PATTERN RECOGNITION FOR ORIENTAL MUSIC SCORES A COMBINATION OF KNOWLEDGE ANALYSIS, KALMAN FILTERING AND NEURAL NETWORKS

T. FELFEL[1], J.-M. CHOUVEL[2], J.-P. RICHARD[1]
[1] LAIL, CNRS, UPRESA 8021
Ecole Centrale de Lille
BP 48, 59651 Villeneuve d'Ascq Cedex, FRANCE
[2] Université de Reims, Champagnes Ardennes
departement de musicologie, campus Croix-Rouge
57, rue Pierre Taittinger 1096 Reims, France…

Abstract: This paper presents a new approach of optical score recognition. The method is based on the analysis of human scoring procedure and on the combination of Kalman filter and neural network techniques.
Keywords: image recognition, music score recognition, image segmentation, Kalman-Bucy filter, RBF neural networks, oriental music. *IMACS/IEEE CSCC'99 Proceedings*, Pages:2891-2896

1 Introduction

This work concerns the recognition of music score sheets (either printed or hand-written) and its application to the oriental music. Note that the proposed method is also efficient for occidental music even if, up to now, only monophonic scores have been considered in the implementation.

The oriental music is characterized by the use of sounds which have neither fixed heights nor intervals (since the mode can be changed within the course of a same piece of music), contrarily to the occidental music. Oriental music is transcribed in the same way than the western one, except that in order to transcribe some degrees, lowered or raised of a subdivision of tone, signs specific to each action have to be used (for instance, half-flat or half-sharp for a quarter of tone). The purpose of our work is to design a software allowing some occidental musicologists having access to this music, listening and analyzing them in an automated way (this means, without having to learn the skill of their performance nor to hand-seize them by mean of an edition software). Furthermore, since this music is often edited in a handwritten way, this work will also lead to a better edition of these documents.

Our recognition system relies on an original process of image analysis, based on the association of three techniques:

- the segmentation,
- the recognition by neural networks,
- and the analysis by use of context.

The hierarchy of use of these techniques results from the assimilation of the way human proceeds to music scoring. Two levels of extraction are to be distinguished:

- The first concerns the separation of the bottom layer and of the music scoring layer: the bottom layer is exclusively constituted by the music stave, as in previous such works [1,2,3].
- The music scoring layer constitutes the second level: the scores (notes, groups of notes), comments (title, …) and signs (flats, sharps, bars, "crescendi"…) that are added on the bottom layer. It can be separated (as in [4]) into two types of entities: symbols and constructs. The ***symbols*** are entities one cannot disconnect (as flat, sharp, numbers or keys); the ***constructs*** are connected entities formed by a group of symbols or of basic entities (as segments or heads of crotchets).

The efficiency of the music document recognition is proportional to the capacity of separating the different layers, of recognizing symbols, of segmenting constructs, and finally of reconstructing by use of the high level knowledge of music notation. Thus, we suggest the following architecture to build our application:

1/ recognition and removal of the bottom layer;
2/ Recognition of the constructs from basic entities;
3/ recognition and classification of the remained entities (symbols);
4/ reconstruction of the musical score.

These four items correspond to the plan of the present paper.

