Robust On-line Beat Tracking with Kalman Filtering and Probabilistic Data Association (KF-PDA)

Yu Shiu, *Student Member*, IEEE, Namgook Cho, *Student Member*, IEEE, Pei-Chen Chang and C.-C. Jay Kuo, *Fellow*, IEEE

Abstract — *A Kalman filtering (KF) approach to on-line musical beat tracking with probabilistic data association (PDA) is investigated in this work. We first formulate the beat tracking process as a linear dynamic system of beat progression, and then apply the Kalman filtering algorithm to the dynamic system in estimating the time-varying tempo and beat locations. Musical beat tracking using traditional Kalman filtering is however not reliable in the presence of tempo fluctuations and expressive timing deviations. To address this problem, we adopt data association techniques to assign probability masses to all possible beat interpretations, and then locate the true beat according to the weighting. Two methods are proposed. The first one (PDA-I) weighs the distance between the candidate observation and the predicted beat location while the second method (PDA-II) considers not only the distance but also the onset intensity in weight selection. Superior performance of the proposed beat tracking algorithm is demonstrated with simulation results on the Music Information Retrieval Evaluation Exchange (MIREX) 2006 beat tracking competition practice dataset and the Billboard Top-10 database*[1].

Index Terms — Musical signal processing, on-line beat tracking, Kalman filter, probabilistic data association, music information retrieval.

I. INTRODUCTION

When listening to music, most people even without musical education can grasp the speed of music and follow it by foot-tapping or hand-clapping along with beats. However, the same is not true for electronic devices. Automatic beat tracking has been an active area of research for more than twenty years. The beat is a fundamental unit of the temporal structure of music, especially to Western music, and beat tracking is an essential task in many musical applications such as musical analysis, synchronization, editing of musical sounds, and human-computer improvisation. This work presents an on-line (or causal) musical beat tracking system, where beat estimation at a given time depends only on past and present data.

Beat tracking is defined by estimating the possibly time-varying tempo and the time location of each beat, where the *beat* is referred to as the foot tapping and *tempo* as the beat rate [1]. Our research goal is to estimate the set of beat locations from musical audio signals sequentially. Ideally, when beat pulses are strong and the duration between adjacent beats is perceptually clear, automatic beat tracking can be done easily. Its performance nevertheless degrades significantly in practice due to several reasons. The first one comes from rest notes and missed-beat syncopation. The rest notes hide beat tracking cues, whereas syncopation does not have an onset pulse on expected beat location but with a small shift. The second one is due to variability in human performance. Even if a performer attempts to keep the duration between two consecutive beats constant throughout the whole music piece, the actual duration tends to vary along time. The last one is that some music pieces have time-varying tempo and, consequently, a time-varying beat period. The performance of beat tracking algorithms is often less robust when dealing with classical music, as compared with that containing drum sounds [1], [2].

Early work on automatic beat tracking was done by researchers in the fields of music perception and computer science [3]. More recently, Brown [4] used the autocorrelation function to examine the pulses in musical scores. Scheirer [5] applied a bank of comb filters to a musical signal at different fixed frequencies and searched for the filter that gives the strongest response for tempo estimation. Afterwards, the beat location was calculated by examining the phase of the filtered signal. Goto [2] developed an on-line beat tracking system that can process music with or without drum sounds. The system recognizes the hierarchical beat structure using three kinds of musical knowledge: onset times, chord changes, and drum patterns. A probabilistic generative model for tempo tracking was examined by Cemgil *et al.* [6],[7]. A Kalman filtering process was used to track beats in [6], which was followed by using the tempogram representation to assign probability masses to all possible beat candidates, while Monte Carlo methods were exploited to infer a hidden tempo variable in [7]. Hainsworth and Macleod [8] used particle filters to associate onsets from an audio signal to a time-varying tempo process so as to determine the beat locations.

Most of earlier work for beat tracking used symbolic or musical instrument digital interface (MIDI) data, *e.g.,* [4],[6],[7]. Audio signals have been examined more recently, *e.g.,* [2],[5],[8]. In addition, most previous beat tracking systems adopt a non-causal method that allows the use of future data and backward decoding, which is not suitable for real-time implementation in consumer electronic applications.

In this work, we present a method that extracts beat locations from acoustic musical signals, not limited to any particular music type, including both classical music and

[1] Part of this work was presented at ICCE2008, Las Vegas, NV USA.

The authors are with Department of Electrical Engineering and Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mails: atoultaro@gmail.com, namgookc@usc.edu, peichenc@usc.edu, and cckuo@sipi.usc.edu).

Manuscript received July 11, 2008

modern music with drums. Our research is motivated by the probabilistic model proposed by Cemgil *et al.* in [6]. More specifically, after pre-processing audio signals and extracting onsets, we formulate the beat tracking process as a linear dynamic system of beat progression by following the framework in [6]. Then, a Kalman filtering process can be applied to the dynamic system to estimate the hidden state, *i.e.*, the beat location and the period. However, beat estimation using *only* Kalman filtering process is not reliable in the presence of tempo fluctuations and expressive timing deviations.

To improve the tracking performance, a probabilistic approach can be used by assigning probability masses to all possible beat interpretations. Following this line of thought, the tempogram representation was adopted in [6]. Here, we adopt an alternative approach known as the probabilistic data association (PDA) technique to enhance the robustness of Kalman filtering. PDA has been widely used in real-time object tracking in computer vision [9]-[11] and radar applications [12],[13]. The proposed method, called KF-PDA, is theoretically elegant and computationally efficient as compared with the KF-tempogram approach [6]. For example, the switching mechanism required by the KF-tempogram approach to handle outliers is not needed in the proposed KF-PDA solution.

The basic idea of KF-PDA is briefly described below. First, a simple strategy, called local maximum selection, is used to choose the location of the predicted beat that has the maximum onset intensity among a set of beat candidates within a fixed window. Then, we consider two PDA methods. The first one (PDA-I) weighs the distance between the candidate observation and the predicted beat location while the second one (PDA-II) considers the distance as well as the onset intensity in weight selection. It is demonstrated by experimental results that the proposed beat tracking system leads to reliable performance even with tempo fluctuations and beat deviations.

