Carnegie Mellon University

MLSP
MachineLearning For SignalProcessing Group

11-755/18-797 Machine Learning for Signal Processing

Final Poster Presentations

Instructor: Bhiksha Raj

TAs: Sourish Chaudhuri, Sohail Bahmani

Fall 2010

UNSUPERVISED FACE CLUSTERING IN VIDEO

Aaron Jaech, Andy Strat, Chenkai Xie

Our goal is to make the tagging and indexing of faces in videos a one-click process. Faces are detected using Viola-Jones supplemented with a tracking algorithm. Since each face is highly similar to the faces on immediately adjacent frames it is possible to group images from the same individual together using a simple clustering algorithm. Since the number of faces per individual is often very large (15 frames per second) it is possible to handle changes in pose, scene and lighting. After clustering the user is presented with one face from each individual for labeling.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

MULTI-RATE EVENT DETECTION FOR ENERGY-AWARE GREEN DESIGN FACILITIES: TECHNIQUES TO IMPROVE EVENT DETECTION IN NONINTRUSIVE SYSTEMS

Amrita Anand, Kedar Mane, Rathi Munukur

The buzz word in today's world is energy and energy management. Current trends show that the total world consumption of marketed energy will increase by 49 percent from 2007 to 2035. If consumers are given access to detailed data about their day-to-day power consumption they can make informed decisions towards managing their own demand thus also achieving huge cost savings. One step to efficiently conserve energy is to first find the energy consumption by each appliance and then accordingly take steps to decrease the overall energy consumption.

The objective of "Multi-rate Event Detection for Energy-Aware Green Design Facilities" is to detect the occurrence of events (on-off of appliances) in order to estimate the distribution of power consumption. This is an extension of "A Framework for Enabling Energy-Aware Facilities through Minimally-Intrusive Approaches" (thesis of Mario E. Berges) which aims at non-intrusive load monitoring i.e., load monitoring without the use of sub meters. Techniques such as multi-rate sampling are explored to improve accuracy of the detected events, apart from using the Generalized Likelihood Ratio algorithm. Different signatures are created to characterize the possible events by using a variety of different methods. These signature libraries are used to assist the detection of events by performing signature matching. This improves the accuracy of the detected events by eliminating false detections. An improvement of around 50% in the number of correct events detected was obtained while using multiple sampling as opposed to using a single sampling rate approach.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

DE-IDENTIFICATION OF SPEECH

Bharat Kandoi, Bhaskar Adavi, Prashant Malani, Sahil Sapre

Telephonic dialogues are a popular means of information retrieval. However, they are also a hidden channel for the invasion of the end-users' privacy. One way of protecting people from privacy invasion is to de-identify the speaker's voice such that the speech still sounds natural and fully intelligible, yet does not reveal information about the identity of the speaker.

In this project, we propose a method of de-identifying voices using existing tools for voice conversion. This task entails both building a robust speaker ID system capable of identifying speakers' voices and an algorithm to determine the transformation model that de-identifies speakers' voices while retaining intelligibility.

We have designed a speaker ID system that builds models for the speakers' voice using Gaussian mixture modeling. We base the models on three features of the speakers' voice: the spectral parameters (mel-frequency cepstral coefficients), the non-zero fundamental frequencies (f0s) and the clustered MFCCs. We have achieved a reasonable accuracy in detecting unconverted voices with the SID.

To transform voices, we use Festvox, a project that is part of the work at Carnegie Mellon University's speech group. We define the 'best' transformation as the most de-identified yet fully comprehensible transformation of the input utterance. We choose the most de-identified transformation using a k-means clustering algorithm and selective cluster-elimination. Intelligibility is a subjective metric that is given to a high degree of variability when presented to different listeners. We propose a possible method of eliminating unintelligible transformations using software. Experimental results show a high degree of confusability when the speaker ID is presented with the voices transformed using models selected by the clustering algorithm.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

EMOTION RECOGNITION AND SYNTHESIS IN SPEECH

Dan Burrows, Ajay Ghadiyaram, Maxwell Jordan, Amandianeze Nwana, Amber Xu

We have researched different techniques to classify emotions based only on the speech waveform. Our method is not reliant on context or any speech to text technologies. We will rely on Gaussian Mixture Models, Support Vector Machines, decision trees and a unique feature set to characterize human emotions. For the synthesis we created new voices based on emotions through the Festival Speech Synthesis System.

