

# Early Recognition of Sequential Patterns by Classifier Combination

Seiichi Uchida and Kazuma Amamoto  
Kyushu University, Fukuoka, Japan  
{uchida,amamoto}@human.is.kyushu-u.ac.jp

## Abstract

*This paper proposes an early recognition method, i.e., a method for recognizing sequential patterns at their beginning parts. The method is based on a combination of frame classifiers prepared at individual frames. The training patterns misrecognized by the frame classifier at a certain frame are heavily weighted for the complementary training of the frame classifier at the next frame. The method was applied to an online character recognition task for showing its usefulness.*

## 1. Introduction

This paper proposes a method of early recognition, where the recognition result of a sequential pattern will be determined at its beginning part. If the proposed method is applied to an online character recognition task of discriminating between “1” and “3”, an input pattern “1” might be recognized at the end of its “P”-shaped part. If the proposed method is applied to gesture recognition, a gesture pattern “raise hand” might be recognized when a hand reaches shoulder height (i.e., before the hand reaches its top position).

Several approaches will realize early recognition. One approach will be partial matching, where the recognition result of an input pattern is determined by the matching cost of the beginning part. This simple approach has been employed in gesture recognition [1] and speech recognition [2].

Another approach is combination of classifiers  $h_1, \dots, h_t, \dots, h_T$ , where  $h_t$  is a *frame classifier* prepared at the  $t$ th frame (i.e., time  $t$ ). The frame classifier  $h_t$  provides a recognition result by only using the feature vector of the  $t$ th frame (or by using  $t$  feature vectors of the first  $t$  frames). The recognition result at the  $t$ th frame will be determined by combining  $t$  recognition results provided by  $h_1, \dots, h_t$ . Clearly, if we prepare the frame classifiers  $h_1, \dots, h_T$ , it is possible to determine its recognition result at an arbitrary frame  $t \leq T$ ,

that is, it is possible to realize early recognition.

This paper proposes an early recognition method belonging to the second approach. The simplest early recognition method of this approach can be realized by “frame independent training”, where each frame classifier  $h_t$  is trained independently of the past classifiers  $h_1, \dots, h_{t-1}$ . This simple method, however, does not fully utilize the potential of the classifier combination approach (as discussed later).

The proposed early recognition method is not frame-independent; the frame classifier  $h_t$  will be trained by considering the training results of the past classifiers  $h_1, \dots, h_{t-1}$ . Specifically, the training patterns misrecognized by the past frame classifiers are heavily weighted at the training of  $h_t$ . If a weighted pattern (i.e., a “hard” pattern) is misrecognized again by  $h_t$ , its weight is further increased; otherwise, its weight is decreased. This *weight propagation* is effective to prepare complementary frame classifiers for earlier recognition.

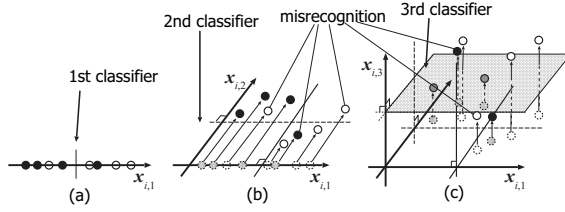
## 2. Classifier Training for Early Recognition

This section discusses the proposed early recognition method. Readers who know AdaBoost [3] will notice that the proposed method for training and utilizing frame classifiers is strongly related to AdaBoost. This relation will be discussed in 2.4.

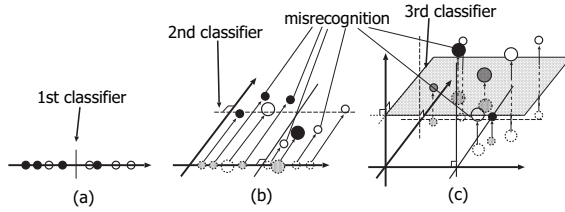
### 2.1. Frame classifier

Consider a set of  $N$  training patterns  $\{\mathbf{x}_i | i = 1, \dots, N\}$  from two categories. Each training pattern has its category label  $y_i \in \{-1, 1\}$  and is a sequence of  $d$ -dimensional vectors,  $\mathbf{x}_i = \mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,T}$ , where  $t$  denotes the frame number. It is assumed that all the training patterns have the same length  $T$ .

