MODELING AUDIO AND VISUAL CUES FOR REAL-WORLD EVENT DETECTION

BY

XIAODAN ZHUANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor Mark A. Hasegawa-Johnson, Chair
Professor Thomas S. Huang
Professor Stephen E. Levinson
Associate Professor J. Stephen Downie

ABSTRACT

Audio-visual event detection aims to identify semantically defined events that reveal human activities. Most previous literature focused on restricted highlight events, and depended on highly ad-hoc detectors for these events. This research emphasizes generalizable robust modeling of single-microphone audio cues and/or single-camera visual cues for the detection of real-world events, requiring no expensive annotation other than the known timestamps of the training events.

To model the audio cues for event detection, we leverage statistical models proven effective in speech recognition. First, a tandem connectionist-HMM approach combines the sequence modeling capabilities of the hidden Markov model (HMM) with the context-dependent discriminative capabilities of an artificial neural network. Second, an SVM-GMM-supervector approach uses noise-robust kernels to approximate the KL divergence between feature distributions in different audio segments. The proposed methods outperform our top-ranked HMM-based acoustic event detection system in the CLEAR 2007 Evaluation, which detects twelve general meeting room events such as keyboard typing, cough and chair moving.

To model the visual cues, we propose the Gaussianized vector representation, constructed by adapting a set of Gaussian mixtures according to the set of patch-based descriptors in an image or video clip, regularized by the global Gaussian mixture model. The innovative visual modeling approach establishes unsupervised correspondence between local descriptors in different images or video clips, and achieves outstanding performance in a video event categoriza-

tion task on ten LSCOM-defined events in the Trecvid broadcast news data, such as exiting car, running and people marching. Following an efficient branch-and-bound search scheme, we further propose an object localization approach for the Gaussianized vector representation.

We jointly model audio and visual cues for improved event detection using multi-stream HMMs and coupled HMMs (CHMM). Spatial pyramid histograms based on the optical flow are proposed as a generalizable visual representation that does not require training on labeled video data. In a multimedia meeting room non-speech event detection task, the proposed methods outperform previously reported systems leveraging ad-hoc visual object detectors and sound localization information obtained from multiple microphones.

To my parents and my wife, for their love and support.

ACKNOWLEDGMENTS

My graduate study at the Beckman Institute is truly a rewarding experience that has exposed me to various research problems reaching beyond the scope of this dissertation. A lot of people have helped make such experiences possible. I'm thankful to every member of my committee for their advice in my dissertation work. I'm particularly grateful to my adviser, Prof. Mark Hasegawa-Johnson, for the guidence and encouragement throughout my study at Illinois and to Prof. Thomas Huang for co-advising much of the work I've done. Many people in the Statistical Speech Technology Group, the Image Formation and Processing Group, the Prosody-Automatic Speech Recognition meetings and Yale Haskins Laboratory have inspired and helped me in various ways. I truly feel honored and fortunate to have had the opportunity to work closely with all of you.

My graduate experiences would not be complete without the helpful interaction with my colleagues at the speech technology groups at Microsoft Research Asia, IBM T. J. Watson and Vlingo, during and after my pleasant internships.

It is impossible to imagine life without the support of my parents Yan Zhuang and Kemin Cheng, and my newly wedded wife and longtime best friend, Zhenghan Qi. This work is dedicated to my deceased father, who I feel is still with the family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Audio and visual information is of significant importance to human perception as well as machine intelligence. Detecting real world events based on such information finds various applications, including security surveillance [1], human computer interaction, video annotation and multimedia retrieval [2]. In aging societies, assistance to dependent people, particularly elderly people, staying in an unsupervised environment also requires such capability [3]. Varying situations determine the availability of information in either or both of the two modalities. While other sensory data has also been studied, this dissertation focuses on modeling audio and visual cues for real-world event detection.

Real-word events present a significant challenge for machine intelligence. Even with predefined categories, the cues can be subtle. Moreover, it is not always possible to pinpoint clear indicators for different event categories. For example, a video clip of a "car exiting" event might not have a complete profile view of the vehicle. A "keyboard typing" event might have low-energy audio footprint and barely visible visual cues from a bird-eye camera.

We study real-world event detection through a set of related problems. First, short-term acoustic event detection aims to reveal the time and category of event occurences in a relatively long audio stream. Second, video event detection provides the event category for video shots, whose boundaries can be obtained by a well-studied task called shot boundary detection. Third, audio-visual event detection performs the same task as acoustic event detection, but with access to

observations in both modalities.

