An Exhaustive Review of Automatic Music Transcription Techniques

Survey of music transcription techniques

Gowrishankar B S
Department of Information Science
Vidyavardhaka College of Engineering
Mysore, Karnataka, India
gowrish.vvce@gmail.com

Dr.Nagappa U Bhajantri
Department of Computer Science
Government College of Engineering
Chamarajnagar, Karnataka, India
bhajan3nu@gmail.com

Abstract— **The main objective of this paper is to review the technologies and models used in the Automatic music transcription system. Music Information Retrieval is a key problem in the field of music signal analysis and this can be achieved with the use of music transcription systems. It has proven to be a very difficult issue because of the complex and deliberately overlapped spectral structure of musical harmonies. Generally, the music transcription systems branched as automatic and semi-automatic approaches based on the user interventions needed in the transcription system. Among these we give a close view of the automatic music transcription systems. Different models and techniques were proposed so far in the automatic music transcription systems. However the performance of the systems derived till now not completely matched to the performance of a human expert. In this paper we go through the techniques used previously for the music transcription and discuss the limitations with them. Also, we give some directions for the enhancement of the music transcription system and this can be useful for the researches to develop fully automatic music transcription system.**

Keywords— ***Automatic music transcription (AMT), Pitch, Note detection, Onset and Offset Detection, Informed music transcription, Instrument specific music transcription and Genre specific music transcription.***

I. INTRODUCTION

Automatic music transcription (AMT) is the process of converting a musical signal into a symbolic notation such as either a musical score, sheet or any other equivalent representation [1]. The applications of AMT include automatic retrieval of musical information, musicological analysis, as well as interactive music systems [2]. The AMT problem can be divided into several subtasks, which include: multi-pitch detection, note onset/offset detection, loudness estimation and quantization, instrument recognition, extraction of rhythmic information, and time quantization [3]. While in some of the traditional audio transcription the techniques such as visual modality is employed to assist the transcription of music. Visual modality is nothing but the transcription method in which the transcription is done there based on the information available from the video recording when the player is playing the instrument [4].

Similar to the speech signal processing schemes, the parameter called pitch plays a vital role in the field of musical signal processing also. While playing most of the musical instruments the performers can produce sounds with easily controlled, locally stable fundamental periods. Such type of signal can be well described by a series of frequency components at multiples of a fundamental frequency which results in the percept of a musical note at a clearly defined pitch [5]. Pitch can be identified only in sounds with clear frequency and stable so that they are distinguishable from noises present in the music signal. Pitch is the most important auditory characteristic of music tones together with loudness, duration and timbre.

In polyphonic mixtures consisting of multiple instruments, the interference of simultaneously occurring sounds is likely to limit the recognition performance. The interference can be reduced by first separating the mixture into signals consisting of individual sound sources [6]. A number of different approaches have been proposed for recognizing instruments in polyphonic music. These include extracting acoustic features directly from the mixture signal, sound source separation followed by the classification of each separated signal, signal model-based probabilistic inference, and dictionary-based methods [7]. The perception of pitch has been extensively studied due to its fundamental significance to the human auditory system [8]. For the detection of pitch several Pitch Detection Algorithms (PDA) are proposed and these algorithms can be classified into methods based on its operation in time-domain, frequency domain or joint time-frequency domain [9].

In semi-automatic transcription the user provides some prior information about the transcription process, such as the instrument identities in the target signal or some correct notes for each instrument [10]. Missing feature approaches provide a general framework for recognizing sound sources based on

978-1-5090-4620-1/16/$31.00 ©2016 IEEE

partial information [11]. In contrast with unsupervised techniques, certain applications can also incorporate score information, such as the emerging field of informed source separation [12]. The Particular benefit of this is the information that facilitates the creation of accurate timbre models also enable the identification of note objects of the underlying instruments in the recording [13]. The main weakness of these kinds of methods is that they lack the capacity to adapt to signals that do not comply with the assumption they make about the sources [14].

The term instrumentation refers to the particular combination of instruments employed in a piece of music and the way in which the music is arranged for the instruments [15]. Instrument identification for polyphonic music is a closely related task to blind source separation, where the objective is to separate the source signals from the mixture when a number of mixture signals is given [16]. The musical instrument identification becomes a crucial process for automatic transcription to give better transcription results when spectrum structure of a sound signal is complicated or spectrum deviation is large [17]. Generally the sound of a musical instrument can be said as timbre that makes it attributable difference from another instrument [18]. Hence the instruments can be identified effectively according to the information from frequency domain as well as the instrument can be selected which has the most appropriate timbre to the current expressing emotion in composition and performance [19].

Various techniques are employed to detect the occurrence of notes. Among them, spectrogram factorization methods such as non-negative matrix factorization (NMF) or probabilistic latent component analysis (PLCA) are widely employed to decompose the spectrogram into meaningful elements and their activations [20]. The genre classification approach itself is rather straight forward. The presented block-level features are combined into a single feature vector that is then used for classification [21].

In [22] an approach for polyphonic transcription using joint multiple-F0 estimation, onset and offset detection is proposed where the onset detection done by two novel descriptors which exploit information from the transcription preprocessing steps, multiple-F0 estimation by a pitch set score function which combines several pitch-related features and offset detection by a novel hidden Markov model-based procedure.

II Review on Different Techniques for Automatic Music Transcription

The Automatic music transcription (AMT) methods are generally classified into three classes based on the knowledge available for transcribing the music. These include informed music transcription (IMT), Instrument specific music transcription (ISMT) and Genre specific music transcription (GSMT). Among them IMT performs music transcription by having the knowledge of information from the user and this can be stated as a semi-automatic mechanism otherwise the method will perform music transcription from the musical score called scored informed mechanism. The ISMT methods require the information about the instrument either to be known in advance or can be inferred from the recordings. Similarly the GSMT methods work on the knowledge of genre of the music. These transcription methods are classified as shown in figure 1.