2 Recognition and removal of the bottom layer

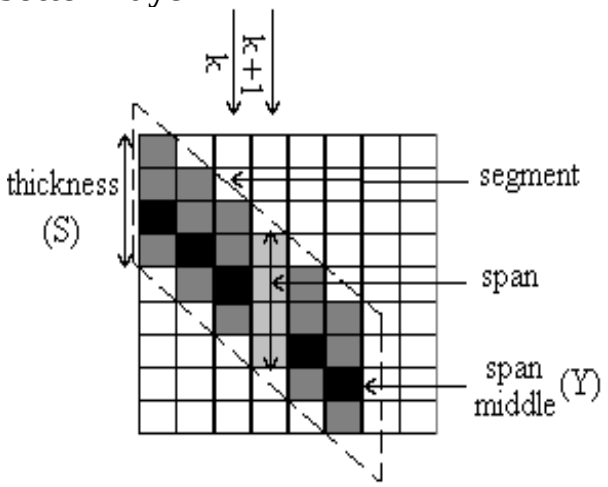


Fig. 1 - Image of a segment

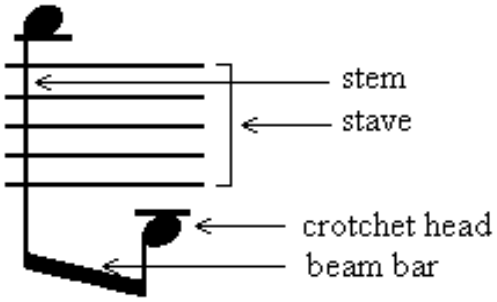


Fig. 2 - Image of group of notes

The action consists in extracting the stave from the image [1,2,3,5]. Long and parallel segments form the stave. So to extract the stave, our computation starts with the recognition of all the segments in the sheet: to remove them we use their characteristics. For this action, we use Kalman-Bucy filtering [6]. Such filter is known to perform a good segment detection and is greatly robust in presence of noise [1,2].

Our Kalman-Bucy filter notations are being introduced. The filter acts by following, step by step (indexed by *k*), a state x_k and a measurement z_k. These variables are related by:

$$z_k = H_k x_k + w_k$$

w_k is the noise on state: white noise of null average and of known variance R_k. H_k is the measurement matrix supposed known. The evolution of the state is:

$$x_{k+1} = F_k x_k + u_k + v_k$$

v_k the noise on measurement: white noise of null average and of known variance Q_k. Matrix F_k is supposed known.

The best prediction of the noise v_k is null, so the best prediction $\hat{x}_{k+1/k}$ of the state x_{k+1} is:

$$\hat{x}_{k+1/k} = F_k x_k + u_k$$

The variance of the error prediction at the step *k* is equal to the noise variance:

$$P_{k/k} = E\left[(x_{k+1} - \hat{x}_{k+1/k})(x_{k+1} - \hat{x}_{k+1/k})^T \right]$$
$$= E\left[v_k v_k^T \right] = Q_k .$$

From an estimation $\hat{x}_{k/k}$, at step *k*, but without measurement, the best prediction for state at step *k*+1 is:

$$\hat{x}_{k+1/k} = F_k \hat{x}_{k/k} + u_k .$$

Then, the variance of prediction error $P_{k+1/k}$ is:

$$P_{k+1/k} = E\left[(x_{k+1} - \hat{x}_{k+1/k}) \cdot (x_{k+1} - \hat{x}_{k+1/k})^T \right]$$
$$= F_k P_{k/k} F_k^T + Q_k .$$

This variance depends on the accuracy of the estimation before $P_{k/k}$ and on the noise magnitude Q_k. Without noise on the measurement $(w_k = 0)$, the optimal prediction is:

$$\hat{z}_{k+1/k} = H_{k+1} \hat{x}_{k+1/k} .$$

After the measurement of z_{k+1} at the step $k+1$, the gap between the measurement z_{k+1} and prediction $\hat{z}_{k+1}$ provides an indication on the estimation error. This error must intervene in the next predictions. In this way, we have to bring a correction to the predicted state $\hat{x}_{k+1/k+1}$, which must be proportional to the gap $z_{k+1} - \hat{z}_{k+1/k}$:

$$\hat{x}_{k+1/k+1} = \hat{x}_{k+1/k} + G_{k+1}(z_{k+1} - \hat{z}_{k+1/k})$$

,

G_{k+1} depending on the noise variances.

Therefore, we have the following Kalman-Bucy filter equations:

$$\hat{x}_{k+1/k+1} = \hat{x}_{k+1/k} + G_{k+1}\left(z_{k+1} - H_{k+1}\hat{x}_{k+1/k}\right)$$

$$P_{k+1/k+1} = P_{k+1/k} - G_{k+1}H_{k+1}P_{k+1/k}$$

$$G_{k+1} = P_{k+1/k}H_{k+1}^T\left(H_{k+1}P_{k+1/k}H_{k+1}^T + R_{k+1}\right)^{-1}$$

Now, this filter is being applied to segment extraction.