1.2 Background

1.2.1 Acoustic event classification and detection

There is growing research interest in audio/acoustic event detection (AED). Although speech is the most informative auditory information source, other kinds of sounds may also carry useful information, such as in surveillance systems [4]. In a meeting room environment, a rich variety of acoustic events, either produced by the human body or by objects handled by humans, reflect various human activities. Detection or classification of acoustic events may help to detect and describe the human and social activity in the meeting room. Examples include clapping or laughter inside a speech discourse, a strong yawn in the middle of a lecture, a chair moving or door noise when the meeting has just started [5]. Detection of the nonspeech sounds also helps improve speech recognition performance [6, 7].

Much research in audio content analysis has typically addressed the problem of segregating a few audio sources [8, 9] or segmenting an audio stream into a small number of acoustically compact categories or scenes [10, 11]. Acoustic event detection (AED), a subtask of audio content analysis, aims to detect specified acoustic events such as gunshots [4], explosions [12, 13], speech/music transitions [10], cough events [14], and audience cheering at a sports event [15]. Such information is very helpful in applications such as surveillance, multimedia information retrieval and intelligent conference rooms.

Acoustic events sometimes intervene between speech or overlap with background speech. Without explicit processing of such phenomena, it is possible to implicitly deal with background speech as noise included in the event observations [16]. Assuming limited overlapping, we can perform voice activity detection first and then identify acoustic events in the non-speech segments. Acoustic

event detection could also be performed tightly coupled with the decoding process of speech recognition. For example, the non-speech events can be included in the language model used in Viterbi decoding, similar to the way silence and noise are modeled in large vocabulary speech recognition. Another possibility is to treat the acoustic event sequence (padded with silence and background) and speech as two separate processes which are decoded simultaneously: the observed audio waveform is the summation of the two processes. Though this approach has not been studied for this particular problem, it is successfully used in multi-talker speech recognition where speech from mutliple speakers overlaps in time [17].

1.2.2 Video event detection and object localization

Video based event recognition is an extremely challenging task due to all kinds of within-event variations, such as unconstrained motions, cluttered backgrounds, object occlusions, environmental illuminations and geometric deformations of objects. While there exists work attempting to detect unusual or abnormal events [18, 19] in video clips, the research on event recognition in the real world is still in its preliminary stage.

Many statistical models, e.g., hidden Markov model (HMM) [20], and coupled HMM [21] were proposed to capture the spatial and temporal correlations of video events, and then the learned models are utilized for pre-defined video event classification or abnormal event detection. On the other hand, appearance-based techniques were also widely used for video event detection and classification. Ke et al. [22] applied the boosting procedure for choosing the volumetric features based on optical flow representations. Niebles et al. [23] adopted the spatio-temporal interest points [24] to extract the features, and other works [24] extracted volumetric features from salient regions. There also exist works that used bag-of-words model to tackle the problem of object/event recognition [25, 26]. In addition, Bagdanov et al. [27] adopted bag-of-SIFTs to detect and

recognize object appearances in videos. Xu and Chang [28] proposed to encode a video clip as a bag of orderless descriptors obtained from mid-level semantic concept classifiers extracted from all of the constituent frames, along with the global features extracted within each video frame.

One problem related to video event detection is video shot boundary detection. A video shot is a fundamental unit for structured video. Video shot boundary detection is a non-trivial task, particularly given that the boundaries could be either gradual or clear cut. The task has been extensively studied in Trecvid 2001-2007, as detailed in [29]. Many video event detection works, including the experiments performed in this dissertation, start with given shot boundaries.

The object localization task involves finding the bounding boxes of an object within an image, thereby leveraging spatially localized visual cues in an image. Different from the image categorization problem that aims to assign one label for the image, object localization needs to evaluate many possible bounding boxes and identify one or several of them that contain the target objects. A natural way to carry out localization is the sliding window approach [30]. However, an exhaustive search in an $n \times n$ image needs to evaluate $O(n^4)$ candidate bounding boxes. Heuristics about possible bounding box locations, widths and heights, or local optimization methods are often used to reduce the search space. The bounding box search speed can be further improved by coarse-to-fine search schemes.