By classifying the emotions of speech we will then have the computer synthesize what was said using the same emotion as the input.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

REHEARSAL AUDIO STREAM SEGMENTATION AND CLUSTERING

Dawen Liang, Guangyu Xia, Mark Harvilla

Rehearsal recordings are valuable for music ensembles to improve their performance. However, since rehearsal recordings are typically hours long and contain disordered, disrupted, and unclassified musical content (mixed with different sections of different music pieces, conductor's talk, all kinds of noise between rehearsal intervals and so on), they're hard to use directly. For example, recordings for a particular song are not easily found. We present a variety of approaches to extract all the musical segments from the disarrayed rehearsal audio stream and then cluster the segments belonging to the same piece of music together. The procedures discussed herein have the potential to be applied in a large scale music database to accomplish the music information retrieval (MIR) tasks.

The whole task is divided into two main sections; the first section covers the initial objective of music and noise discrimination. We tried three different approaches: ADABoost based on eigenmusic in time- and frequency-domain and SVM based on timbral feature. The second section concerns clustering of similar musical segments (including feature representation). We used audio fingerprinting and chroma-based features respectively and different clustering algorithms to accomplish this section. For both sections, we obtained successful results from the experiments.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

Personalization of Head-Related Transfer Functions from a Limited Number of Acoustic Measurements

Griffin Romigh

Virtual 3D audio, or the technology by which sound sources can be presented in 3D space virtually over headphones, has useful applications from military targeting systems to immersive entertainment. Despite the age of the technology and its potential influence, relatively few applications have left the laboratory because of the need for individualized Head-Related Transfer Functions (HRTFs) to achieve high fidelity spatialization. In general, HRTF measurements require dedicated anechoic chambers and speaker arrays capable of placing sound sources with a density of more than 15 degrees in both azimuth and elevation, with typical collection times of greater than an hour. The main goal of this project was to take advantage of the spatial structure inherent within a set of HRTFs in order to provide mechanism by which an entire set of individualized HRTFs could be estimated from a small set of measurement locations. Several methods were investigated for this purpose including missing feature techniques such as latent variable model decomposition and k-Nearest Neighbor averaging, as well as naïve methods such as spherical harmonic decomposition and linear interpolation. The most successful method investigated was a Linear Minimum Mean Square Error (LMMSE) estimation procedure which showed near complete reconstruction of the 232 HRTFs from a subset of less than 50 evenly distributed locations and promising results for as few as 5 locations. More practical measurement schemes like measurements only taken on the horizontal plane proved to be less successful than even distributions; however they still provided benefit over other rapid HRTF personalization techniques such as derivation from anthropometric measurements.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

SOURCE SEPARATION WITH CHARACTER MATCHING

Hongfei Wang, Zhong Zhang

The cocktail party effect describes human's ability to follow one voice source from a mixture of conversations, and often with the addition of background noises. These conversation may happen simultaneously.

However, it is tricky for computers to handle this sort of auditory source separation problem. One relatively successful approach is to use Independent Component Analysis (ICA) to separate voice sources from a mixture of them.

Specifically in this project, FastICA, an efficient and popular algorithm for ICA is implemented. Then we use a correlation filter called the Minimum Average Correlation Energy (MACE) filter to match the separated voices with their characters for identification.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

NON-INTRUSIVE LOAD MONITORING

Hugo Goncalves, Adrian Ocneanu, Manuel Tragut, Aditi Pandya

Non-Intrusive Load Monitoring (NILM) is a technique that determines the load composition of a household through a single point of measurement at the main power feed. The current focus of NILM is the disaggregation of load states by means of supervised learning algorithms that use transition signatures. Recorded signatures of appliances are matched with real-time power readings for this purpose.