The rest of this paper is organized as follows. Sec. II describes the pre-processing of music signals for beat tracking. A linear dynamic model of beat progression and Kalman filtering algorithm are given in Sec. III. In Sec. IV, beat selection techniques based on PDA are discussed. Finally, experimental results are given to compare the performance of two PDA methods in Sec. V, followed by concluding remarks and future research directions in Sec. VI.

II. MUSICAL DATA PRE-PROCESSING

The block-diagram of the proposed musical beat tracking algorithm is illustrated in Fig. 1. Given the acoustic waveform of a musical signal, musical data pre-processing is performed to extract temporal locations of musical onsets by two modules: 1) onset detection and 2) periodicity estimation. Typically, these tasks can be done within a local temporal interval of an analysis window and updated from one interval to the other. The temporal locations and intensities of onsets are used as the input data to the next module, *i.e.*, the Kalman filtering process. It should be noted that although there exist many techniques for onset detection and periodicity estimation, we choose classical algorithms in our implementation for their simplicity. Next, the Kalman filtering process is used to estimate the hidden state: temporal locations of beats and their period. At each step, we validate only measurements whose predicted probability is sufficiently high, and then select the best beat estimate by assigning probability masses to the validated measurements. In this section, we focus on music data pre-processing.

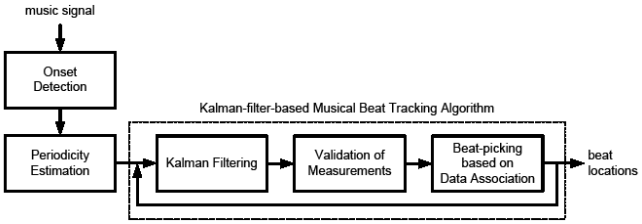


Fig. 1. Overview of the proposed musical beat tracking system.

A. Onset Detection

The aim of onset detection is to extract a detection function to indicate the locations of the most salient features of an audio signal [14]. These events are particularly crucial to beat perception and provided as an input to the proposed musical beat tracking system as shown in Fig. 1. The onset detection task falls into two categories: detection of percussive events and harmonic changes [8]. The transient events, usually coming from drum sounds, have strong energy changes while the change of musical piches/harmonies, usually due to the arrival of a new note, is associated with small energy changes.

Here, we adopt a cepstral distance method to calculate the musical onset detection function as described below. The discrete Fourier transform of the input audio signal is calculated for every 20-ms time frame, which is Hanning-windowed and overlapping with each other by 50%. In each frame, the spectrum is mapped onto the mel-scale using the triangular mel-scale filter bank. Then, the mel-frequency cepstral coefficients (MFCCs) are calculated by taking the cosine transform of a log power spectrum on the mel-scale frequency, denoted by the *m*th cepstral coefficient in the *n*th time frame, $c_m(n), m = 0, \cdots, L$, where L is the order of the cepstral coefficients.

Since low-order MFCCs are highly correlated to the mel-scale energy envelope of audio signals, we choose four low-order coefficients to represent the energy change of audio signals. Specifically, the 0th order coefficient, $c_0(n)$ represents exactly the mel-scale energy while three low-order coefficients, $c_1(n)$, $c_2(n)$, and $c_3(n)$, capture well the energy change of harmonic sounds. Then, the chosen coefficients are averaged over p consecutive time frames, *i.e.*, $c_m(n), \cdots, c_m(n-p+1)$ to represent the smoothed coefficients, $\hat{c}_m(n), m = 0, \cdots, 3$ at time frame n. We selected

$p = 3$ in our experiments so that even a fast energy change in the underlying music signal can be captured well. Finally, the musical onset detection function is calculated by finding the difference between the two consecutive smoothed cepstral coefficients as

$$d(n) = \frac{1}{L}\sum_{m=1}^{L}\left(\hat{c}_m(n) - \hat{c}_m(n-1)\right)^2. \quad (1)$$

Low amplitude peaks in the detection function can be discarded by thresholding [14]. Then, onsets are localized in the detection function by identifying local maxima above the threshold.

B. Period Estimation

The detection function in (1) at the output of the onset detection is observed as a quasi-periodic and noisy pulse-train. The aim of periodicity estimation is to find the coarse periodicity of the detection function for the next step, *i.e.*, the Kalman filtering process. For periodicity estimation, we assume that the tempo of the music signal is constant over the time interval of the analysis window (yet it can evolve slowly from one to the other). To this end, we use the classical autocorrelation function as

$$ACF(\tau) = \sum_n d(n)d(n-\tau), \quad (2)$$

where $d(n)$ is the onset detection function and τ is a delay parameter.

However, the autocorrelation function of real-world musical onset signals does not exhibit ideal periodicity so that it is not easy to find the exact peak of a period. For example, there often exists confusion between the real period and its double/half-period (or triple/one-third-period for the triplet case) in the musical sound.

III. Beat Tracking with Kalman Filtering

Musical tempo can be modeled as a hidden state variable of a stochastic dynamic system. Here, we employ the linear dynamic model proposed by Cemgil *et al.* [6] to formulate the beat tracking problem in a probabilistic framework, where beats are estimated by a Kalman filter under the noisy environment. Kalman filtering, widely used in object tracking applications [12],[13],[15], exploits the dynamics of the target to remove the noise effect and obtain a good estimate of the target location.

A. Linear Dynamic Model of Beat Progression

A perfect metronome can be described as a dynamical system with two state variables [6]: beat τ and period $\hat{\Delta}$. Letting the state variable at the *k*th step be $\mathbf{x}_k = [\hat{\tau}_k, \hat{\Delta}_k]^T$, the linear state transition model can be written as

$$\mathbf{x}_k = \mathbf{\Phi}_k \mathbf{x}_{k-1} + \mathbf{w}_{k-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \mathbf{w}_{k-1}, \quad (3)$$

where $\mathbf{w}_k$ is the process noise which is modeled as a zero-mean multivariate normal distribution with covariance, $\mathbf{Q}_k$, *i.e.*, $\mathbf{w}_k \sim N(\mathbf{0}, \mathbf{Q}_k)$.