1.2.3 Audio-visual fusion

It has been shown that in many applications with both audio and visual information, modeling of the two modalities improves performance compared with either modality. Chu and Huang [31] and Hasegawa-Johnson et al. [32] both used the coupled hidden Markov model for audio-visual speech recognition. Hasegawa-Johnson et al. [32] also explored using a more general dynamic Bayesian network to better model the coupling between audio and vision, based

on articulatory phonology. Sadlier and O'Connor [33] studied detection of field sports scoring events, using a support vector machine with various audio-visual features informative across various sports types. Canton-Ferrer et al. [34] and Butko et al. [35] both performed audio-visual event detection using not only audio information, but also output from well trained specialized visual object trackers, and fused the two modalities at score level and at feature level respectively.

One way to classify audio-visual integration strategies [36] views them as three categories. The first is early integration, which extracts feature vectors from both audio and visual observations and concatenates them into one feature vector sequence for use in one model with the same structure as for one modality. The second is late integration, which extracts feature vector sets separately and uses two sets of models generating reliability weights to be combined across modalities. This is also referred to as decision fusion or separate identification. The third is intermediate integration, e.g., product hidden Markov model or coupled hidden Markov model.

Besides audio-visual integration, the availability of audio-visual data also enables multi-view learning, which leverages the relation between the different modalities to improve the learning. Canonical correlation analysis (CCA) is an unsupervised feature transform learning method that finds a subspace where the audio and visual cues achieve maximum correlation. One modality can be viewed as "soft labels" for the other, when finding the optimal projection onto the CCA subspace. This has been shown to improve speaker recognition and clustering, even when the visual cues are not available at testing, in [37] and [38] respectively. When both audio and visual cues are available at testing, we can apply CCA for both modalities to obtain two versions of the projected feature vectors. It is pointed out by [39] that these projected vectors can be further decomposed into uncorrelated elements, so that an early integration strategy can be applied to correlated corresponding audio-visual elements and a late integration strategy to the uncorrelated elements.

1.2.4 Audio-visual pattern recognition in general and realistic data

Real-word audio and visual data present much more variation than restricted lab data. Many times, even for the same task, approaches that work on restricted lab data are not necessarily suitable for realistic data. One example is from the acoustic event detection literature. While most of the work in event detection focuses on a few highlight events, the 2006 and 2007 AED Evaluations sponsored by the project "Classification of Events, Activities and Relationships (CLEAR)" [5, 1] were mainly performed on a continuous audio database recorded in real seminars [40]. Systems attempted to identify both the temporal boundaries and labels of twelve acoustic events (door slam, paper wrapping, foot steps, knocking, chair moving, phone ringing, spooncup jingle, key jingle, keyboard typing, applause, cough, and laughter). Instead of being exclusively highlight events, many of the events in CLEAR evaluations were either subtle (low SNR, e.g. steps, paper wrapping, keyboard typing), or/and overlapping with speech, making the task particularly challenging. The real environment factor added to the variation of the events as well as the difficulty of segmenting the audio-visual input stream. In the 2006 CLEAR AED Evaluation, the participants delivered superb performance on acoustic event detection on clean audio with performed events, while the same teams struggled with realistic seminar data [41].

In 2007 CLEAR AED Evaluation, with only audio information available to the systems, although different system architectures and feature sets have been explored [5, 1], even the top rated AED system, which was developed by the author of this dissertation together with other members of our UIUC team, left much space for improvement [42]. The evaluations highlighted the challenges in the detection of a large set of general acoustic events in a real world environment.

With the significant challenge from audio-only event detection, the research community has explored leveraging additional visual information to improve AED performance [43, 12, 44]. Leveraging additional visual cues for audio

signal analysis has also been explored for other applications, such as speech recognition [45] and person identification [46]. In particular, the multi-stream HMM and the couple HMM (CHMM) are two effective models for audio-visual fusion.

Video event detection presents a major challenge, when the concerned data is from real broadcast news video. Video event detection in this genre differs from previous studies of more constrained video in various ways. First, the camera is often in motion, introducing blur and movement of the views. Second, the same event category may present itself in dramatically different visual content or layout. Third, it is hard to pinpoint particular problem-specific audio-visual characteristics in order to identify different categories. One way to deal with the realistic video data is to leverage lower-level semantic concepts, with the assumption that such concepts well summarize the visual cues and enable convenient comparison between different video clips [47].

1.3 Contributions

This dissertation tackles the problem of identifying both timestamps and types of real world events, providing a comprehensive description of the real world audio and/or visual stream. Moreover, this research emphasizes robust and generalizable modeling of audio cues and video cues, either separately or jointly, with no use of highly ad-hoc detectors trained using separate labeled data. The proposed framework for audio-visual event detection takes advantage of known timestamps of the training events and requires no expensive location annotation of the visual cues.