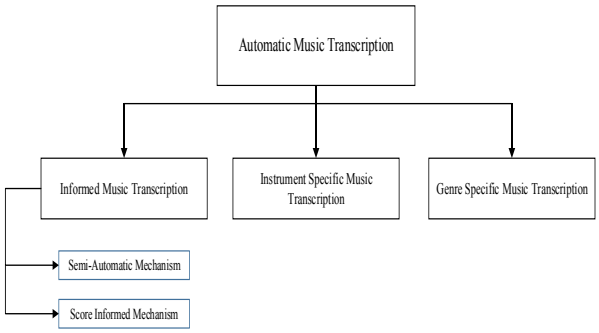

Fig.1.1 Classification of automatic music transcription methods

2.1. Overview on Informed music transcription Methods

Ye Wang and Bingjun Zhang [23] had proposed an Automatic music transcription System (AMT) based on the Human-Computer Interface (HCI) which allows the user to tackle the weakness of the computer on transcribing the music. Their work comprises three strategies like they initially generated the HCI interface to combine the strength of the human with that of computer. They let the users for providing the contextual information such as the instrument type to be played which was the prior information used by them. They achieved this by the instrument modeling approach. This allowed their system to select the corresponding instrument model and simplified the operation. They have provided the feedback to the user by availing the multimodal information from both audio and video streams by means of multimedia fusion in order to enhance the transcription process. They have conducted their research on the violin tutoring scenario by incorporating the settings such as HCI, Instrument modelling and multimedia fusion. In their existing works, they have developed a timbre model that yielded better performance with that of keyboard instruments not for bowed string instruments.

Shlomo Dubnov [24] had proposed a method for the analysis of musical structure which captured local prediction and global repetition properties of audio signals in one information processing framework. They employed constant-Q RTFI to represent the music signals in the form of time-frequency representation and also proposed a method to quash the noise present in the music signal by means of cepstral smoothening as well as pink noise supposition. They adopted new measures called data and model Information rate (IR) for the characterization of those properties within and across the blocks of musical features. They have developed the visualization of the music structures of the two measures mentioned for their easy interpretation. Finally they have used those structural features to detect points of the high music interested. The analysis operated on individual recordings and revealed their internal organization which seemed to be related to the listening experience and reflected some of their compositional design. The analysis described by them was

limited to spectral represented by small cepstral coefficients as well as the computational complexity in the vector IR estimation of the data and model IR.

J. J. Carabias-Orti et al. [25] had proposed an unsupervised process for obtaining the multi-scene adaptive spectral patterns for each MIDI note. They obtained the prior information about the instrument directly from the music file to be analyzed. They have used some clues from the perceptual significance, spectral smoothness, stability, and distance between spectral envelopes at different frames of the same MIDI note in order to determine the spectral envelope of the isolated notes. In this way they defined the adaptive harmonic atoms for the instrument as well as the music scene in which the instrument was played. Their method was accessed adaptive spectral envelopes for mono-timbre polyphonic music files. That harmonic spectral envelopes knowledge was applied to matching pursuits in order to obtain harmonic decompositions adapted to the music scene. These harmonic decompositions were established the input for the note event detection algorithm with promised accuracy and error rate results.

Emmanouil Benetos and Simon Dixon [26] had proposed a method for automatic transcription of music signals based on joint multiple-F0 estimation. They employed constant Q-resonator time-frequency image for the time-frequency representation of the music signal and in the preprocessing stage they suppressed noise based on pink noise assumption. The optimal tuning and inharmonicity parameters were computed in the multiple F0 estimation stage and to select the pitch candidates a salience function was proposed. Then the harmonic envelope of each pitch candidate combination was determined by an overlapping partial treatment procedure. For the optimal pitch candidate combination they proposed one score function for the optimal pitch combination which together combines their spectral and temporal characteristics and suppressed the harmonic errors. They employed Hidden Markov Models (HMM) and Conditional Random Fields (CRF) trained on MIDI data in the post processing stage to enhance the transcription accuracy. They have trained this system on isolated piano sounds from the MAPS database and was tested on classic and jazz recordings from the RWC database, as well as on recordings from a Disklavier piano.

Namgook Cho and C.-C. Jay Kuo [27] had proposed a source-specific dictionary method for efficient music representation and applied it to the separation of music signals which coexist with background noise such as speech or environmental sounds with an impression to calculate a set of rudimentary functions known as atoms which captured music signal characteristics in an efficient manner. They have built a source specific dictionary to capture the inherent music characteristics. There were three steps in the construction of the dictionary. In first step they decomposed the basic components of the music signal into set of source independent atoms called Gabor atoms. Then these atoms were prioritized according to their approximation capability to music signals of interest. Finally they synthesized new atoms from these prioritized Gabor atoms to build a compact dictionary. The number of atoms required for the representation of the music signals using the dictionary proposed by them is minimum compared with the Gabor dictionary which leads to a sparse music representation. The sparse representation of the music signal would produce the system capable of producing the complete representation of music. They applied this technique to the approximation of musical sounds as well as to the music signal separation from single-channel mixtures.

Jean-Louis Durrieu et al. [28] had proposed a source/filter signal model for the mid-level representation which made the pitch content of the signal as well as some available timbre information by keeping much information from the raw data for the broad range of signals. The mid-level representation was formed by the higher level semantics of the music signals such as pitch, timbre or phoneme after normal representation of the music signal which had a tendency to improve the discriminative characteristics as well. An algorithm for the estimation of decomposition parameters for both the single and multiple channel cases was also described. They incorporated specific dictionary element for the decomposition which allowed the representation of unvoiced or noise components in the leading musical source. They used this model within a main melody extraction system and a lead instrument/accompaniment separation system. And both of the two contexts gained topmost outputs at numerous international estimation crusades. They mentioned that their proposed model can also be used with lyrics recognition, Chroma computation or multiple pitch extraction systems.