At time *k,* observation $\mathbf{y}_k$ of true state $\mathbf{x}_k$ can be written as

$$\mathbf{y}_k = \mathbf{M}_k \mathbf{x}_k + \mathbf{v}_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k + \mathbf{v}_k, \quad (4)$$

where $\mathbf{v}_k$ is the observation noise which is modeled as the zero-mean Gaussian white noise with covariance, $\mathbf{R}_k$, *i.e.*, $\mathbf{v}_k \sim N(\mathbf{0}, \mathbf{R}_k)$.

The initial state and noise vectors at each step are assumed to be mutually independent. With this formulation, hidden state $\hat{\mathbf{x}}_k$ can be estimated sequentially in the time domain based on the estimated state from the (*k-1*)th time step and the current measurement.

B. Beat Estimation via Kalman Filtering

Given the linear dynamic model in (3) and (4), the Kalman filtering process can be used to estimate the hidden state of beats efficiently. In what follows, we use $\hat{\mathbf{x}}_{n|m}$ to represent the estimate of $\mathbf{x}$ at time *n* given observations up to and including time *m*.

The Kalman filter has two distinct phases (*i.e.,* predict and update) [16] as summarized below. The predict phase uses the state estimate from the previous time step to produce an estimate of the state at the current time step. It can be written as

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{\Phi}_k \hat{\mathbf{x}}_{k-1|k-1} \quad (5)$$

$$\mathbf{P}_{k|k-1} = \mathbf{\Phi}_k \mathbf{P}_{k-1|k-1} \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{Q}_{k-1} \quad (6)$$

where $\hat{\mathbf{x}}_{k|k-1}$ and $\mathbf{P}_{k|k-1}$ represent the predicted state and covariance, respectively. In the update phase, the measurement information at the current time step is used to refine this prediction to derive a more accurate state estimate as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{M}_k^{\mathrm{T}}\left(\mathbf{M}_k \mathbf{P}_{k|k-1}\mathbf{M}_k^{\mathrm{T}} + \mathbf{R}_k\right)^{-1} \quad (7)$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\left(\mathbf{y}_k - \mathbf{M}_k \hat{\mathbf{x}}_{k|k-1}\right) \quad (8)$$

$$\mathbf{P}_{k|k} = \left(\mathbf{I} - \mathbf{K}_k \mathbf{M}_k\right)\mathbf{P}_{k|k-1} \quad (9)$$

where $\mathbf{K}_k$ is called the Kalman gain. These two phases are updated alternatively to estimate the hidden state of the dynamic system from a series of incomplete and noisy measurements.

IV. Beat Selection with Probabilistic Data Association (PDA)

In a real world situation, there exist non-beat onsets in musical notes and percussive sounds although beats tend to have large onset values. The beat tracking performance based on *only* Kalman filtering with onset inputs is however poor.

To improve its performance, we need a more intelligent way to locate true beats from the audio waveform. In this section, we apply the probabilistic data association (PDA) technique, which was proposed for real-time object tracking in computer vision [9]-[11] and radar applications [12],[13], to estimate beat locations robustly in the presence of noisy measurements.

A. Local Maximum (LM) Selection Rule

The Local Maximum (LM) selection rule is a simple strategy that chooses the beat location at the *k*th step from a set of possible beat interpretations. Basically, it selects the onset that has the maximum magnitude within a fixed window around the predicted beat location $\hat{\tau}_{k|k-1}$ in the Kalman filtering process as shown in Fig. 2.

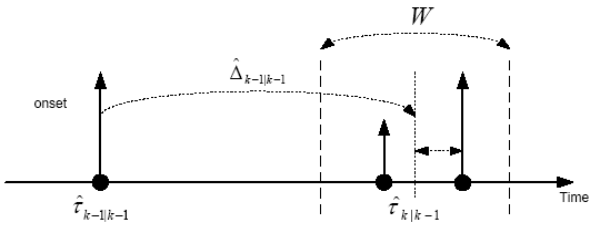


Fig. 2. Beat selection in the Kalman filtering process with the LM selection rule. The arrows represent onsets obtained from the musical data pre-processing modules.

Mathematically, the LM selection rule can be written as

$$\tau_k = \underset{|n-\hat{\tau}_{k|k-1}|<W/2}{\arg\max}\ d(n), \tag{10}$$

where τ_k is the beat at time step *k* obtained from the LM selection rule, *W* is a fixed-window length and the predicted beat location can be obtained by

$$\hat{\tau}_{k|k-1} = \hat{\tau}_{k-1|k-1} + \hat{\Delta}_{k-1|k-1}.$$

In practice, when the LM selection rule is adopted, its performance can still be confused by onsets of non-beat notes or percussive sounds, which have stronger energy than true beat onsets.

Fig. 3 shows such an example, where a segment from *We Didn't Start the Fire* by Billy Joel is illustrated. The top sub-figure of Fig. 3 shows the spectrogram of the music sound from 30s to 35s, which is re-labeled as 0 to 5s. The bottom sub-figure illustrates the corresponding detection function obtained by the musical data pre-processing procedure as discussed in Sec. II. The music sound has strong beats from percussion in the first 3 seconds. We see that the detection function behaves like a pulse train with a fixed interval between consecutive pulses. However, beat notes do not have stronger energy values than non-beat notes in the time interval between 3.0s and 4.5s. When the beat note does not have the strongest musical onset in the neighborhood of predicted beat location $\hat{\tau}_{k|k-1}$, the LM selection rule fails. For example, there are 3 strong pulses around 3.7s in Fig. 3, denoted by A, B and C, where A and C are true beats while B is a note of the half-beat metrical structure. In this situation, since $\hat{\tau}_{k|k-1}$ is around 3.7s and onset B has a musical intensity larger than onset A within a window *W*, the Kalman filtering process based on the LM selection rule selects onset B as new observation $\mathbf{y}_k$. Thus, the newly estimated beat location $\hat{\tau}_{k|k}$ will be located between onsets A and B (instead of onset A, which should be the case if the observation is chosen to be onset A). Then, at time step *k*, it is possible to select onset D as observation $\mathbf{y}_{k+1}$. If this happens, the Kalman filtering process starts to track wrong beat locations from this point on. In conclusion, the performance of beat tracking using the Kalman filter with the LM selection rule degrades due to the existence of non-beat notes and/or percussive sounds.