Statistical models proven effective in the speech recognition literature are used for audio cue modeling. First, a tandem connectionist-HMM approach combines the sequence modeling capabilities of the HMM with the high-accuracy context-dependent discriminative capabilities of an artificial neural network trained us-

ing the minimum cross entropy criterion. Second, an SVM-GMM-supervector approach uses noise-adaptive kernels approximating the KL divergence between feature distributions in different audio segments. These methods show that a better temporal context modeling improves AED based on HMMs, and modeling the audio segment via one distribution for all frame-based vectors provides useful complimentary information for the task.

In this dissertation, visual cue modeling uses an innovative Gaussianized vector representation for images and video clips, applied in object categorization and localization algorithms. The Gaussianized vector representation summarizes an image or a video clip with the distribution of patch-based descriptors, approximated by a Gaussian mixture model. This representation establishes unsupervised correspondence between different images through the set of Gaussian components adapted from a global set of Gaussian components according to the maximum a posteriori (MAP) criterion. A linear kernel based on this representation approximates the KL divergence between patch descriptor distributions from different images or video clips, and can be used not only for categorization but also for localization in an efficient branch-and-bound search scheme. These methods show that it is possible to effectively model real world image and video data without developing supervised lower level semantic concept detectors, and achieve state-of-the-art performances for broadcast news video event detection.

I also study improving the detection and classification of the events using cues from both audio and visual modalities requiring only labels available for audio training. Optical flow based spatial pyramid histograms are used as a generalizable visual representation that does not require training on labeled video data. Multi-stream HMMs or coupled HMMs (CHMM) are used for audio-visual joint modeling. To allow the flexibility of audio-visual state asynchrony, I explore effective CHMM training via HMM state-space mapping, parameter tying and different initialization schemes. The proposed methods successfully improve acoustic event classification and detection on a multimedia meeting room dataset containing eleven types of general non-speech events without using extra data

resource other than the video stream accompanying the audio observations. The audio-visual event classification and detection system outperforms a previously reported system engaging multiple supervisedly-trained visual object detectors and location estimators based on microphone array signal processing.

The rest of this dissertation is organized as follows. Chapter 2 presents the work in acoustic event detection, which has been published in [48, 49]. Chapter 3 details the Gaussianized vector representation and its applications in video event detection and visual object localization, most of which have been published in [50, 51]. Chapter 4 presents the work on improving acoustic event detection using general visual cues, to be published in [52]. The dissertation concludes with discussion and conclusion in Chapter 5.

CHAPTER 2

AUDIO MODELING FOR ACOUSTIC EVENT DETECTION

Acoustic event detection (AED) in realistic data differs from classification of isolated events in a silent environment, calling for different statistical models. While SVMs were shown to be optimal for the latter [53], the former saw most leading CLEAR AED Evaluation participants using dynamic Bayesian networks [5, 1]. In particular, our top-rated AED system in CLEAR Evaluation 2007 used a set of left-to-right hidden Markov models (HMMs), each for one event. HMMs owe their success to the Viterbi algorithm [54], which allows them to compute simultaneously optimal segmentation and classification of the audio stream. Noise in individual frames is alleviated by the HMM's learned hysteresis, i.e., its typical learned preference for self-transitions rather than non-self-transitions in the hidden finite state machine.

To take advantage of this proven approach, we leverage a framework in which HMMs are used to achieve audio segmentation and event classification simultaneously. To alleviate HMM's problem that each hidden state models only local observations, we propose to use the tandem connectionist-HMM approach [55], where an artificial neural network (ANN) outputs posterior probabilities of event types based on very-long-duration, temporally overlapping observation vectors, leading to better contextual modeling and event discrimination. To further refine the event detection result, we propose using Gaussian mixture model (GMM) supervectors [56] to abstract the noisy features in the training audio segments and the hypothesized segments obtained by the tandem model. An SVM with kernels built on these GMM supervectors, namely the SVM-GMM-supervector classifier, is used to replace the labels proposed by the first-pass tandem model,

when such replacement is desirable according to held-out development data.

We perform acoustic event detection experiments on the same setup as the AED evaluation in CLEAR 2007. It is demonstrated that the tandem connectionist-HMM approach and the SVM-GMM-supervector approach for refining the result both contribute to performance improvement, and the proposed system significantly outperforms our submission system in the CLEAR 2007 AED Evaluation, which was the best ranked in the challenging AED task, outperforming other participating systems by 50% relative in detection accuracy. We also show that the acoustic event detection methods, in particular the HMM-based AED system and the complimentary SVM-GMM supervector rescoring can be effectively applied in a human falling detection system using a single microphone as the sensor.