Akira Maezawa et al. [29] had presented a method to recover fingerings for a given piece of violin music for the recreation of the timbre of a given audio recording of the piece. They achieved this by first analyzing the audio signal to determine the most likely sequence of two-dimensional fingerboard locations which recovers elements of violin fingering relevant to timbre. The sequence was then used as a constraint for finding an ergonomic sequence of finger placements that satisfied both the sequence of notated pitch and the given fingerboard-location sequence. The fingerboard-location sequence estimation was done by based on Hidden Markov Model. Then they estimated the relative strengths of the harmonics from the polyphonic mixture using score-informed source segregation which compensated for discrepancies between observed data and training data through mean normalization. They performed fingering estimation based on cost function model for a sequence of finger placements. The finger board-location estimator performance was evaluated with a polyphonic mixture and with recordings of a violin whose timbre characteristics differed significantly from that of the training data.

Benoit Fuentes et al. [30] had proposed a Harmonic Adaptive Latent Component Analysis (HALCA) for the consideration pitch and spectral envelope variations of notes simultaneously. They modeled each note in the constant Q transform as a weighted sum of fixed narrowband harmonic spectra, spectrally convolved with some impulse that defined the pitch. They estimated all the parameters in the Probabilistic Latent Component Analysis framework by using Expectation Maximization (EM) algorithm. They also introduced interesting priors over the parameters to converge the EM towards meaningful solution. They inferred the onset time,

duration and pitch of each note in the audio file from the estimated parameters on the application of their proposed model to the automatic music transcription. On the other hand in the HALCA model, each polyphonic source had its own time varying spectral envelope, and the noise element was intended such that it could not consider harmonic notes. They evaluated the system on two different databases and produced reasonable results.

Stanisław A. Raczyński et al. [31] had proposed a family of probabilistic symbolic polyphonic pitch models based on three layer Dynamic Bayesian Networks (DBN) with two hidden layers corresponding to the chords and the notes which accounted for both the "horizontal" and the "vertical" pitch structure. Those prototypes were expressed as linear or log-linear interpolations of up to five sub-models, each of which was accountable for modeling a different type of relative. The capability of the prototypes to forecast symbolic pitch data was estimated in terms of cross-entropy and of their proposed "contextual cross-entropy" measure. They evaluated their proposed model on two different experiments: the framework was first evaluated in symbolic experiments where the modeling power quantified in terms of cross-entropy and the contextual cross-entropy and in the acoustic experiments they performed multi-pitch estimation using harmonic NMF model as the acoustic model.

2.2. Overview on Instrument specific music transcription Methods

Anssi Klapuri et al. [32] had proposed a method for extracting the fingering configurations automatically from a recorded guitar performance. They considered 330 different fingering configurations corresponding to the different versions of the major, minor, major 7th, and minor 7th chords played on the guitar fretboard. They formulated the transcription framework as Hidden Markov Model where the hidden states were different fingering configurations and the observed acoustic features were obtained from a multiple fundamental frequency estimator that measured the salience of a range of candidate note pitches within individual time frames. The transitions between consecutive fingerings were constrained by a musical model trained on a database of chord sequences and a heuristic cost function that measured the physical difficulty of moving from one configuration of finger positions to another. They evaluated their proposed method on acoustic, electric and the Spanish guitar. They also showed that their proposed model outperformed a non-guitar-specific reference chord transcription method despite the fact that the number of chords considered there was significantly larger.

Anssi Klapuri and Tuomas Virtanen [33] had proposed a computationally efficient algorithm for the modeling and representation of the time varying musical sounds. They have encoded the individual sounds rather than the statistical properties of various sounds representing a certain class. The acoustic feature vectors were modeled by such anchor points in the feature space which was representing the input data by interpolating between them. This model was generic and used to represent any multi-dimensional data sequence. They applied this model to represent musical instrument sounds in a compact and accurate form. The method was outperformed the conventional vector quantization approach where the acoustic feature data was k-means clustered and the feature vectors were exchanged by the corresponding cluster centroids. Their proposed algorithm achieved computational complexity as a function of the input sequence length T was O (T log T).

Mohammad Akbari and Howard Cheng [34] had proposed a real time piano music transcription method based on computer vision called claVision. The system performed music transcription only from the video performance instead of processing the music audio. They showed that the claVision system had a high accuracy (F1 score over 0.95) and a very low latency (about 7.0 ms) in real-time music transcription, even under different illumination conditions and used for other keyboard musical instruments also. They stages included in the claVision were keyboard registration, illumination normalization, pressed keys detection, and note transcription.

Vipul Arora and Laxmidhar Behera [35] had proposed a method for Clustering and Identification of musical source in polyphonic audio using unsupervised as well as semi-supervised algorithms. The algorithms were based on auditory scene analysis theory, which dealt with how the simultaneous acoustic streams were perceived and segregated. Their proposed clustering of sound streams took place in three levels, viz., pitched event decomposition, group object formation and source streaming. The proposed semi-supervised approach makes use of the source labels of a few pitched events provided by a human annotator.

Alfonso Perez-Carrillo and Marcelo M. Wanderley [36] had proposed a novel indirect acquisition method for the estimation of continuous violin controls from audio-signal analysis based on the training of statistical models with a database contained synchronized streams of audio features and instrumental controls of previously recorded violin performances. The method was able to model the characteristics of a specific violin, which was equipped with a vibration transducer built into its bridge, and once trained, they performed indirect acquisition from analysis of the transducer signal without the need for the sensors any more. Their devised method was able to predict continuous sequences of instrumental controls in a time-aware approach using a combination of Hidden Markov Models (HMM) with observation distributions parameterized as Multivariate Gaussian Mixtures (GM).