Our application sweeps the first column of the image from top to bottom, then goes to the next columns (left to right): when it finds a black pixel, it makes the hypothesis of a possible starting segment and begins its recognition. The action aborts when the information extracted in the final iteration does not agree with hypothesis emitted in the last one.
The variables used by the filter are now to be defined (see Fig. 1). The integer *k* denotes column indices. *S* is the thickness of the span in the segment, *Y* the position of the middle pixel of the span and $\dot{Y}$ its derivative (speed) defined as $\dot{Y}_k = Y_k - Y_{k-1}$. In fact, the tracking of the segment is made by identifying the evolution $(Y, \dot{Y})$ of the trajectory and thickness *S* of the span. v_k represents the disturbances on the state, this means, due to some mismatch in the writing or scanning of the score. Further, w_k will denote the measurement perturbations (supposed to be a white noise). This leads to the following state model:

$$x_{k+1} = F_k x_k + v_k \text{ with}$$

$$F_k = F = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\quad x_k = \begin{bmatrix} Y_k \\ \dot{Y}_k \\ S_k \end{bmatrix},\quad v_k \begin{bmatrix} v_{1k} \\ v_{2k} \\ v_{3k} \end{bmatrix}.$$

The filter can be separated into two disconnected subsystems corresponding to the position vector $(Y_k, \dot{Y}_k)^T$ and the span thickness *S*. This decomposition decreases the complexity of calculation without tampering the results.

- *First subsystem: position vector*

$\begin{bmatrix} Y_k \\ \dot{Y}_k \end{bmatrix}$, with matrix $F_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and the state noise $V = \begin{bmatrix} v_{1k} \\ v_{2k} \end{bmatrix}$. For this noise, a second-order development has been proceeded, considering the acceleration $\ddot{Y}$ of *Y*:

$$\begin{cases} V_1 = \frac{1}{2}\ddot{Y}_k \\ V_2 = \ddot{Y}_k \end{cases} \text{ variance } Q_{Y,k} = V_{\ddot{Y}}\begin{bmatrix} 1/4 & 1/2 \\ 1/2 & 1 \end{bmatrix} \text{ with } V_{\ddot{Y}} \text{ t}$$

he acceleration variance, generally constant.

Second subsystem: span thickness

S, with matrix F_2=1 and the state noise V_3. The noise variance $Q_{E,k}$ is normally constant, too.

The following measurement systems arise, with W_i (i=1,2) the measurement noises:

- *Position measurement:* $z_{1,k} = Y_k + W_1$ with $H_{1,k} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and the variance R_Y fixed.

- *Thickness measurement:* $z_{2,k} = S_k + W_2$ with $H_{2,k} = 1$ and the variance R_S fixed.

Here we shall take $W_1 = W_2$, a same white noise of null average and of known variance R_k. H_k is the matrix binding the state to the measurement, it is supposed to be known. Prediction needs to compute the matrices of error covariance $P_{\hat{Y},k/k}$ (for the position) and $P_{\hat{S},k+1/k}$ (for the span thickness). If σ_Y denotes the root mean-square (deviation) between state and prediction, then

$$P_{\hat{Y},k/k} = \begin{bmatrix} \sigma_Y^2 & \sigma_Y\sigma_{\dot{Y}} \\ \sigma_Y\sigma_{\dot{Y}} & \sigma_{\dot{Y}}^2 \end{bmatrix},$$

$$P_{\hat{Y},k+1/k} = \begin{bmatrix} \left(\sigma_Y^2 + 2\sigma_Y\sigma_{\dot{Y}} + \sigma_{\dot{Y}}^2\right) & \left(\sigma_Y\sigma_{\dot{Y}} + \sigma_{\dot{Y}}^2\right) \\ \left(\sigma_Y\sigma_{\dot{Y}} + \sigma_{\dot{Y}}^2\right) & \left(\sigma_{\dot{Y}}^2\right) \end{bmatrix} + V_{\ddot{Y}}\begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