2.1 Segmentation and Classification: HMM-Based System

Audio event detection requires both segmentation of the audio stream, and classification of the segments. Following our experience in the AED task of CLEAR 2007, we perform simultaneous segmentation and classification using a Bayesian inference procedure similar to state-of-the-art methods for continuous speech recognition [57, 58].

We formulate the goal of acoustic event detection as follows: to find the event sequence that maximizes the posterior probability of the event sequence $W = (w_1, w_2, ..., w_M)$, given the observations $O = (o_1, o_2, ..., o_T)$:

$$\hat{W} = \arg\max_{W} P(W|O) = \arg\max_{W} P(O|W)P(W) \tag{2.1}$$

The acoustic model $P(O|W)$ is one HMM for each acoustic event, with three emitting states connected using left-to-right and self-loop transitions. For background silence and speech, we use a HMM with additional transitions between

the first and third emitting states, to account for the increased internal complexity. The structure of the HMMs can model some of the non-stationarity of acoustic events. The observation distributions of the states are incrementally-trained Gaussian mixtures. The HMM for an acoustic event is trained to represent all training data segments carrying the same event label.

In order to capture short-term soft constraints on the sequence of event labels, the probability of an event label sequence $(w_1, \ldots, w_m)$ is represented by a bigram language model:

$$P(w_1 w_2 \cdots w_m) = P(w_1) \prod_{i=2}^{m} P(w_i | w_{i-1}). \tag{2.2}$$

A bigram "language model" in AED favors recognized acoustic event sequences with sequence statistics similar to those in the training data. Although the language model here does not have the same linguistic implications as in speech recognition, it does improve performance. One of the possible reasons is that it suppresses long sequences of identical event labels, thus forcing the HMMs to better learn the internal temporal structure of the acoustic events.

2.2 Acoustic Context: Tandem Connectionist-HMM Approach

The tandem connectionist-HMM approach is composed of two major components, as shown in Figure 2.1: an artificial neural network (ANN) that observes feature vectors in a context window and outputs posteriors of different acoustic event types, and an HMM component that uses a transformed and normalized version of the output of the ANN, optionally together with the original features, as input features. This approach has been shown to improve HMM-based automatic speech recognition [55]. We use the same framework to boost performance of acoustic event detection by drawing evidence from a wider time context window and emphasizing the difference between confusable feature vectors

across acoustic events by discriminative training.

Two lessons from its application in speech recognition are particularly relevent for using the approach in AED. First, the ANN improves recognition performance in high noise conditions [59, 60]. The AED task is characterized by low SNR, in particular with backgrounds that have high variation. Second, the ANN benefits speech recognition when context independent models are used [60]. To limit the complexity of the ANN, it is used to distinguish only between different context-independent models. As pointed out by [60], if the generative (HMM) part of the tandem system leverages context-dependent models, the ANN may end up counterproductive by increasing overlap and confusion between different context-dependent models that correspond to the same context-independent model. Consistent with the above findings, we have used the tandem architecture successfully for speech recognition from tract variables in an architecture based on articulatory phonology [61, 62]. In this work, we use the HMMs to model different acoustic events that are indeed context-independent.

Consecutive frames within the context window are concatenated to form the input X to the ANN, each dimension corresponding to one input node. The number of output nodes equals the number of acoustic event types. The ANN is discriminatively trained, by back-propagating a minimum cross entropy criterion, to targets that set the output node corresponding to the ground truth event as one and all other output nodes as zero. During testing, for each context window, the ANN presents estimated posterior probabilities across all acoustic events. All context windows centered at every consecutive feature frame are evaluated in the same way, resulting in a sequence of posterior probability vectors.

With these posterior probabilities, we could perform classification using two different approaches. The first approach just directly uses the ANN output: either to assign to each frame its maximum a posteriori event label, or to generate probabilities that will be smoothed by a Viterbi decoder. However, experiments in automatic speech recognition suggest that better results may be obtained by transforming the posteriors into a pseudo-observation, which is then used as the

input to a Gaussian mixture HMM.

In order for ANN posterior probability vectors to be better modeled by the Gaussian mixture likelihood model of an HMM, three transformation are applied as suggested by previous work in tandem speech recognition [55]. First, we take the log of each posterior probability to reduce the skewedness of the distributions. Second, principal component analysis (PCA) is applied on the log probabilities to decorrelate the HMM input, so that we may use diagonal covariance matrices in the Gaussian mixture models. Third, mean and variance normalization is applied on each of the decorrelated dimensions, within each audio session.