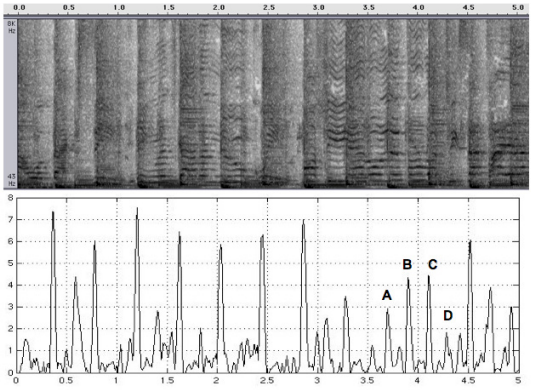


Fig. 3. Billy Joel's *We Didn't Start the Fire*: (top) the spectrogram of a music segment from 30:00 to 35:00 seconds, where the y-axis represents the frequency from 0 to 8 kHz; (bottom) the detection function as a function of time (in the unit of seconds) for the same music segment.

B. Measurement Validation

To overcome the weakness of the LM selection rule, we adopt the PDA technique that has been widely used for target tracking in a cluttered environment, where a clutter denotes a set of objects that are close to the target yet random in location or intensity. Any non-target in such an environment may cause confusion in the tracking of real targets. The PDA technique [9]-[12] provides a probabilistic method to associate observations (or measurements) with the target of interest in the cluttered environment. Ideally, PDA should choose any measurement that is originated from the target of interest and discard other measurements contributed by random noise and/or interference. Instead of using the deterministic framework as discussed in Sec. IV-A, we will employ the Bayesian probabilistic approach to consider all candidate observations simultaneously and probabilistically.

Here, a two-step approach is implemented to track musical beats robustly: *validation* and *association of measurements*. The first step is to validate measurements whose predicted probability is high. In other words, the observation validation process aims to remove measurements that are unlikely to be a correct target in the next step. In this subsection, the observation validation method for musical beat tracking is discussed and data association techniques to choose beats robustly will be described in the following subsections.

A validation region can be viewed as a subspace in the observation space and often obtained by applying a multi-dimensional probabilistic threshold [12]. To derive the validation region for musical beat tracking, we begin with the predicted observation

$$\hat{\mathbf{y}}_{k|k-1} = \mathbf{M}_k \hat{\mathbf{x}}_{k|k-1}, \tag{11}$$

and the associated covariance matrix can be written as

$$\begin{aligned}\mathbf{S}_k &= E\left[\left(\mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}\right)\left(\mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}\right)^{\mathrm{T}} | \mathbf{Y}^{k-1}\right] \\ &= \mathbf{M}_k \mathbf{P}_{k|k-1} \mathbf{M}_k^{\mathrm{T}} + \mathbf{R}_k,\end{aligned} \tag{12}$$

where $E[\xi]$ is the expected value of ξ and $\mathbf{Y}^{k-1}$ is a set of validated measurements up to time (*k-1*): $\mathbf{Y}^{k-1} = \{\mathbf{y}_j\}_{j=1}^{k-1}$.

If the true measurement at time *k* conditioned upon $\mathbf{Y}^{k-1}$ is normally distributed, the probability distribution of $\mathbf{y}_k$ can be expressed as

$$p\left[\mathbf{y}_k | \mathbf{Y}^{k-1}\right] = N\left(\hat{\mathbf{y}}_{k|k-1}, \mathbf{S}_k\right). \tag{13}$$

Then, a region can be defined in the measurement space, which has a higher probability [12],

$$\tilde{V}_k(\gamma) = \left\{\mathbf{y} : \tilde{\mathbf{y}}_k^{\mathrm{T}} \mathbf{S}_k^{\mathrm{T}} \tilde{\mathbf{y}}_k \le \gamma\right\}, \tag{14}$$

where γ is a threshold and $\tilde{\mathbf{y}}_k$ is a measurement residual or innovation defined by

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}. \tag{15}$$

By choosing a proper value of γ, we can determine the validation region, $\tilde{V}_k(\gamma)$. Under (15), measurements that lie inside the region are considered valid while those outside are discarded.

It was shown in [12] that the weighted norm of the measurement residual from (14) is chi-square distributed with the degree of freedom equal to the dimension of the measurement vector. By choosing $\gamma = 9$ in (14), the probability for the region to contain true measurements is 99.7%, whereas the choice of $\gamma = 4$ produces a probability of 95.4 % [12].

Based on (12) and (14), we derive the validation region for the proposed musical beat tracking algorithm as

$$\tilde{V}_k(\gamma) = \left\{\mathbf{y} : \frac{\left|\mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}\right|^2}{p_{11} + \sigma_v^2} \le \gamma\right\}, \tag{16}$$

where p_{11} and σ_v^2 are the variance of beat location τ_k and observation noise $\mathbf{v}_k$, respectively. Given γ, p_{11}, and σ_v^2, the validation region can be calculated for each predicted observation $\hat{\mathbf{y}}_{k|k-1}$ as

$$\hat{\mathbf{y}}_{k|k-1} - \gamma\left(p_{11} + \sigma_v^2\right) \le \mathbf{y} \le \hat{\mathbf{y}}_{k|k-1} + \gamma\left(p_{11} + \sigma_v^2\right). \tag{17}$$

To summarize, the validation procedure limits the region in the measurement space to search for the beat of interest robustly. After the validation, PDA adopts a strategy to associate valid measurements with probabilities as described below.