Paul H. Peeling et al. [37] had proposed a framework for probabilistic generative models of time–frequency coefficients of audio signals, using a matrix factorization parameterization to jointly model spectral characteristics such as harmonicity and temporal excitations and activations. The models represented the observed data as the superposition of statistically independent sources, and they considered variance-based models used in source separation and intensity-based models for non-negative matrix factorization. They derived a generalized expectation-maximization algorithm for inferring the parameters of the model and then this algorithm was

adapted for the task of polyphonic transcription of music using labeled training data.

Emmanouil Benetos and Simon Dixon [38] had proposed a probabilistic model for multiple-instrument automatic music transcription. The model extended the shift-invariant probabilistic latent component analysis method, which was used for spectrogram factorization. The extensions were support the use of multiple spectral templates per pitch and per instrument source, as well as a time-varying pitch contribution for each source. Thus this method used for multiple-instrument automatic transcription. In addition, the shift-invariant aspect of the method exploited for detecting tuning changes and frequency modulations, as well as for visualizing pitch content. For note tracking and smoothing, pitch-wise hidden Markov models were used. For training, pitch templates from eight orchestral instruments were extracted, covering their complete note range. The transcription system was tested on multiple-instrument polyphonic recordings from the RWC database, a Disklavier data set, and the MIREX 2007 multi-F0 data set.

Nancy Bertin et al. [39] had proposed a model of superimposed Gaussian components including harmonicity while temporal continuity was enforced through an inverse-Gamma Markov chain prior. They exhibited a space-alternating generalized expectation-maximization (SAGE) algorithm to estimate the parameters. Computational time was reduced by initializing the system with an original variant of multiplicative harmonic NMF, which also described as well. The algorithm was then applied to perform polyphonic piano music transcription. Convergence issues were also discussed on a theoretical and experimental point of view. Bayesian NMF with harmonicity and temporal continuity constraints is shown to outperform other standard NMF-based transcription systems, providing a meaningful mid-level representation of the data. However, temporal smoothness had its drawbacks, as far as transients were concerned in particular and detrimental to transcription performance when it was the only constraint used.

Graham Grindlay and Daniel P. W. Ellis [40] had proposed a general probabilistic model for transcribing single-channel music recordings containing multiple polyphonic instrument sources. The system required no prior knowledge of the instruments present in the mixture other than the number, although it can benefit from information about instrument type if available. This approach explicitly modeled the individual instruments and was thereby assign detected notes to their respective sources. They used training instruments to learn a set of linear manifolds in model parameter space which were then used during transcription to constrain the properties of models fit to the target mixture. This lead to a hierarchical mixture-of-subspaces design which made it possible to supply the system with prior knowledge at different levels of abstraction. They evaluated the technique on both recorded and synthesized mixtures containing two, three, four, and five instruments each.

2.3. Overview on Genre specific music transcription Methods

Anssi Klapuri et al. [41] had proposed a segmentation technique to the multi-track audio comprised different genres with main focus on pop and rock music. They have calculated the audio features for frames of audio using Beat tracking algorithms which analyzed single channel and produced list of temporal beat locations. They used a simple algorithm based on four audio features viz., self-distance matrices and homogeneity detection according to the instrumentation or musical function of each track. Their hypothesis was that having access to the multitrack version of a recording enables them to avoid the loss of appropriate information by calculating features from all of the individual source tracks, rather than just the final mixdown as was usually the case in that research area.

Olivier Gillet and Gaël Richard [42] had proposed a method for transcription and separation of drum signals from polyphonic music which combined information from the original music signal and a drum track enhanced version obtained by source separation. They have built a complete and accurate drum transcription system integrating a large set of features were optimally selected by feature selection approaches. They have fused the transcription results obtained on the original music signal and on a drum-enhanced version estimated by source separation to improve the performance. Finally a thorough evaluation of the transcription and separation methods introduced by taking advantage of a large and fully annotated database of drum signals. Their proposed algorithms were of relatively low complexity and run in near real time on standard personal computers.

Amelie Anglade et al. [43] had proposed a new genre classification framework using both low-level signal-based features and high-level harmony features with a first-order logic random forest based on chord transitions and built using the Inductive Logic Programming algorithm TILDE. Three-class genre classification experiments were performed by them on two commonly used datasets using harmony-based classifier, combined with a low-level feature set using support vector machines and multilayer perceptrons. For both datasets when the SVM classifier was used, the improvement over the standard classifier was found to be statistically significant when the highest classification rate is considered. Also it was shown that the combination of high-level harmony features with low-level features lead to genre classification accuracy improvements and was a promising direction for genre classification research.

Vishweshwara Rao and Preeti Rao [44] had proposed a system for voice pitch contour extraction in polyphonic music with a focus on improving pitch accuracy in the presence of strong pitched accompaniment. They used some of the existing methods and the novel methods for the voice pitch tracking. The novel aspects of their proposed system involved the separation of the F0 candidate selection and salience computation into two distinct steps, the joint tracking of two F0 contours by the Dynamic Programming (DP) algorithm with a harmonic-relationship constraint on F0 pairing and the final identification of the voice pitch contour from the dual-F0 tracking output using a voice-feature that exploits the temporal instability (in frequency) of voice harmonics and the existing methods involved the use of a main-lobe matching method for the identification of sinusoids from the short-time magnitude

spectrum of a signal and the TWM error as the F0 candidate salience measure.