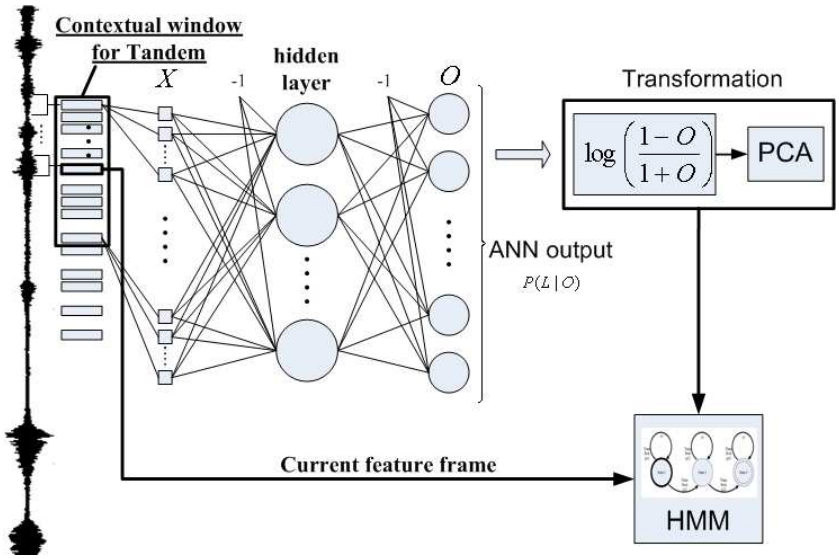

Figure 2.1: Classification using a tandem model (ANN+HMM).

2.3 Complimentary Rescoring: SVM-GMM-Supervectors for Audio Segments

Researchers in automatic speaker identification have recently developed a set of algorithms that boost classification performance by feeding the likelihood out-

put of a generative model (usually an adapted Gaussian mixture model) to the input of a discriminative classifier (usually an SVM) [56]. The SVM-GMM-supervector approach is not practical as a first-pass segmenter for AED, because it requires some type of hypothesized segment boundaries. Given the boundaries chosen by a connectionist-HMM first-pass system, the SVM-GMM is able to efficiently compute confidence scores for each of the proposed segment labels. The SVM-GMM is robust to background noise owing to the parametric modeling of frame-level feature distribution. It discriminates between the candidate classes, with scores normalized by adaptation of a common multi-mode Gaussian mixture distribution.

We refer to the audio observation between two adjacent boundaries as an *audio segment*. The SVM-GMM-supervector approach approximates the joint distribution of all feature vectors in *each audio segment* with a GMM, from which a GMM supervector is constructed as a summary of the segment. The pairwise Euclidean distances between these supervectors characterize the difference between the audio segments. Kernels derived from these distances are used in an SVM for classification.

Figure 2.2 demonstrates that each audio segment is represented as an ensemble of frame-based feature vectors, whose distribution is approximated by a set of Gaussians adapted from the global Gaussian mixtures, or the universal background model.

2.3.1 Universal background model and segment-specific Gaussian mixture models

We estimate a GMM for the distribution of all feature vectors in each audio segment. Instead of separately estimating a GMM for each audio segment, we estimate a GMM for each audio segment by adapting, to each audio segment, the parameters of a universal background model (UBM): a GMM that has been previously trained to represent all types of audio. Adaptive training creates a