C. Basic Probabilistic Data Association (PDA-I)

Within the validation region, the next step is to perform the probabilistic data association (or weighting), which is derived with the following assumption

$$p\left[\mathbf{x}_k | \mathbf{Y}^{k-1}\right] = N\left(\hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1}\right), \tag{18}$$

i.e., the state is Gaussian-distributed with a mean and an error covariance matrix. Mathematically, PDA decomposes an estimate into a linear combination of estimate from measurements inside the validation region as

$$\begin{aligned}\hat{\mathbf{x}}_{k|k} &= E\left[\mathbf{x}_k | \mathbf{Y}^k\right] \\ &= \sum_{i=0}^{m_k} E\left[\mathbf{x}_k \mid \theta_i(k), \mathbf{Y}^k\right] p\left[\theta_i(k) \mid \mathbf{Y}^k\right] \\ &= \sum_{i=0}^{m_k} \hat{\mathbf{x}}_{i,k|k} \beta_i(k),\end{aligned} \tag{19}$$

where m_k is the number of measurements in the validation region. $\hat{\mathbf{x}}_{i,k|k}$ is the updated state estimate conditioned on event $\theta_i(k)$ which means that $\hat{\mathbf{y}}_{i,k}$ is the measurement originating from the target beat, while $\theta_0(k)$ represents the event that none of the measurements originating from the target beat. $\beta_i(k)$ in (19) is the probability of even $\theta_i(k)$ conditioned on given measurements*, i.e.,*

$$\beta_i(k) = p\left[\theta_i(k) \mid \mathbf{Y}^k\right], \quad i = 0, \cdots, m_k. \tag{20}$$

These events are mutually exclusive and $\sum_{i=0}^{m_k} \beta_i(k) = 1$.

To make the basic Kalman filtering process compatible with PDA, some steps in the Kalman filter algorithm as described in Sec. III-B have to be modified. We begin with the update of state vector $\hat{\mathbf{x}}_{k|k}$ from $\hat{\mathbf{x}}_{k|k-1}$ as

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k\left(\mathbf{y}_k - \mathbf{M}_k \hat{\mathbf{x}}_{k|k-1}\right). \tag{21}$$

With (20) and (21), $\hat{\mathbf{x}}_{k|k}$ can be represented by

$$\begin{aligned}\hat{\mathbf{x}}_{k|k} &= \sum_{i=0}^{m_k} \hat{\mathbf{x}}_{i,k|k} \beta_i(k) \\ &= \hat{\mathbf{x}}_{k|k-1} \sum_{i=0}^{m_k} \beta_i(k) + \mathbf{K}_k \sum_{i=0}^{m_k} \left(\mathbf{y}_{i,k} - \mathbf{M}_k \hat{\mathbf{x}}_{k|k-1}\right) \beta_i(k) \\ &= \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \boldsymbol{\Psi}_k,\end{aligned}$$

where

$$\boldsymbol{\Psi}_k = \sum_{i=1}^{m_k} \left(\mathbf{y}_{i,k} - \mathbf{M}_k \hat{\mathbf{x}}_{k|k-1}\right) \beta_i(k). \tag{22}$$

The state vector update formula in (21) is kept intact except for the replacement of the prediction residual with its weighted version, $\mathbf{\Psi}_k$. Since the quantity, $\mathbf{y}_{i,k} - \mathbf{M}_k \hat{\mathbf{x}}_{k|k-1}$, is called the innovation, $\mathbf{\Psi}_k$ can be viewed as the equivalent innovation for PDA, which is the weighted average of innovations from all validated measurements. Another needed modification is to compute the error covariance matrix $\mathbf{P}_{k|k}$. We only state the results below and refer to [12] for details:

$$\mathbf{P}_{k|k} = \beta_0(k)\mathbf{P}_{k|k-1} + \left[1-\beta_0(k)\right]\mathbf{P}^0_{k|k} + \widetilde{\mathbf{P}}_k, \tag{23}$$

where $\mathbf{P}^0_{k|k} = \left(\mathbf{I} - \mathbf{K}_k\mathbf{M}_k\right)\mathbf{P}_{k|k-1}$ is the covariance of $\hat{\mathbf{x}}_{k|k}$, and

$$\widetilde{\mathbf{P}}_k = \mathbf{K}_k\left[\sum_{i=1}^{m_k}\beta_i(k)\widetilde{\mathbf{y}}_{i,k}\widetilde{\mathbf{y}}^{\mathrm{T}}_{i,k} - \widetilde{\mathbf{y}}_{i,k}\widetilde{\mathbf{y}}^{\mathrm{T}}_{i,k}\right]\mathbf{K}^{\mathrm{T}}_k, \tag{24}$$

and where $\widetilde{\mathbf{y}}$ is the measurement residual defined in (15).

For musical beat tracking, a validation region is calculated at each sample time. If more than one measurement is found in the validation region at a given time for a beat, then PDA is applied to all of them. In other words, all measurements in the validation region are used for beat estimation.

D. Enhanced Probabilistic Data Association (PDA-II)

PDA-I exploits the weighted average of innovations from all validated measurements $\mathbf{y}$ as shown in (22). As defined in (20), weight $\beta_i(k)$, also known as the association probability, is related to the distance between candidate measurement $\mathbf{y}_{i,k}$ and $\mathbf{M}_k\mathbf{x}_{k|k-1}$. The smaller the distance between them, the larger the probability is. However, in musical beat tracking, human uses not only the closeness of the measurement and the predicted beat location but also the onset intensity as cues to select the next beat location. Motivated by the observation, we propose an enhanced PDA method.

It is worthwhile to mention that modification of the association probability has been considered by researchers before in various contexts to improve the tracking performance, *e.g.*, in visual object tracking [9], [11] and radar applications [13]. The former uses both the prediction residual and image similarity as cues for object tracking while the latter uses the prediction residual and the intensity of the reflected radar signal as cues for airplane tracking. Here, we use both the prediction residual and the intensity of the measurement to improve the musical beat tracking performance .