As noted in Section 1, early recognition is realized by combining frame classifiers  $h_t(\mathbf{x}) \in \{-1, 1\}$ . The  $t$ th frame classifier is a two-category classifier prepared at the  $t$ th frame for providing the recognition result at  $t$ . The frame classifier  $h_t(\mathbf{x})$  is assumed to be trained



**Figure 1. Frame classifiers at (a)  $t = 1$ , (b)  $t = 2$ , and (c)  $t = 3$ .**



**Figure 2. Frame classifiers with weight propagation.**

by only using the feature vectors of the  $t$ th frame, i.e.,  $\{\mathbf{x}_{i,t} | i = 1, \dots, N\}$ . (An extension of using the past feature vectors  $\{\mathbf{x}_{i,\tau} | i = 1, \dots, N, \tau = 1, \dots, t\}$  is also examined in Section 3.3.)

Figure 1 shows a feature space for representing the frame classifiers of the first three frames for  $d = 1$ . Its  $t$ th axis represents the feature vector of the  $t$ th frame. Consequently,  $h_t(\mathbf{x})$  only concerns with the  $t$ th frame and therefore its discrimination boundary is formally depicted as a hyperplane perpendicular to the  $t$ th axis in Fig. 1.

## 2.2. Early recognition using frame classifiers

The early recognition result at an arbitrary frame  $t$  is obtained from a weighted combination of the frame classifiers  $h_1(\mathbf{x}), \dots, h_t(\mathbf{x})$ , i.e.,

$$H_t(\mathbf{x}) = \text{sign} \left( \sum_{\tau=1}^t \alpha_{\tau} h_{\tau}(\mathbf{x}) \right), \quad (1)$$

where the weight  $\alpha_t$  is an importance parameter of  $h_t(\mathbf{x})$ . One possible definition of  $\alpha_t$  is

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right), \quad (2)$$

where  $\varepsilon_t$  is a (weighted) error rate of  $h_t(\mathbf{x})$  whose detail will be given later. If the combined classifier  $H_t(\mathbf{x}) > 0$ ,  $x$  is classified into 1 (−1, otherwise).

## 2.3. Weight propagation

The simplest procedure for training the frame classifiers is the “frame independent training” where each

frame classifier  $h_t(\mathbf{x})$  is trained independently of the past classifiers  $h_1(\mathbf{x}), \dots, h_{t-1}(\mathbf{x})$ . This procedure, unfortunately, does not fully utilize the merit of the classifier combination (1). Let us consider an extreme case where every sequential pattern is a time-invariant pattern, i.e.,  $\mathbf{x}_{i,1} = \dots = \mathbf{x}_{i,t} = \dots = \mathbf{x}_{i,T}$ . In this case, all the frame classifiers become identical and therefore the combined classifiers are reduced to a single classifier.

Another procedure which can utilize the combination (1) is a complementary training of frame classifiers by *weight propagation*. When training the frame classifiers from  $t = 1$  to  $T$ , the patterns misrecognized by the last classifier  $h_{t-1}(\mathbf{x})$  are largely weighted not to be misrecognized by  $h_t(\mathbf{x})$ . Let  $D_t(i)$  denote the weight of the  $i$ th training pattern at  $t$ . Then  $h_t(\mathbf{x})$  is trained to minimize

$$\varepsilon_t = \Pr_{D_t(i)} [h_t(\mathbf{x}_{i,t}) \neq y_i] = \sum_{i: h_t(\mathbf{x}_{i,t}) \neq y_i} D_t(i). \quad (3)$$

Since a training pattern misrecognized by  $h_{t-1}(\mathbf{x})$  has a large weight  $D_t(i)$ , it might be correctly recognized during the minimization of  $\varepsilon_t$ . Figure 2 illustrates this procedure.

The entire training procedure is as follows:

1. Set  $t = 1$  and  $D_t(i) = 1/N$ .
2. Repeat the following steps for  $t = 1$  to  $T$ .
3. Obtain  $h_t(\mathbf{x})$  which minimizes (3).
4. Update the weight as

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(\mathbf{x}_{i,t}))}{Z_t}, \quad (4)$$

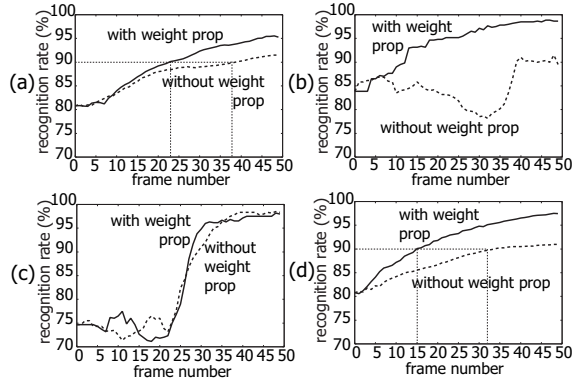
where  $Z_t$  is a constant to satisfy  $\sum D_{t+1}(i) = 1$ .