Vishweshwara Rao et al. [45] had proposed signal driven window-length adaptation technique to sinusoid identification on real musical signals and investigated the use of signal sparsity for adapting analysis window lengths. They conducted the experiment on two datasets, sampled at 22.05 kHz, each of about 9.5-minutes duration of which the singing voice is present about 70% of the time. The first dataset contains excerpts of polyphonic recordings of nine Western pop songs of singers such as Mariah Carey and Whitney Houston, who are known for using extensive vibrato in their singing. The second dataset contains five Indian classical vocal music recordings. Another result of this work was that the window main-lobe matching sinusoid detection method outperformed an amplitude envelope and phase-based sinusoid detection method.

Matija Marolt [46] had proposed a method for automatic transcription of bell chiming recordings, where the goal was to detect the bells that were played and their onset times. They presented an algorithm that estimate the number of bells in a recording and their approximate spectra. The algorithm used a modified version of the intelligent k-means algorithm, as well as some prior knowledge of church bell acoustics to find clusters of partials with synchronous onsets in the time–frequency representation of a recording. They used cluster centers to initialize non-negative matrix factorization that factorized the time–frequency representation into a set of basis vectors (bell spectra) and their activations. The recording was transcribed by the proposed probabilistic framework that integrated factorization and onset detection data with prior knowledge of bell chiming performance rules. Both parts of the algorithm were evaluated on a set of bell chiming field recordings.

Jia-Min Ren and Jyh-Shing Roger Jang [47] had proposed the use of time-constrained sequential patterns (TSPs) as effective features for music genre classification. They performed an automatic language identification technique to tokenize each music piece into a sequence of hidden Markov model indices. Then TSP mining was applied to discover genre-specific TSPs, followed by the computation of occurrence frequencies of TSPs in each music piece. Finally, support vector machine classifiers were employed based on these occurrence frequencies to perform the classification task. They conducted the experiments on two widely used datasets for music genre classification, GTZAN and ISMIR2004 Genre which showed that the proposed method could discover more discriminative temporal structures and achieved a better recognition accuracy than the unigram and bigram-based statistical approach.

Yizhar Lavner and Dima Ruinskiy [48] had proposed an algorithm for speech/music classification based on Decision Tree Framework. The Algorithm contained learning phase and classification phase. In the learning phase, predefined training data was used for computing various time-domain and frequency-domain features, for speech and music signals separately, and estimating the optimal speech/music thresholds, based on the probability density functions of the features. They employed an automatic procedure for the selection of best features for separation. In the test phase initial classification was performed for each segment of the audio signal by using three-stage sieve-like approach with the application of both Bayesian and rule-based methods. Erroneous rapid alternations in the classification were avoided by a smoothing technique and averaging the decision on each segment with past segment decisions.

Junyong You et al. [49] had proposed a semantic framework for weakly supervised video genre classification as well as for the event analysis with the use of probabilistic models for MPEG video streams. They proposed Hidden Markov Model (HMM) and Naive Bayesian classifier (NBC) based analysis algorithm for the video genre classification. They also built Gaussian Mixture Model (GMM) for the detection of contained events using the same semantic features as well as an event adjustment strategy was proposed according to an analysis on the GMM structure and pre-definition of video events. Subsequently they recognized a special event was based on the detected events by another HMM. They simulated the experiments on video genre classification and event analysis using a large number of video data sets.

III Performance Comparison

The comparison and performance analysis of the different techniques used in Automatic music transcription method is detailed in this section. We compared these different techniques for automatic transcription in terms of the technique used for transcription, achieved accuracy by them and the conclusion of the research. Table 1 below shows the comparative performance analysis of the different methods used in AMT.

Author	Dataset	Classification	Transcription method	Accuracy	Conclusion
Ye Wang and Bingjun Zhang [23]	Violin tutoring scenario	IMT	Human-Computer Interface (HCI), Instrument modelling and multimedia fusion	TP= 83% FP= 28%	Note segmentation in violin and singing sounds is most challenging task
Shlomo Dubnov [24]	Musical examples from classical piano literature	IMT	Music Structure Analysis using statistical properties of the signal	Not Specified	High computational complexity of $O(n^3)$
J. J. Carabias-Orti *et al.* [25]	Classical Piano Midi Page, http://www.piano-midi.de/	IMT	Unsupervised process for multi-scene adaptive MIDI note spectral patterns	40.5%	Provided promising accuracy and error rate results
Emmanouil Benetos and Simon Dixon [26]	Isolated piano chords from MAPS database for training and recordings from RWC database, Disklavier database and MIREX multipitch estimation for testing	IMT	Joint multiple-F0 estimation	60.5%	System performance improved by performing joint multiple-F0 estimation and note tracking
Namgook Cho and C.-C. Jay Kuo [27]	RWC Musical Instrument Sound Database, McGill University Master Samples Library and the University of Iowa Musical Instrument Samples	IMT	Source-specific dictionary method	Not Specified	Obtaining Speech-specific dictionaries is challenging
Jean-Louis Durrieu *et al.* [28]	Polyphonic excerpt from "Three views of a secret"	IMT	Source/filter signal model the mid-level representation	79.9% precision	Mid-level representation of the mixture displayed polyphonic pitch content
Akira Maezawa *et*	Three pieces of classical music, using two significantly different	IMT	Fingerboard-location sequence estimation	79.3% Recognition	Sequential modeling drastically improved the