The intensity of the observed signal is introduced to the association probability calculation via

$$\begin{aligned}\beta_i^{enh}(k) &= p\left[\theta_i(k) \mid I_Y(k), \mathbf{Y}^k\right] \\ &\propto p\left[I_Y(k) \mid \theta_i(k), \mathbf{Y}^k\right] p\left[\theta_i(k) \mid \mathbf{Y}^k\right],\end{aligned} \tag{25}$$

where $I_Y(\cdot)$ is the distribution function of the onset intensity. As shown in (25), new weight $\beta_i^{enh}(k)$ is the product of two terms. The first term is contributed by the onset intensity while the second term is associated with the prediction residual that is actually equal to the weight defined in (20). More specifically, the first term contributed by the onset intensity can be further decomposed as [11]

$$p\left[I_Y(k) \mid \theta_i(k), \mathbf{Y}^k\right] = \frac{I_i(\mathbf{y}_i)}{I_0(\mathbf{y}_i)}\prod_{j=1}^{m_k} I_0(\mathbf{y}_j), \tag{26}$$

where $I_i(\mathbf{y}_i), i = 1, \cdots, m_k$, is the probability distribution of validated measurement $\mathbf{y}_i$ and $I_0(\mathbf{y}_j)$ is the probability distribution of $\mathbf{y}_j$ when it is not the measurement corresponding to the target beat. We can re-write (26) as

$$p\left[I_Y(k) \mid \theta_i(k), \mathbf{Y}^k\right] = I_i(\mathbf{y}_i)\prod_{j=1, j\neq i}^{m_k} I_0(\mathbf{y}_j). \tag{27}$$

It is difficult to compute $I_i(\mathbf{y}_i), 1 \le i \le m_k$, in (27) efficiently and accurately for two reasons. First, the number of candidate measurement, m_k, is determined dynamically by validated region $\widetilde{V}_k(\gamma)$ at each time step *k*. Second, for particular *i* and *k*, there are not enough samples in estimating probability distribution $I_i(k)$ accurately. To address this issue, $I_i(k)$ is replaced by a fixed probability distribution of the validated measurement for all *i* and *k*. That is, we approximate

$$I_i(k) \approx I_B, \quad i = 1, \cdots, m_k, \tag{28}$$

where I_B is the probability distribution of onset intensities at beat locations, which can be determined statistically. On the other hand, $I_0(\mathbf{y}_i)$ is the probability distribution of observations $\mathbf{y}_i$ at non-beat locations. Similarly, since it is difficult to get the accurate distribution specific *i* and *k*, we approximate it by a fixed probability distribution

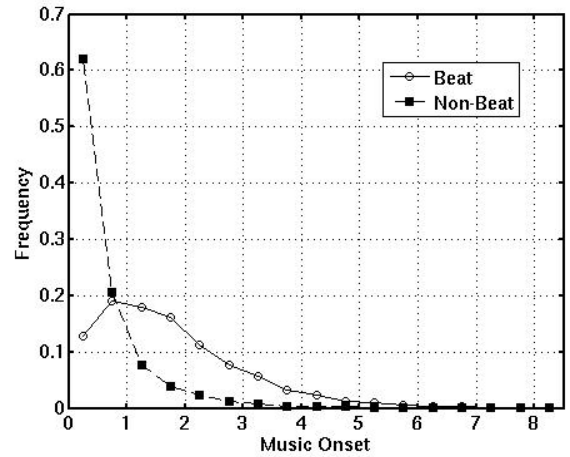
$$I_0(k) \approx I_N, \tag{29}$$

where I_N is the probability distribution of onset intensities at non-beat locations.

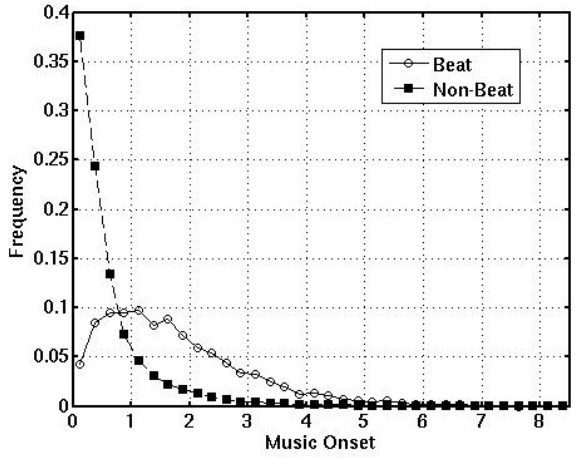
E. Probability Distribution Estimation

To estimate probability distributions I_B and I_N in (28) and (29), we adopt a non-parametric approach and use part of the MIREX 2006 beat tracking competition dataset [17] as the training data. This database consists of twenty 30s music clips with annotated beat locations. The first 10s music signals were used to determine I_B and I_N. The histograms of I_B and I_N as functions of the onset intensity are shown in Fig. 4, where the y-axis is the frequency of occurrence and the x-axis is the onset intensity. Bin centers are located uniformly between 0.25 and 8.25 seconds with bin width of 0.5 second in (a) and

between 0.125 and 8.125 seconds with bin width of 0.25 second in (b), respectively. We see from Fig. 4 that the onset distribution for non-beats heavily concentrates on small onset values with few large onset values. The non-beat type has a much higher probability than the beat type in the first bin in Fig. 4 (a). For the second bin centered at 0.75, the beat and the non-beat types have comparable probabilities around 20% (0.190 for beats and 0.207 for non-beats). For the third bin centered in 1.25 and above, the onset probability for beats is larger than that for non-beats. By refining the bin width from 0.5 to 0.25, we can get a better resolution for small onset values as shown in Fig. 4 (b).



(a) Bin width of 0.5 second.



(b) Bin width of 0.25 second.

Fig. 4. Probability distributions of I_B and I_N as a function of music onset intensities.

V. EXPERIMENTAL RESULTS

A. Experimental Setup and Evaluation Metrics

Musical sounds in our experiments were obtained from two music databases. The first one was the MIREX 2006 beat tracking competition dataset that has twenty 30-second music clips of diverse genres including classic, pop, rock, blues, and foreign-language pop songs. Beats in each audio clip were listened and verified by more than 40 people. The tempo of the audio clip was not known to the listeners so that they might use a different period to label beats. The second dataset used in our experiments was twenty Billboard Top-10 songs in 80's [18], which contains various genres including pop, rock, and some adult contemporary from singer-songwriters such as Billy Joel. For each song, a 60-second music clip was segmented from the original. The candidate tempos were estimated from the autocorrelation method and manually selected. Then, beats were labeled based on the selected period to serve as the ground truth. To be compatible with the MIREX 2006 dataset, only the first 30 seconds of the Billboard Top-10 dataset were used. All audio signals were sampled at 44.1-kHz rate with 16-bit resolution.