Note that we can stop the above repetition at  $t = \bar{t} (< T)$  if any stopping condition is satisfied. Then,  $\bar{t}$  becomes the number of frame classifiers for early recognition. In this paper, the condition will not be discussed. Future work will focus on this condition while referring to the Fu [4] and its application called WaldBoost [5].

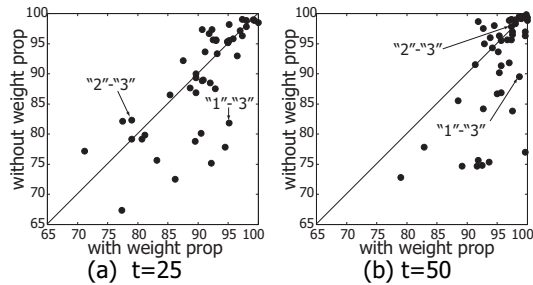
## 2.4. Relation to AdaBoost

The proposed method is closely related to AdaBoost [3]; the frame classifier  $h_t(\mathbf{x})$  corresponds to a so-called weak learner and the combined classifier  $H_t(\mathbf{x})$  corresponds to a strong learner. The weight updating scheme (4) seems the same as that of AdaBoost.

The proposed method, however, is not the same as AdaBoost. In AdaBoost, all features are always available when training each weak learner. This condition holds even in AdaBoost-based feature selection methods such as [6]. In contrast, the frame classifier  $h_t(\mathbf{x})$  cannot use any feature of the  $t' (> t)$  th frame in the proposed method.



**Figure 3. Recognition rate at each frame.** (a) Average of all the 45 category pairs. (b) Category pair “1”-“3.” (c) Category pair “2”-“3.” (d) Average for multiple-frame classifier.



**Figure 4. Recognition rates of all the 45 category pairs.**

### 3. Application to Online Character Recognition

#### 3.1. Experimental setup

The proposed early recognition method was applied to online character recognition, which is an appropriate evaluation task because of its intelligibility and visibility. Isolated digit patterns were obtained from the public database called Ethem Alpaydin Digit, which includes 500 training patterns and 300 test patterns for each of 10 categories (“0”- “9”). Performance was evaluated on all the 45 ( $=_{10}C_2$ ) category pairs. After preprocessing, each pattern was represented as a sequence of two-dimensional feature vectors (i.e., pen-position feature vectors) with the equal length  $T = 50$ .

A simple linear classifier was used as the frame classifier  $h_t(x)$  which minimizes (3). The discrimination boundary of  $h_t(x)$  was the equidistant line from  $\tilde{x}_t^{-1}$  and  $\tilde{x}_t^y$ , where  $\tilde{x}_t^y$  is the weighted center of the category

$$y \in \{-1, 1\}, \text{ that is, } \tilde{x}_t^y = \sum_{i:y_i=y} x_{i,t} D_t(i).$$

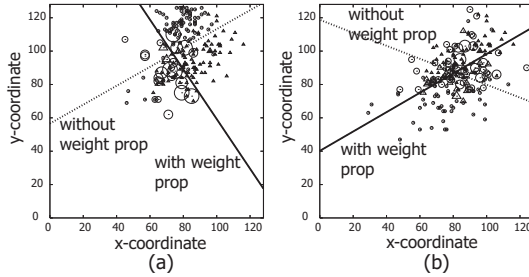
#### 3.2. Results

Figure 3 (a) shows the average recognition rate by the proposed method (“with weight propagation”) and that by the frame independent training (“without weight propagation”) at each frame  $t$ . This figure indicates that the weight propagation was effective for early recognition. For example, the proposed method attained 90% at  $t = 23$  whereas the frame independent training attained it at  $t = 37$ .

Figures 4 (a) and (b) show the recognition rates for all the 45 category pairs at  $t = 25$  and 50, respectively. In these figures, a point around the lower and/or right side represents a category-pair whose recognition rate was improved by the weight propagation. For example, the pair “1”-“3” is an improved pair. Accordingly, these figures also show that the weight propagation was effective for most category pairs. In contrast, several category pairs, such as “2”-“3”, cannot have any improvement.

Figure 3 (b) shows the recognition rate of the category pair “1”-“3”, which is one of pairs with remarkable improvement by the weight propagation. The reason of this improvement will be explained by Fig. 5 (a), which illustrates the discrimination boundary of  $h_{17}(x)$  (depicted as the solid line) on  $x - y$  coordinate feature space. From this figure, it can be observed that there are several patterns with large weights and the frame classifier  $h_{17}(x)$  discriminates them correctly as possible. Those patterns are misrecognized at the 16th (or more past) frame and therefore this figure indicates that  $h_{17}(x)$  works complementary to past frame classifiers for better recognition accuracy. The dashed line in Fig. 5 (a) depicts the boundary by the frame classifier trained without weight propagation. Since this boundary is determined according to the feature distribution of this frame (i.e., regardless of the results of past frame classifiers), it is impossible to expect complementary effect.