al. [29]	fingerings (207 notes for three pieces, yielding a total of 414 notes).		using HMM	Accuracy	accuracy
Benoit Fuentes *et al.* [30]	3307 isolated notes from the Iowa database	IMT	Harmonic Adaptive Latent Component Analysis (HALCA)	40.3% Precision	1. Automatic estimation of the threshold of threshold-based onset detection w.r.t. the input signal can be included. 2. Hypothesis of redundancy was not necessary for a TFR factorization technique
Stanisław A. Raczyński *et al.* [31]	RWC Classical Music Database and the Mutopia Project data set	IMT	Probabilistic symbolic polyphonic pitch models based on three layer Dynamic Bayesian Networks (DBN)	83.4% Precision	Interpolation of n-gram models with n>2 can be used in music processing
Anssi Klapuri *et al.* [32]	Training data set consisted of 22 recordings & test data consisted of 14 recordings	ISMT	Fingering configurations extraction from a recorded guitar performance using Hidden Markov Model	PM 88% & PHY 79% with 210 possibilities	Developing music transcription systems for more narrowly targeted contexts lead to significantly improved performance.
Anssi Klapuri and Tuomas Virtanen [33]	Samples McGill University Master Samples collection	ISMT	Algorithm for modeling and representation of time varying musical sounds	SNR = 33 dB for the model order (K) 30	Achieved better modeling accuracy than the k-means clustering method
Mohammad Akbari and Howard Cheng [34]	Sample videos including different slow and fast pieces of music	ISMT	Piano music transcription based on computer vision (claVision)	97.4% key detection Precision	Illumination changes, drastic camera views, covered keys, and expressive aspects of music are the limitations
Vipul Arora and Laxmidhar Behera [35]	VOCAL dataset of the vocal songs from MIR-1k database and instrumental dataset IMIDI of 10 songs	ISMT	Unsupervised as well as semi-supervised algorithms for Clustering and Identification of musical source	For Polyphony order 4, vocal 45.5% & IMIDI 55.2% in unsupervised approach 79.3% & 75.2% in supervised approach	Higher level features could be incorporated

Alfonso Perez-Carrillo and Marcelo M. Wanderley [36]	Violin performances recorded with a measurement system	ISMT	Indirect acquisition method for the estimation of continuous violin controls	23.2% RAE (Relative Absolute Error) and 25,7% RRSE (Root Relative Squared Error)	Restricted to indirect acquisition with recordings
Paul H. Peeling *et al.* [37]	Poliner and Ellis training and test data	ISMT	Generative Spectrogram Factorization Model	67.7%	Generalized expectation-maximization algorithm experience slow convergence to local maxima.
Emmanouil Benetos and Simon Dixon [38]	Set of twelve classic and jazz music excerpts from the RWC database	ISMT	Shift-Invariant Latent Variable Model	62.5%	Sparsity constraints were enforced and note tracking was performed using HMMs.
Nancy Bertin *et al.* [39]	MAPS (MIDI-Aligned Piano Sounds)	ISMT	Non-negative matrix Factorization in Bayesian Network	Precision: 63.4% (synthesized) & 43.3% (real)	Refinement of the temporal prior, which suits for modeling the sustain and decay parts of the note
Graham Grindlay and Daniel P. W. Ellis [40]	Data set used in the MIREX Multiple Fundamental Frequency Estimation and Tracking evaluation task	ISMT	General Probabilistic model	F -measure : 0.73 (synthesized) 0.67 (real)	Many instruments have complex time-varying structures within each note that would seem to be important for recognition.
Anssi Klapuri *et al.* [41]	Annotations and audio files from http://www.eecs.qmul.ac.uk/~stevenh/multi_seg.html	GSMT	Segmentation technique with Beat tracking algorithms	Precision 57%	Greater segmentation accuracy and/or reduced computational complexity by selecting audio features according to instrumentation or musical function of each track.
Olivier Gillet and Gaël Richard [42]	Minus one sequences of the ENST-drums database	GSMT	Drum transcription system integrating feature selection approaches	79.8% Precision	Transcription is easier when the isolated signal is available
Amelie Anglade *et al.* [43]	1. Perez-9-genres Corpus (Training) 2. GTZAN database and ISMIR 2004 (Testing)	GSMT	Chord Transitions with Inductive Logic Programming algorithm (TILDE)	1. GTZAN-41.67% (symbolic training) and 44.67% (synthesize training)	Provide improved results when combined with several other descriptors

				2. ISMIR 2004- 57.49% (symbolic) and 59.28% (synthesize)	
Vishweshwara Rao and Preeti Rao [44]	1. 25 clips from ten songs [50] 2.13 clips from MIR-1k database 3. Excerpts from two North Indian classical vocal performances	GSMT	Voice pitch contour extraction with Dynamic Programming (DP) algorithm & main-lobe matching method	1. PA= 84.1% CA= 88.8% 2. PA = 69.1% CA =74.1% 3. PA =73.9% CA =76.3%	Single-F0 tracking system made pitch tracking errors caused by the output pitch contour switching between tracking the voice and instrument pitches
Vishweshwara Rao *et al.* [45]	Nine Western pop songs and five Indian classical vocal music recordings	GSMT	Signal driven window-length adaptation technique	Precision 88.6% to 100%	Window main-lobe matching sinusoid detection method outperformed amplitude envelope and phase-based sinusoid detection method
Matija Marolt [46]	Field recordings from the digital archive of Slovenian folk music and dances EthnoMuse	GSMT	Modified intelligent k-means algorithm with acoustic features	91% Precision	Onset detection and better handling of beating being two directions
Jia-Min Ren and Jyh-Shing Roger Jang [47]	GTZAN and ISMIR2004 Genre	GSMT	Automatic language identification technique with time-constrained sequential patterns (TSPs)	81.7% for GTZAN and 79.7% for ISMIR2004 Genre	Accuracy is approximately 2% to 6% higher than that of the text-categorization-based approach
Yizhar Lavner and Dima Ruinskiy [48]	Classical, folk, pop, rock, metal, new age and others.	GSMT	Decision Tree Based algorithm	98.6%	Generic training procedure allowed testing any number of additional features.
Junyong You *et al.* [49]	Seven typical video genres : sports, news, ordinary film, cartoon video, music video, Documentary and game	GSMT	Semantic video genre Classification	93.4%	More suitable video features and optimization of probabilistic models needed

	video.				