$$P_{\hat{S},k+1/k} = P_{\hat{S},k/k} + Q_S.$$

The coefficient $G_{.,k+1}$ balancing the measurement and prediction is :

$$G_{Y,k+1} = \frac{1}{\sigma_Y^2 + 2\sigma_Y\sigma_{\dot{Y}} + \sigma_{\dot{Y}}^2 + Q_Y} \begin{bmatrix} \sigma_Y^2 + 2\sigma_Y\sigma_{\dot{Y}} + \sigma_{\dot{Y}}^2 + \frac{1}{4}V_{\ddot{Y}} \\ \sigma_Y\sigma_{\dot{Y}} + \sigma_{\dot{Y}}^2 + \frac{1}{2}V_{\ddot{Y}} \end{bmatrix}$$

$$G_{S,k+1} = \frac{P_{\hat{s},k/k} + Q_s}{P_{\hat{s},k/k} + Q_s + R_s}$$

The initial substates and the filtering rules are now to be fixed. For all new span, after measurement, the following hypotheses (H) and criterion (C) are chosen:

(H) Let Y_0 the span middle; $\dot{Y}_0 = 0$ for the tracking of horizontal tendency segments, $P_{Y,0}$ and $P_{S,0}$ are set to the thickness S_0 of the span, in order to have a large adaptability for the slope determination. The variance of noise measurement R is set to 2 pixels, corresponding to the sampling error.

(C) For the filtering stages, the correspondence between the observed state and the predicted one is made by application of the normal law (for a probability 0.99), the gap between observed and predicted must be inferior to 2.8σ (σ_Y and σ_S are given by the $P_{\hat{\cdot},k/k}$). In the contrary case, the observation is left but the hypothesis follows up until meeting a correspondence with (H), or stops after a chosen number n of iterations (for a 300 dpi image, we took n=15 for the stave, 5 for the group bars, 3 or 4 for the vertical segments).

The filtering process begins by sweeping the image, as said before: each time a black pixel is encountered, it is considered as a belonging to the first span of a possible segment. Using initial hypothesis (H), the purchase of the segment starts. The action stops when the criterion (C) fails. Each time a span is verifying hypothesis, its pixels are eliminated from the image. The filtering stops when all pixels of the image are treated.

Because of the connection of the group bars in our case of scores, we make the filtering twice: to detect, firstly, the stave, secondly, the group bars. In addition, before the second filtering, we make an image thinning so to disconnect the group bars.

At the end of filtering, all kinds of segments in the image are recovered. To have the bottom layer, only longest horizontal segments are extracted. The remaining vertical segments are stems and measure bars. Those with horizontal tendency correspond to group-of-notes bars and others.

3 Recognition of the constructs from basic entities

After the first stage, an intermediate one is dedicated to detection and extraction of possible notes heads (crotchets heads) by erosion of the image. A music knowledge analysis then compares the expected crotchets heads positions with the detected segments so to identify the constructs of the musical scoring [5,7,8,9,10]. For instance (see Fig. 2), two crotchet heads + two vertical stem bars + one beam bar = a beamed group, which validates the detection. The remaining entities will be treated by the third stage.

4 Recognition and classification of the remaining symbols

The remaining basic entities (symbols) are treated as independent characters by use of the classical OCR technique (Optical Character Recognition, [11]). For the classification, a RBF neural network is chosen (Radial Basis Function, [12,13]). This classifier is well adapted to our system architecture: it has a good classification reliability even not well extracted entities like those deteriorated by the extraction of the bottom layer. Besides, the answer of the classifier is not one class like other neural classifiers, but it returns for every class a probability to match the symbol in entry, so we can choose the most probable class in agreement with the high level knowledge.