The first 10 seconds of the music clips were used to determine the probability distributions I_B and I_N as discussed in Sec. IV-E. The next 5 seconds were exploited to initialize the state vector of the Kalman filter such as $\mathbf{x}_0$. The performance of the proposed musical beat tracking system was evaluated with the remaining 15 seconds of these musical clips. A detection function is derived, as discussed in Sec. II, at a sampling rate of 100 Hz, and the peak-picking scheme [14] is used to locate onsets.

Two well-known metrics were exploited to evaluate the musical beat tracking performance. Being similar to the *P*-score evaluation in MIREX 2006 [17], we choose the first evaluation metric as

$$P = \frac{1}{N_{\max}} \sum_{n=-\infty}^{\infty} \sum_{m=-W}^{W} \delta_d(n)\delta_g(n-m), \tag{28}$$

where δ_d and δ_g are unit pulse functions in the detected and the ground-truth beat locations, respectively. Note that δ_d and δ_g take binary values only (*i.e.*, 0 and 1). In (28), $2W$ is the tolerable window size and $N_{\max}$ is defined as

$$N_{\max} = \max\left(N_d, N_g\right),$$

where N_d and N_g are the detected and the ground-truth beat numbers, respectively. The window size $2W$ was chosen to be 20% of the beat duration throughout the experiment. From the definition of the metric, the *P* value lies between 0 and 1, where the higher the *P* value, the better the performance. Note that if there are false alarms, the *P* value will be penalized by a larger value of N_d.

The second evaluation metric is the Longest Tracked Music Segment Ratio (LTMSR) [1], [2]. In the metric, we normalize the longest music segment with all its beats correctly tracked by the total duration of the same clip, *i.e.*, 15 seconds in the MIREX dataset. It shows how long the beat tracking algorithm can maintain the accurate tracking once it starts to track. Thus, its value lies between 0 and 1. It is noticed that even if a single beat is missed in the tracking process, this metric drops significantly. For example, for a clip of 15

seconds, if all beats are correctly detected except one at the 9th second, the performance will drop from 100% (=15/15) to 60% (=9/15). In contrast, the *P*-score metric can still be as high as 96%.

B. Performance Evaluation with P-Score Metric

Table I shows the average beat tracking performance for the MIREX dataset, where the proposed KF-based beat tracking algorithm was used, associating with various beat selection strategies. We see that PDA-II improves the performance greatly over LM by an average of 13.25%. In contrast, PDA-I and LM have performance similar.

The performance of LM and PDA-II for each music clip from the MIREX dataset is further compared by the scatter plot shown in Fig. 5, where each diamond-shape dot represents a music clip and its x- and y-coordinates represent the performance using LM and PDA-II, respectively. 14 dots are concentrated in the top-right corner of the figure, where the performance of LM is between 80% to 90% and that of PDA-II ranges from 95% to 100%. For the remaining six cases, three dots are above the 45-degree line (*i.e.*, the dotted line), which implies that PDA-II still performs better than LM. LM performs slightly better than PDA-II in the two cases, where two dots are below the 45-degree line. Finally, there is a one dot in the bottom-left corner of the figure, for which both LM and PDA-II perform poorly. It is "train13.wav'' in the MIREX dataset, which has a very fast tempo (178 beat per minutes) and strong onsets from the metrical level of the half beat. Consequently, they result in serious confusion on beat detection.

The *P*-score performance for the Billboard Top-10 dataset is shown in Table II along with that for the MIREX dataset for comparison. For the Billboard Top-10 dataset, the *P*-scores of LM, PDA-I and PDA-II all improve, as compared with the MIREX dataset, but with a different degree. The improvement of PDA-I is most significant while PDA-II still offers the best performance. Actually, except for poorer performance with Elton John's *Candle in the Wind*, PDA-II can achieve a *P*-score higher than 94%. In contrast, LM has similar performance for the MIREX dataset and the Billboard Top-10 dataset. There is only a small difference of 3.72% between the two datasets when using LM. The results can be explained as follows. The MIREX dataset has rather diverse genres while songs in the Billboard Top-10 dataset are more homogeneous. For the latter, since all of them received great commercial success, they resorted to average people's music taste. Generally speaking, the Billboard dataset has more regular beats throughout each music clip than the MIREX dataset.

C. Performance Evaluation with LTMSR Metric

Next, we tested the proposed beat tracking algorithms on the same datasets with the second evaluation metric, LTMSR, which emphasizes the robustness of the tracking performance in the presence of beat variation, rest notes and noisy measurements. The results are shown in Table III. For the MIREX dataset, the performance of PDA-I is worse than that of LM, which implies that the onset intensity may play a more important role than the prediction residual in beat tracking for the MIREX dataset. PDA-II outperforms LM and PDA-I by 12.36% and 26.71%, respectively. For the Billboard top-10 dataset, PDA-I has slightly better performance than LM. The performance of PDA-II is 90.88%, which is significantly better than LM and PDA-I. As compared with the MIREX dataset, LM, PDA-I and PDA-II all have better LTMSR performance for the Billboard Top-10 dataset, which can be explained by the homogeneity of beats in the Billboard Top-10 dataset.

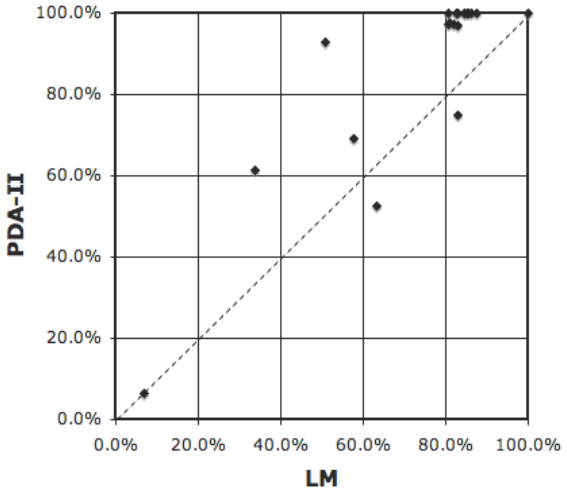


Fig. 5. The performance of the Kalman-filter-based beat tracking algorithm with LM and PDA-II for each music clip.