TABLE 1.COMPARATIVE ANALYSIS OF DIFFERENT TECHNIQUES IN AUTOMATIC MUSIC TRANSCRIPTION (AMT)

From these discussions we can observed that most of the transcription methods employs Hidden Markov Model (HMM) in extracting the fingering information as well as note tracking. On classifying the piano notes Support Vector Machine (SVM) is the most widely used classifier. Similarly most of the methods also uses machine learning method called unsupervised algorithms to make the transcription system to function as a completely automatic. The transcription was also done in a simple way by using the technique called Non negative Matrix Factorization (NMF) which will factorize the time-frequency coefficients of the signal into a codebook of spectral templates and an activation matrix from which the transcription can be inferred. As well as most of the transcription systems proposed in the literature generate the transcription system based on the extraction of fingering configurations, and finger location estimation. Similarly most of the existing works had their main concern on the multiple pitch and note detection. Hence several approaches have to be proposed to extract as more possible musical information from the recording to make the AMT system as fully automatic. Moreover, the finest transcription results will be obtained when machine learning algorithms are mixed with specialist knowledge on the musical information.

IV. FUTURE ENHANCEMENT

Most of the preceding works on automatic transcription systems leverage the problem of multiple pitch estimation (multiple F0) as well as note onset and offset detection. In order to make a fully automatic music transcription system in such a way to provide the result similar to that of a sheet music furthermore issues need to be taken into consideration. The issues can be said as meter induction, rhythm parsing, note spelling, dynamics, articulation (formation of clear and distinct sounds in speech) and typesetting. On seeing the classification based approaches over the first two transcription methods genre-specific music transcription is still under development and steps can be taken to improve such type of transcription as it can be beneficial for different types of music genres. The efficient system for the automatic music transcription should be in such a way that it should be suitable for all types of music with great accuracy and with reduced errors associated with the recognition of the musical instruments, pitch estimation, note tracking, onset and offset detection. The use of different optimization algorithms can also be incorporated into the music transcription system for the optimal selection of audio features from the music files as well as artificial intelligence based techniques which would considerably increase the performance of the system.

V. CONCLUSION

Automatic music transcription is the fastest growing field for the extraction of musical information from the different musical recordings and still investigations on different approaches going on. Even though several works proposed in this area the results of those works are not suitable for different applications and more musical information need to be taken into account for the efficient music transcription system. In this paper, we have reviewed the methods so far proposed by several authors for transcribing the music with different information available for the transcription and classified those works under three conspicuous classifications as informed music transcription, Instrument specific and Genre specific music transcription. Besides, a vast comparison of different transcription methods and the technologies used within each classification also given in this review. The review demonstrates that the music transcription in the field of Instrument Specific transcription come up with latest technologies and extra innovations need to be created for other two classifications of transcription methods. And finally a complete AMT system need to developed with better accuracy and adoptable for all type of music instruments as well as different music genres with minimum available information.

VI. REFERENCES

[1] Fabrizio Argenti, Paolo Nesi and Gianni Pantaleo, "Automatic transcription of polyphonic music based on the constant-Q bispectral analysis", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 6, pp. 1610-1630, 2011.

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff and Anssi Klapuri, "Automatic music transcription: Breaking the Glass Ceiling", In Proceedings of ISMIR, pp. 379-384, 2012.

[3] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen, "Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation", In Proceedings of ISMIR, 2009.

[4] Marco Paleari, Benoit Huet, Antony Schutz and Dirk Slock, "A multimodal approach to music transcription", In Proceedings of 15th IEEE International Conference on Image Processing (ICIP 2008), 2008.

[5] Meinard Muller, Daniel PW Ellis, Anssi Klapuri, Gaël Richard, and Shigeki Sagayama, "Introduction to the special issue on music signal processing", IEEE Journal of Selected Topics in Signal Processing, Vol. 6, No. 5, pp. 1085-1087, 2011.

[6] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff and Anssi Klapuri, "Automatic music transcription: challenges and future directions", Journal of Intelligent Information Systems, Vol. 41, No. 3, pp. 407-434, 2013.

[7] Meinard Müller, Daniel P. W. Ellis, Anssi Klapuri, and Gaël Richard, "Signal processing for music analysis", IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 6, pp. 1088-1110, 2011.

[8] Md. Omar Faruqe, Md. Al-Mehedi Hasan, Shamim Ahmad and Farazul Haque Bhuiyan, "Template music transcription for different types of musical instruments", In Proceedings of 2nd IEEE International Conference on Computer and Automation Engineering (ICCAE), Vol. 5, 2010.

[9] Chetan Pratap Singh and T. Kishore Kumar, "Efficient pitch detection algorithms for pitched musical instrument sounds: A comparative performance evaluation", In Proceedings of IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014.

[10] Holger Kirchhoff, Simon Dixon, and Anssi Klapuri, "Shift-variant non-negative matrix deconvolution for music transcription", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[11] Dimitrios Giannoulis, Anssi Klapuri, and Mark D. Plumbley, "Recognition of harmonic sounds in polyphonic audio using a missing feature approach", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[12] Emmanouil Benetos, Anssi Klapuri, and Simon Dixon, "Score-informed transcription for automatic piano tutoring", In Proceedings of the 20th European IEEE Signal Processing Conference (EUSIPCO), 2012.

[13] Holger Kirchhoff, Sam Dixon, and Anssi Klapuri, "Missing template estimation for user-assisted music transcription", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[14] Kazuki Ochiai, Hirokazu Kameoka, and Shigeki Sagayama, "Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[15] Hiromasa Fujihara, Anssi Klapuri, and Mark D. Plumbley, "Instrumentation-based music similarity using sparse representations", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[16] Dimitrios Giannoulis, Emmanouil Benetos, Anssi Klapuri, and Mark D. Plumbley, "Improving instrument recognition in polyphonic music through system integration", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[17] Yasunori Uchida and Shigeo Wada, "Melody and bass line estimation method using audio feature database", In Proceedings of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2011.