In this project, we have attempted to identify an unsupervised learning alternative for load disaggregation, also called Blind Source Separation (BSS). Given an observation dataset **O** composed of a variety of overlapping sources **S**, we try to decompose it into the constituent sources by using a weighted sum W of individual observations. Our goal is to iteratively explain **O** in terms of a 'best-fit' solution from an approximated Weight Matrix **W** and the Source Matrix **S**.

We distinguish between ON/OFF states of the household appliances, and form clusters based on the Real power(R) and Reactive power(Q) readings of the appliances. We have tested various clustering techniques based on agglomeration and genetic algorithms for this purpose. We attach a power consumption weight to each of these appliance clusters, to ultimately reconstruct the individual energy consumption of each appliance. We explored the use of a simple Matching Pursuit algorithm, a COSAMP algorithm to exploit the sparsity of the signals and classical integer programming for this purpose. These algorithms try to minimize the signal reconstruction error, while our objective is to find the real profile of the states of each appliance. We conclude that although they accurately reconstruct the signal, the profiles found are apparently different from the reality. The lack of ground truth of the data makes it difficult to use meaningful metrics for evaluation of our solution.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

DYNAMIC FOREGROUND/BACKGROUND EXTRACTION BASED ON SEGMENTED IMAGE

Huimin Yang, Tianyi Chen, Yulian Xu, Ran Ye

This paper addresses the problem of extracting dynamic foreground regions from a relatively complex environment within a collection of images or a video sequence. By using image segmentation code, we can first convert our traditional pixel-wise image collection into a collection of image with multiple monochrome image segments. Our approach in this study consists of four steps. First of all, we uniformly extract patches from the first frame of segmented image collection. And then, we manual tag the foreground and background within the first segmented image. After this step all the patches in the first frame will form two bags of patches. For one bag, the patches in it can model the features of the foreground. Meanwhile, the patches in the other bag can describe the features of the background. In this case, we call them foreground patches and background patches respectively. Third, for an incoming frame, we perform the segmentation and then extract the patches. For both the patches from the new frame and the previous known patches, in order to reduce the dimension, we perform PCA and LDA. Then use KNN and KDE to find the nearest patch bag for the incoming patches. At this point, we can differentiate the foreground and background for the new image frame. Finally, we perform a bidirectional consistency check between the patches we already get and the patches from incoming image frame make the model adapt to new incoming image frames. In this report, we practice a novel, clear and easy way to extract dynamic foreground from the complex background.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

SUPPORT VECTOR CORRELATION FILTERS

Jonathon Smereka and Andres Rodriguez

For pattern recognition including automatic target recognition (ATR) two distinct approaches prevail. One is based on detecting regions of interest, extracting features and classifying the feature vectors using classification schemes such as support vector machines (SVMs). The second approach involves the use of correlation filters (CF) that are designed to yield sharp correlation peaks for desired targets while exhibiting low response to clutter (noise and background). Both approaches have benefits and drawbacks. SVMs can offer good generalization, but require the segmentation of the target and it is not known a priori which features should be used. Attractive properties of correlation filters include shift-invariance (i.e., the targets do not need to be centered or segmented) and graceful degradation (i.e., targets can be partially occluded). However, correlation filters depend significantly on the training sets used, making their generalization capabilities not as good as those of SVMs. For our class project we developed and evaluated a new framework called the Support Vector Correlation Filter (SVCF) that combines SVM and CF approaches in order to obtain the best features of these two distinct approaches. Our results show improve performance of SVCF to both SVMs and CFs.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

ROBUST IMAGE LOGO REMOVAL

Miriam Cha, Pooya Khorrami, and Matt Wagner

The latest trend in advertising is to display video logos in a context devoid of ads such as movies or television shows where the product is featured in story line of the show. Such embedded marketing strategy often causes the problem of decreasing viewing pleasure for audiences. This paper presents a method for automatically detecting and removing logos from digital images given the logo of interest. The logo is localized using Speed Up Robust Features (SURF) technique and the logo is completely removed after further refinement by active contour based on a Mumford-Shah segmentation method. In the removal of the logo we use exemplar-based inpainting to fill in the marked location in a visually plausible way. The combination of these three methods allows robust detection and removal of the target region without manual selection.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

MUSIC INFORMATION RETRIEVAL

Philippe De Wagter, Yuqian Zhao, Quan Chen

This project relies on computer vision techniques to build a practical music retrieval system.