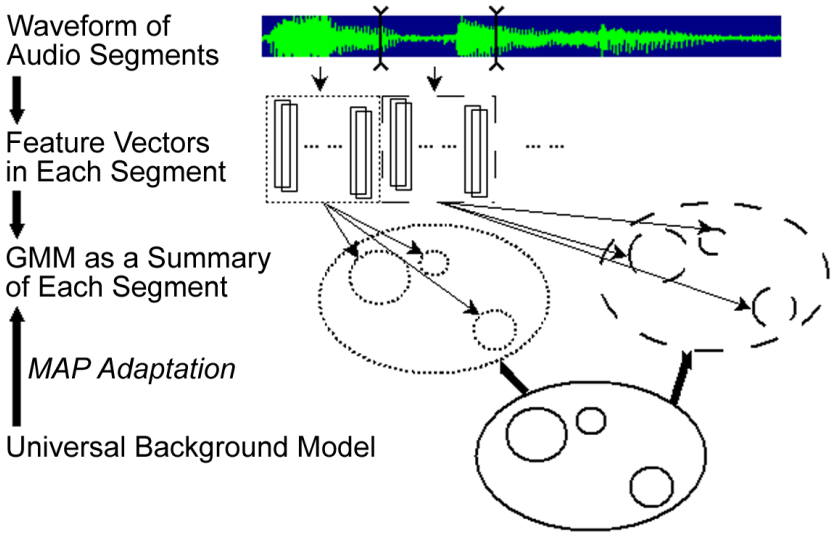

Figure 2.2: GMMs (indicated by the ovals) summarize audio segments using multiple unimodal Guassians (indicated by the circles).

regularized estimate of the true, underlying likelihood function governing each audio segment. Regularization (adaptative training based on a UBM) reduces the effects of outliers, e.g., noisy frames in an audio segment. Adaptive training also provides a natural measure of the difference between any given audio segment and the UBM, since each Gaussian component in the segment-specific likelihood has been adapted from a particular component of the UBM. Conversely, the use of a GMM allows arbitrarily precise representation of the acoustic feature likelihood, with large enough number of Gaussian components. Finally, the GMM clusters similar frames, by assigning them to the same kernel in the GMM.

We first estimate a UBM using feature vectors extracted from all training audio segments, regardless of their event labels. Then the distribution model of the feature vector for a certain audio segment is adapted from the UBM in order to maximize the a posteriori probability of the adapted model [63].

Here we denote $z \in \mathbb{R}^d$ as a feature vector, where d is the dimension of the

feature vector. The GMM distribution of variable z is

$$p(z;\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(z;\mu_k,\Sigma_k), \tag{2.3}$$

where $\Theta = \{w_1, \mu_1, \Sigma_1, \cdots\}$, w_k, μ_k and Σ_k are the weight, mean, and covariance matrix of the kth Gaussian kernel, respectively, and K is the total number of Gaussian kernels.

The density is a weighted linear combination of K unimodal Gaussian densities, namely,

$$\mathcal{N}(z;\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)}. \tag{2.4}$$

We obtain maximum likelihood parameters for the UBM using expectation-maximization (EM). For computational efficiency, the covariance matrices are restricted to be diagonal, which proves to be effective and computationally economical.

The UBM, learned from all training audio, specifies a feature domain, of which each segment-specific GMM span a subset. The subset constraint can be enforced by interpreting the UBM parameter set, Θ, as a set of conjugate-prior PDFs governing the distribution of segment-specific GMM parameters, θ, i.e., the segment-specific GMM has the a priori PDF $p(\theta;\Theta)$. The a posteriori probability of the segment-specific GMM parameters is obtained by multiplying $p(\theta;\Theta)$ by the data likelihood, $p(Z|\theta)$, where $Z = \{z_1, \ldots, z_H\}$ are the frames observed belonging to the segment of interest, and by then dividing by a normalizing constant; the normalizing constant is irrelevant to computation of the model parameters, and may be omitted. Thus, for example, MAP adaptation

selects the segment-specific mean parameters $\hat{\mu}_k$ to maximize

$$\begin{aligned} \ln p(\hat{\theta}, Z) &= \sum_{k=1}^{K} \ln \mathcal{N}(\hat{\mu}_k; \mu_k, \Sigma_k/r) \\ &\quad + \sum_{i=1}^{H} \ln \sum_{k=1}^{K} w_k \mathcal{N}(z_i; \hat{\mu}_k, \Sigma_k), \end{aligned} \tag{2.5}$$

where $\hat{\theta} = \{\hat{\mu}_1, \ldots, \hat{\mu}_K\}$ is the set of segment-specific GMM parameters, and $\Theta = \{w_1, \mu_1, \Sigma_1, \ldots\}$ are the parameters of the global GMM.