The neural network RBF is characterized by a transition function which output is not a Boolean {0,1} (as in [14,15]), but a density of probability (between 0 and 1). Most of the time [9], Gaussian functions $\Psi(x) = e^{-\|x-C\|^2 \rho^{-2}}$ are used. Such functions correspond to the distance from the vector x (the symbol to be classified) to the model C (the class) and is parameterized by a coefficient ρ representing the field depth.

The number of neurons in the classifier is closely linked to the number of learning models. If these last are well chosen, the two numbers are equal. The size of the model image is normalized, in the present application, to a 24×24 pixels matrix. This size permits an optimal classification [13].

5 Reconstruction of the music score

The reconstruction uses the context and the rules of musical notation, which constitutes the high-level knowledge [5,10,16]. This stage permits to synthesize and analyze the results of the former ones

in order to reconstruct the original score. The analysis is done measure by measure, for each measure type it uses a specific analyzer, according to the diagram of Fig. 3 that has been worked out from musicologist knowledge.

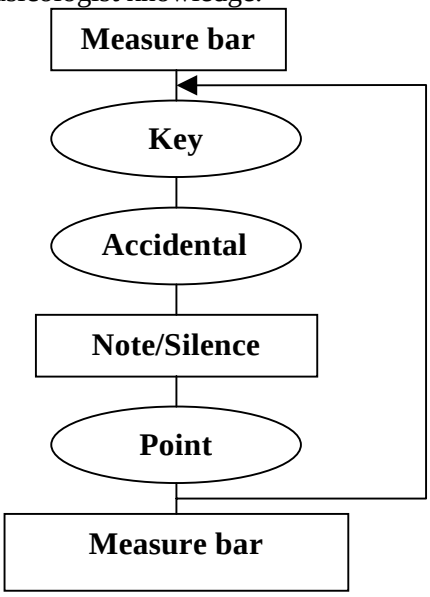


fig3- Diagram of a simple measure analyzer

In case this analysis diagram does not provide a coherent reconstruction of the score, there is to go back to lower levels in order to determinate whether an error occurred in the pattern recognition. This error checking takes advantage of the RBF neural networks chosen at the third step of recognition, since alternative symbols were provided at this step (with a lower probability): these other proposals can be automatically checked to achieve the coherence. The software user (musician or musicologist) will intervene only in last recourse, in case this analysis process fails.

6 An illustrative example

The following section illustrates the hierarchy of the segmentation and treatment of the image. The example is worked on a sample of an oriental score with a rather rough scanning accuracy.

Original image:

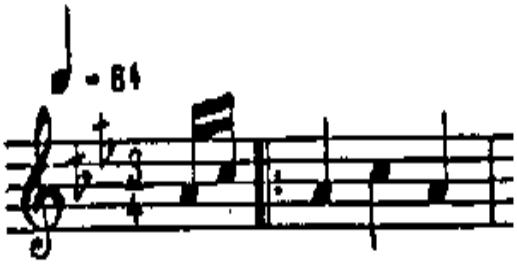


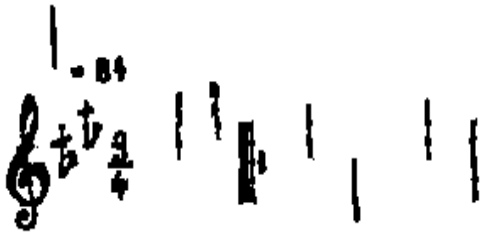
Image after extraction of the bottom layer (by Kalman-Bucy filtering)

erosion: after erosion the detection of the head crotchets is made by criterion of size and shape

Three successive images of a one-shot action: after detection of segments with horizontal tendency and not belonging to the stave, of vertical segments and of crotchet heads, the application uses its music knowledge analysis, eliminating entities in the following way:

1. beam bars: horizontal tendency segments near a vertical one, which must be near a crotchet head.

2. crotchet heads : full ellipsis near a vertical segment.



3. stems: vertical segment near a crotchet head.