Our approach tackles traditional music identification as a corrupted sub-image retrieval problem from the 2-D spectrogram representation of the original songs. More specifically, a query snippet spectrogram is matched against our database using its descriptor representation.

We utilize a novel pairwise boosting method to learn a robust and compact subset of Viola-Jones filters. This subset of filters captures distinctive information of each spectrogram while being resistant to distortion. By applying these filters, we can transform a spectrogram into a low-dimension binary representation.

During the query phase, we search for all the candidates that locally match the descriptors of the query snippet. Then, we select the most likely candidate using an occlusion model based on a 2-state HMM.

In order to test the performance of our music retrieval system, we rely on a case study: the Reggae music genre of the iTunes Store. The recognition is both fast and accurate, even with noisy background and a short song snipped as query.

We believe that this project can have a significant impact on the way people organize and research music. Indeed, this system tags songs automatically, adding music meta data such as album, artist, genre, and art cover. Furthermore, this project also opens up new avenues for live music search and identification.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

TALK-ALONG KARAOKE

Takshak Desai, Swetha Chigurupati, Anish Menon, Jason Andersen

Karaoke systems usually prompt users to "sing" out the lyrics of selected songs. The idea of a Talk-Along Karaoke system is to have the user "talk" out the lyrics of a song and output a singing version of the recorded speech. This essentially means that the speech waveform must be modified to resemble the corresponding sound waveform. Speech modification for such an application requires the pitch and duration to be modified in order to match the pitch and duration of the song. Several methods can achieve such modification. One simple technique to perform this is Pitch Synchronous Overlap Add (PSOLA). Modifying the pitch of the speech signal to approximate the note frequency in a song produces a satisfactory representation of "singing speech". Further effects can be added by incorporating techniques like vibrato while modifying the pitch. This project opens up several interesting possibilities in speech transformation and speech-to-song conversion.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

SONG RETRIEVAL SYSTEM USING HIDDEN MARKOV MODELS

Xie Yiting, Akshay Chandrashekaran, Nidhi Kohli, Ge Yang, Abhishek Jain, Anoop Ramakrishna, Tejas Chopra, R. Swaminathan

This project proposes a novel method to retrieve the song corresponding to a recording of a snippet of the same. The method used converts the song into a string of music phones using HMMs. The sequence of these phones are then stored. A similar procedure is performed upon the snippets. The snippet sequence and song sequence are then compared using a sliding window method with the Levenshtein edit distance as the score metric. This effectively reduces the problem to a string search problem. The usage of HMMs adds robustness against noise. The application for these projects include user based content search, song retrieval systems like Shazam©" etc.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

DAMAGE RECOGNITION FOR STRUCTURAL HEALTH MONITORING

Yujie Ying and Joel Harley

In the field structural health monitoring, researchers focus on the design of systems and techniques capable of detecting damage in structures such as buildings, bridges, airplanes, ships, and pipes. However, it is difficult to develop robust detection schemes that are invariant to environmental and operational conditions. In this project, we discuss how signal processing and machine learning techniques can be used to develop such a robust system.

We focus on damage detection through the measurement and analysis of ultrasonic guided waves produced by piezoelectric transducers. Ultrasonic guided waves travel through the thickness of the structure and are sensitive to structural changes caused by cracks, corrosion, or other forms of damage. However, these waves are also sensitive to benign effects, such as changes in temperature or air pressure. As a result, most traditional detection techniques fail under variable environmental conditions. Our system takes advantage of an advanced signal processing tool known as the Mellin transform to extract robust, scale-invariant features. This is useful since uniform variations in the wave velocity translate into a time-scaling effect.

From experimental data of a pressurized pipe with varying pressure, we extract a set of 212 different data features and use them to implement three different classification algorithms for detecting and localizing damage: ADABoost, support vector machines, and a combination of the two. The third algorithm shows the best overall performance in terms of accuracy, ranging from 81% to 100% in damage detection tests and 70% to 100% in damage localization tests. We also demonstrate the effectiveness of the Mellin transform for robust feature extraction by implementing ADABoost as a feature selection tool.

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):