Figure 3 (c) shows the recognition rate of the category pair “2”-“3”, which achieved no improvement by the weight propagation. This graph can be separated into two parts: the first half with low recognition rate and the latter half with high recognition rate. Figure 5 (b) illustrates the discrimination boundary at the 17th frame, i.e., a frame in the first half. This figure indicates that patterns with large weights are still overlapped and hard to discriminate them. In fact, the first half of “2” is very similar to that of “3” and their difference has no specific tendency. Consequently, hard patterns do not form some clusters but show a random distribution in Fig. 5 (b), and therefore no improvement



**Figure 5. Discrimination boundary and pattern distribution at  $t = 17$ . (a) Category pair “1”-“3.” (b) Category pair “2”-“3.”**

has been achieved even with the weight propagation. In contrast, their latter halves of “2” and “3” have a distinct difference and therefore the frame classifier can discriminate them without weight propagation.

### 3.3. Multiple-frame classifier

In the above discussion, the frame classifier  $h_t(\mathbf{x})$  is concerned only with the feature vector of the  $t$ th frame  $\{\mathbf{x}_{i,t} | i = 1, \dots, N\}$ . This is a very hard situation for the frame classifier. In fact, in the task of on-line character recognition, this implies that each frame classifier  $h_t(\mathbf{x})$  distinguishes character patterns only using the single pen position at a certain instant  $t$ .

This situation can be relaxed by using the past feature vectors  $\{\mathbf{x}_{i,\tau} | i = 1, \dots, N, \tau = 1, \dots, t\}$  for training the frame classifier  $h_t(\mathbf{x})$ . That is, each frame classifier  $h_t(\mathbf{x})$  itself can distinguish character patterns by using the pen movements (i.e., partial strokes) during  $\tau = 1, \dots, t$ .

Figure 3 (d) shows the average recognition rate when this multiple-frame classifier was used. Comparison with Fig. 3 (a) indicates that the performance was improved by the multiple-frame classifier. The multiple-frame classifier could improve recognition accuracy without the weight propagation ( $t = 37 \rightarrow 32$  for 90% accuracy). The improvement, however, was larger with the weight propagation ( $t = 23 \rightarrow 15$  for 90% accuracy). This is because the multiple-frame classifier can form its discrimination boundary in a higher-dimensional ( $d \times t$ -dimensional) feature space and thus could utilize the merit of the weight propagation.

## 4. Conclusion and Future Work

An early recognition method was proposed for recognizing sequential patterns at their beginning parts. The method trains frame classifiers prepared at individual frames and provides a final recognition result by

combining the recognition results of the frame classifiers. The key idea for improving early recognition performance is weight propagation that patterns misrecognized by the frame classifier at a certain frame are heavily weighted at the training of the frame classifier at the next frame. The usefulness of weight propagation was confirmed through an online character recognition experiment.

Future work will focus on the application to various recognition problems. Although online character recognition was chosen as the evaluation task for its intelligibility and visibility, the automatic learning ability of the proposed method may be appreciated in more complicated and less visible tasks. As noted in Section 2.3, future work will also focus on the condition for accepting the early recognition result at the  $t$ th frame. Wald’s sequential probability ratio test [4, 5] may give a condition. Incorporation of some nonlinear time-warping scheme will be necessary for dealing with sequential patterns having different lengths. The number of frame classifiers per frame can be increased for improving the performance of early recognition.

**Acknowledgment** This work was supported in part by the Research Grant (No.19650042) of The Ministry of Education, Culture, Sports, Science and Technology in Japan and Microsoft Research Asia Mobile Computing in Education Theme Program.

## References

- [1] A. Mori, S. Uchida, R. Kurazume, R. Taniguchi, T. Hasegawa and H. Sakoe, “Early recognition and prediction of gestures,” Proc. ICPR, vol. 3 of 4, pp. 560–563, 2006.
- [2] M. Mohri, F. C. N. Pereira, and M. Riley, “Weighted finite-state transducers in speech recognition,” Computer Speech and Language, vol. 16, no. 1, pp. 69–88, 2002.
- [3] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” J. Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
- [4] K. S. Fu, Sequential methods in pattern recognition and machine learning, Academic Press, 1969.
- [5] J. Sochman and J. Matas, “WaldBoost — Learning for time constrained sequential detection,” Proc. CVPR, vol. 2, pp. 150–156, 2005.
- [6] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” Proc. CVPR, vol. 1, pp.511–518, 2001.