TABLE I
P-SCORE COMPARISON WITH MIREX DATASET

	LM	PDA-I	PDA-II
MIREX	74.08%	72.67%	87.33%

TABLE II
P-SCORE COMPARISON WITH MIREX AND BILLBOARD DATASETS

	LM	PDA-I	PDA-II
MIREX	74.08%	72.67%	87.33%
BILLBOARD	77.80%	87.81%	94.68%

TABLE III
LTMSR COMPARISON WITH MIREX AND BILLBOARD DATASETS

	LM	PDA-I	PDA-II
MIREX	66.18%	51.83%	78.54%
BILLBOARD	73.42%	78.98%	90.88%

VI. CONCLUSION AND FUTURE WORK

A Kalman-filter approach to musical beat tracking with three measurement selection rules were examined in this work. The simple LM measure selection rule chooses the onset that has the maximum intensity within a fixed window. The basic PDA method (PDA-I) considers only the prediction residual value while the enhanced PDA method (PDA-II) incorporates both the prediction residual and the onset intensity. PDA-II gives the best performance among the three for two test music databases: the MIREX 2006 competition dataset and the Billboard Top-10 dataset.

We should mention that our experimental results are still preliminary. More extensive test of the proposed KF-PDA method should be conducted in the near future. In particular, it is worthwhile to compare the proposed scheme with the scheme proposed by Cemgil *et al.* [6],[7] on a wide range of test datasets.

REFERENCES

[1] A. P. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Speech Audio Process.,* vol. 14, no. 1, pp. 342-355, Jan. 2006.

[2] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.,* vol. 30, pp. 159-171, 2001.

[3] C. S. Lee, *The perception of metrical structure: experimental evidence and a mode*, Academic Press: New York, 1991.

[4] J. C. Brown, "Determination of the meter of musical scores by autocorrelation," *J. Acoust. Soc. Amer.,* vol. 94, pp. 1953-1957, 1993.

[5] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.,* vol. 103, pp. 588-601, 1998.

[6] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: tempogram representation and Kalman filtering," *J. New Music Res.,* vol. 28, pp. 259-273, 2001.

[7] A. T. Cemgil, and B. Kappen, "Monte Carlo methods for tempo tracking and rhythm quantization," *J. Artif. Intell. Res.,* vol. 18, pp. 45-81, 2003.

[8] S. Hainsworth, and M. Macleod, "Beat tracking with particle filtering algorithms," in *Proc. IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics,* New Paltz, NY, 2003, pp. 91-94.

[9] C. Rasmussen, and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 23, pp. 560-576, 2001.

[10] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 25, pp. 564-575, 2003.

[11] C.-M. Huang, D. Liu, and L.-C. Fu, "Visual tracking in cluttered environments using visual probabilistic data association filter," *IEEE Trans. Robot.,* vol. 22, pp. 1292-1297, Dec., 2006.

[12] Y. Bar-Shalom and T. E. Fortmann, *Traking and data association*, Academic Press: Orlando, FL, 1988.

[13] Y. Bar-Shalom and X. R. Li, *Multitarget-multisensor tracking: applications and advances*, Artech House: Norwood, MA, 2000.

[14] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.,* vol. 13, pp. 1035-1047, Sep. 2005.

[15] Y. Boykov and D. Huttenlocher, "Adaptive Bayesian recognition in tracking rigid objects," in *Proc. IEEE Conf. on Computer Vision and Pattern Recog.,* Hilton Head, SC, 2000, pp. 697-704.

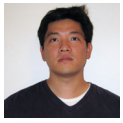
[16] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. of Basic Engineering,* vol. 82, pp. 35-45, 1960.

[17] Available at http://www.music-ir.org/mirexwiki/index.php.

[18] Available at http://www.billboard.com/bbcom/index.jsp.

Yu Shiu (S'06) received the B.S. and M.S. from Electrical Engineering Department in National Taiwan University in 1996 and 1998, respectively. After getting another M.S. specializing in speech signal processing from UCLA in 2002, he joined Prof. Kuo's Multimedia Communications Lab in USC. His research interests include musical information retrieval, musical beat tracking, content-based audio/video signal segmentation, video indexing via audio features, and audio source separation.

Namgook Cho (S'07) received the B.S. and M.S. degrees in electrical engineering from Inha University, Incheon, Korea, in 1994 and 1996, respectively. From 1996 to 2001, he was with Research & Development Division in Hyosung Corporation, Seoul, Korea. He is currently a Ph.D. candidate in the Ming Hsieh Department of Electrical Engineering at the University of Southern California, and a member of Media Communications Lab. led by Prof. C.-C. Jay Kuo. His research interests include efficient representations of audio signals with content-adaptive dictionaries and multichannel audio source separation in noisy environments.



Pei-Chen Chang is a Ph.D. student in the Ming Hsieh Department of Electrical Engineering, University of Southern California. He received his B.S. Degree in Computer Science from the National Chiao-Tung University, Hsinchu in 1999 and the M.S. degree in Computer Science from the National Taiwan University, Taipei in 2001, respectively. He was a R & D engineer in MediaTek Inc. from 2002 to 2007. He is currently a member of Prof. Kuo's research group. His research interests include audio signal processing, audio coding, and audio information retrieval.

C.-C. Jay Kuo (S'83-M'86-SM'92-F'99) received the B.S. degree from the National Taiwan University, Taipei, in 1980 and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively, all in Electrical Engineering. He is presently Director of the Signal and Image Processing Institute (SIPI) and Professor of Electrical Engineering, Computer Science and Mathematics at the University of Southern California (USC). His research interests are in the areas of digital image/video analysis and modeling, multimedia data compression, communication and networking and multimedia database management. Dr. Kuo has guided about 90 students to their Ph.D. degrees and supervised 20 postdoctoral research fellows. He is a co-author of about 150 journal papers, 750 conference papers and 9 books. Dr. Kuo is a Fellow of IEEE and SPIE. He is co-Editor-in-Chief for the Journal of Visual Communication and Image Representation, and Editor for the Journal of Information Science and Engineering, LNCS Transactions on Data Hiding and Multimedia Security and the EURASIP Journal of Applied Signal Processing. Dr. Kuo received the National Science Foundation Young Investigator Award (NYI) and Presidential Faculty Fellow (PFF) Award in 1992 and 1993, respectively.