[18] Peter Somerville and Alexandra L. Uitdenbogerd, "Multitimbral musical instrument classification", In Proceedings of IEEE International Symposium on Computer Science and its Applications (CSA'08), 2008.

[19] Jijun Wang, and Chengdong Lin, "Research on instrument timbre characteristics for music emotional expression", In Proceedings of IEEE Symposium on Robotics and Applications (ISRA), 2012.

[20] Dooyong Sung and Kyogu Lee, "Transcribing Frequency Modulated Musical Expressions from Polyphonic Music Using HMM Constrained Shift Invariant PLCA", In Proceedings of Tenth IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014.

[21] Klaus Seyerlehner, Markus Schedl, Tim Pohle and Peter Knees, "Using block-level features for genre classification, tag classification and music similarity estimation", In Proceedings of Submission to Audio Music Similarity and Retrieval Task of MIREX 2010, 2010.

[22] Emmanouil Benetos and Simon Dixon, "Polyphonic music transcription using note onset and offset detection", In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.

[23] Ye Wang and Bingjun Zhang, "Application-specific music transcription for tutoring", IEEE MultiMedia, Vol. 15, No. 3, pp. 70-74, 2008.

[24] Shlomo Dubnov, "Unified view of prediction and repetition structure in audio signals with application to interest point detection", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, pp. 327-337, 2008.

[25] J. J. Carabias-Orti, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes, "Music scene-adaptive harmonic dictionary for unsupervised note-event detection", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 3, pp. 473-486, 2010.

[26] Emmanouil Benetos and Sam Dixon, "Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription", IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 6, pp. 1111-1123, 2011.

[27] Namgook Cho and CC Jay Kuo, "Sparse music representation with source-specific dictionaries and its application to signal separation", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 2, pp. 326-337, 2011.

[28] Jean-Louis Durrieu, Barak David, and Guilhem Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation", IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 6, pp. 1180-1191, 2011.

[29] Akira Maezawa, Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automated violin fingering transcription through analysis of an audio recording", Computer Music Journal, Vol. 36, No. 3, pp. 57-72, 2012.

[30] Benoit Fuentes, Roland Badeau, and Guilhem Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 9, pp. 1854-1866, 2013.

[31] Stanislaw Raczynski, Emmanuel Vincent, and Shigeki Sagayama, "Dynamic Bayesian networks for symbolic polyphonic pitch modeling", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 9, pp. 1830-1840, 2013.

[32] Ana M. Barbancho, Anssi Klapuri, Lorenzo J. Tardón, and Isabel Barbancho, "Automatic transcription of guitar chords and fingering from audio", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 3, pp. 915-921, 2012.

[33] Anssi Klapuri and Tuomas Virtanen, "Representing musical sounds with an interpolating state model", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 3, pp. 613-624, 2010.

[34] Mohammad Akbari and Howard Cheng, "Real-Time Piano Music transcription Based on Computer Vision", IEEE Transactions on Multimedia, Vol. 17, No. 12, pp. 2113-2121, 2015.

[35] Vipul Arora, and Laxmidhar Behera, "Musical source clustering and identification in polyphonic audio", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 6, pp. 1003-1012, 2014.

[36] Alfonso Perez-Carrillo and Marcelo M. Wanderley, "Indirect Acquisition of Violin Instrumental Controls from Audio Signal with Hidden Markov Models", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 5, pp. 932-940, 2015.

[37] Paul H. Peeling, and Simon J. Godsill, "Generative spectrogram factorization models for polyphonic piano transcription", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 3, pp. 519-527, 2010.

[38] Emmanouil Benetos, and Simon Dixon, "A shift-invariant latent variable model for automatic music transcription", Computer Music Journal, Vol. 36, No. 4, pp. 81-94, 2012.

[39] Nancy Bertin, Roland Badeau, and Emmanuel Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 3, pp. 538-549, 2010.

[40] Graham Grindlay, and Daniel PW Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments", IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 6, pp. 1159-1169, 2011.

[41] Steven Hargreaves, Anssi Klapuri, and Mark Sandler, "Structural segmentation of multitrack audio" IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 10, pp. 2637-2647, 2012.

[42] Olivier Gillet and Gaël Richard, "Transcription and separation of drum signals from polyphonic music", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 3, pp. 529-540, 2008.

[43] Amelie Anglade, Emmanouil Benetos, Matthias Mauch and Simon Dixon, "Improving music genre classification using automatically induced harmony rules", Journal of New Music Research, Vol. 39, No. 4, pp. 349-361, 2010.

[44] Vishweshwara Rao and Preeti Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 2145-2154, 2010.

[45] Vishweshwara Rao, Pradeep Gaddipati, and Preeti Rao, "Signal-driven window-length adaptation for sinusoid detection in polyphonic music", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 342-348, 2012.

[46] Matija Marolt, "Automatic transcription of bell chiming recordings", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 3, pp. 844-853, 2012.

[47] Jia-Min Ren and Jyh-Shing Roger Jang, "Discovering time-constrained sequential patterns for music genre classification", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1134-1144, 2012.

[48] Yizhar Lavner and Dima Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation", EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2, 2009.

[49] Junyong You, Guizhong Liu, and Andrew Perkis, "A semantic framework for video genre classification and event analysis", Signal Processing: Image Communication, Vol. 25, No. 4, pp. 287-302, 2010.

[50] Yipeng Li and DeLiang Wang, “Separation of singing voice from music accompaniment for monoaural recordings”, IEEE Transactions on Audio, Speech, Language Process, Vol. 15, No. 4, pp. 1475–1487, 2007.