{if } \mathbf{T} \subset_S \mathbf{I} \\ -1, & \text{if } \mathbf{T} \not\subset_S \mathbf{I} \end{cases} \quad (1)$$

We then approach the task of selecting the optimal SP weak classifier in terms of a tree-based search. Here, sequential patterns are arranged into a tree-based structure[1]. Each tree node corresponds to a particular sequential pattern that increase in complexity as we traverse deeper into the tree. Traversing the tree is achieved using two methods: adding a new element to a node's sequential pattern last set; appendding another sequential pattern to its end.

In order to avoid having to search the entire sequential pattern tree (i.e. exhaustive search), a number of tree-pruning strategies are introduced. Importantly, none of these tree-pruning strategies stop us from finding the optimal weak classifier. We start by observing that the more complex a sequential pattern becomes, the number of training examples containing it reduces. This has important implications on how an increase in complexity of a sequential pattern result in monotonically increasing positive errors and monotonically decreasing negative dataset errors. These properties are then used to form pruning criteria that uses the sequential pattern weak classifiers error lower bound and subsequent data-mining based rules. These are then integrated into a queue-based search algorithm for locating the optimal sequential pattern classifiers. Further improvement is achieved by dividing an input sequential pattern into a number of equal sized segments and train separate classifiers for each segments. The final result is obtained by summing the results of these segment classifiers.

Experiments were performed on the OuluVS database [2], containing video sequences of 20 subjects reading 10 phrases 5 times (Figure 2). The aim of the experiments is to recognise which of the 10 phrases was spoken and performance evaluated in a subject-dependent setting using leave-one-out cross-validation. The visual features used take the form of simple comparative binary features, and are similar to LBPs in that relative intensity differences are used. However, our visual features also allow us to capture non-local intensity relations.

The recognition accuracy of the proposed method for different parameter configurations can be seen in Figure 3. We note that as the pattern length and segment numbers are increased, the recognition accuracy significantly increased as well. The recognition rate was 64.3% with pattern length 1 and segment number 1, increasing to a rate of 86.2% when the pattern length was 7 with 5 segments. The highest average test recognition rate was 86.2% compared to 70.2 in [2].

[1] J. Ayres, J. Gehrke, T. Yiu, and J. Flannick. Sequential pattern mining using bitmaps. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2002.

[2] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7), 2009.