The remaining, last image is to be processed directly by neural network. The connected entities are introduced at the entry of the neural classifier, leading to the last stage of the recognition low level. Finally, it is possible to proceed the high level analysis of the document.

7 Conclusion

Presently, the first two steps are completely implemented and work in an automated way. The efficiency of the computation depends on the quality of the hand-written sheets. In the running tests we analyzed, segmentation met no problem, but we presume some problems may arise if the score is very bad and if there are many accidental connections between entities. With the hand-written scores we checked, the efficiency of the two-layer separation (stave detection) was of 100% and the segmentation reconstruction, up to 80%.

The neural network step is under implementation: we guess at a high recognition efficiency since the images resulting from the first steps are well extracted, without much noise: in such cases, RBF classifiers are known to be quite efficient [15].

References

[1] V. Poulain d'Andecy, J. Camillerapp, I. Leplumey *« Détecteur robuste de segments : application à l'analyse de partitions musicales. »* RFIA, neuvième congrès Reconnaissance des Formes et Intelligence Artificielle- Paris, France, janvier 1994

[2] V. Poulain d'Andecy, J. Camillerapp, I. Leplumey *« Analyse de partitions musicales »* L'écrit et le document, trairement du signal 1995- Volume 12-N° 6

[3] Kia C.Ng, Roger D. Boyle *« Segmentation of Music Primitives »* Proceeding of the British Machine Vision Conference, Leeds, 22-24 Septembre 1992

[4] B. Coüanon *« Formalisation grammaticale de la connaissance a priori pour l'analyse de documents : applications aux partitions d'orchestre. »* dixième congrès Reconnaissance des Formes et Intelligence Artificielle - Rennes, France, janvier 1996

[5] Kia C.Ng, Roger D. Boyle *« Recognition and reconstruction of primitives in music scores »* Image and Vision Computing 14 (1996) 39-46.

[6] M. Labarrere, J.P. Krief, B. Gimonet *« Le filtrage et ses applications »* Cepaudes éditions 1982

[7] Kia C.Ng, Roger D. Boyle David Cooper «Automated Optical Score Recognition and its Enhacement using High-level Musical Knowledge» XI Colloquio di informatica Musicale – Bologna 1995.

[8] B. Ostenstad *« Décomposition des objets d'une partition musicale numérisée en entités classables »*, Institut d'Informatique de l'Université d'Oslo, Rapport Nr. 31, Octobre 1988

[9] H. Miyao et Y. Nakano *« Head and stem from printed music scores using a neural network approach »* International conference on document analysis and recognition, Montreal Canada, 1995

[10] H. Kato et S. Inokochi, *« A recognition system for printed piano Music Using Musical Knowledge and constraints »*, Structured document Image Analysis, ed H. Baired, H Bunke et K. Yamamoto, pp. 435-445, springer-verlag, 1992

[11] P. Martin *« Réseaux de neurones artificiels : application à la reconnaissance optique de partitions musicales. »* Thèse de l'université Joseph Fourier- Grenoble I, Avril 1992

[12] D.E. Rumelhart, G.E. HINTON, R.J. Williams *« Learning internal representations by error propagation. Parallel distributed processings :explorations in the microstructure of cognition »*, vol 1, chap 8, MIT press, 1986.

[13] Y. Lecun, L.D. Jachel, H.P. Graf, B. Boser, J.S. Denker, I. Guyon, D. Henderson, R.E. Howard, W. Hubbard, S.A. Solla *«Optical charachter recognition and neural-net chips. »* INNC, p31, Boston 1988

[14] D. Broomhead, D. Lowe, *«Multivariable functional interpolation and adaptative networks»* Complex systems, vol 2, N°3, pp 321-355, 1998.

[15] S. Lecoeuche *«econnaissance de caratères industriels par application d'un système de réseaux de neurones à boucle de rétroaction»* PhD thesis, university des sciences et technologies de Lille, nov. *1998.*

[16] N.P. *Carter « Automatic recognition of printed music in context of electronic publishing »*, PhD thesis, University of Surrey, Depts of Physics and Musics, UK 1989.