The joint distribution function $p(\hat{\theta}, Z)$ has the same form as the likelihood function $p(Z|\hat{\theta})$, and may therefore be optimized in the same way as a likelihood function, i.e., using EM with the hidden variable $Pr(k|z_i)$ as the posterior probability of the Gaussian component k for given feature vector z_i [64]. In the E-step, we compute the posterior probability as

$$Pr(k|z_i) = \frac{w_k \mathcal{N}(z_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(z_i; \mu_j, \Sigma_j)}, \tag{2.6}$$

$$n_k = \sum_{i=1}^{H} Pr(k|z_i), \tag{2.7}$$

and then the M-step updates the mean vectors, namely,

$$E_k(Z) = \frac{1}{n_k} \sum_{i=1}^{H} Pr(k|z_i) z_i, \tag{2.8}$$

$$\hat{\mu}_k = \alpha_k E_k(z) + (1 - \alpha_k)\mu_k, \tag{2.9}$$

where $\alpha_k = n_k/(n_k + r)$. MAP adaptation using conjugate priors is useful because it interpolates, smoothly, between the hyper-parameters μ_k and the maximum likelihood parameters $E_k(Z)$. In this work, r is adjusted empirically. If a Gaussian component has a high probabilistic count, n_k, then α_k approaches 1 and the adapted parameters emphasize the new sufficient statistics; otherwise, the adapted parameters are determined by the global model.

2.3.2 Approximating Kullback-Leibler divergence

Two *segment-specific* GMMs adapted from the same UBM are denoted as g_a and g_b. A natural similarity measure between these two GMMs is the Kullback-Leibler divergence,

$$D(g_a||g_b) = \int_z g_a(z) \log \frac{g_a(z)}{g_b(z)} \, dz \; .$$

The Kullback-Leibler divergence does not satisfy the conditions for a metric function. Instead, we can use its upper bound obtained by the log-sum inequality,

$$D(g_a||g_b) \leq \sum_{k=1}^{K} w_k \, D(\mathcal{N}(z; \mu_k^a, \Sigma_k)||\, \mathcal{N}(z; \mu_k^b, \Sigma_k)) \, ,$$

where μ_k^a and μ_k^b denote the adapted means of the k^{th} component from the segment GMMs g_a and g_b , respectively. Since the covariance matrices are shared across all adapted GMMs and the UBM, the right-hand side is equal to

$$d(a,b)^2 = \frac{1}{2} \sum_{k=1}^{K} w_k (\mu_k^a - \mu_k^b)^T \Sigma_k^{-1} (\mu_k^a - \mu_k^b) \; .$$

We can consider $d(a,b)$ as the Euclidean distance between the normalized GMM supervectors in a high-dimensional feature space [65],

$$d(a,b) = \|\phi(Z_a) - \phi(Z_b)\|_{\,2} \; , \tag{2.10}$$

where

$$\phi(a) = [\, \sqrt{\frac{w_1}{2}} \Sigma_1^{-\frac{1}{2}} \mu_1^a \, ; \cdots ; \sqrt{\frac{w_K}{2}} \Sigma_K^{-\frac{1}{2}} \mu_K^a \,] \; . \tag{2.11}$$

2.3.3 Kernel for SVM

GMM supervectors are used in an SVM for acoustic event classification. This multi-class classification task is implemented as binary classification problems

via the one-vs.-one method using LibSVM [66]. The distance defined in (2.10) can be evaluated using kernel functions, as

$$d(a,b) = \sqrt{K(a,a) - 2K(a,b) + K(b,b)}\ . \tag{2.12}$$

It is straightforward that kernel function $K(a,b) = \phi(a) \bullet \phi(b)$ satisfies (2.12), where $\phi(a)$ and $\phi(b)$ are defined as in (2.11).

2.4 Hybrid Architecture of the AED System

Both the HMM-based approach and the tandem HMM-connectionist approach engage the maximum a posteriori (MAP) decoding for AED: the recognizer outputs a sequence of hypothesized acoustic events corresponding to the highest $sequence$ a posterior probability, as discussed in Section 2.1. However, the best acoustic event sequence obtained by the MAP decoding is not optimal according to the performance measure for AED, $AED - ACC$, i.e. the acoustic event F-score (harmonic mean of precision and recall). For example, Mangu, Brill and Stolcke [67] proposed solving a similar problem using localized confidence rescoring: the MAP decoder defines a reduced search space, within which a new hypothesis is chosen explicitly to minimize the target performance measure. Confidence scoring also allows us to apply methods such as SVM-GMM-supervector classification, which are difficult to apply in a MAP decoding paradigm because of computational complexity and model structure limitations.

In this work, our AED system uses a two-stage hybrid architecture (Figure 2.3). In [67] a rescoring paradigm aligns all of the edges in an event lattice to the times marked in the MAP hypothesis. In the AED task, the number of labels is small enough to obviate lattice rescoring; therefore, we can take a route that is straightforward, yet effective and computationally inexpensive. The MAP decoding outputs a one-best result with boundaries of events and background, as well as hypothesized event types. The SVM-GMM-supervector approach is

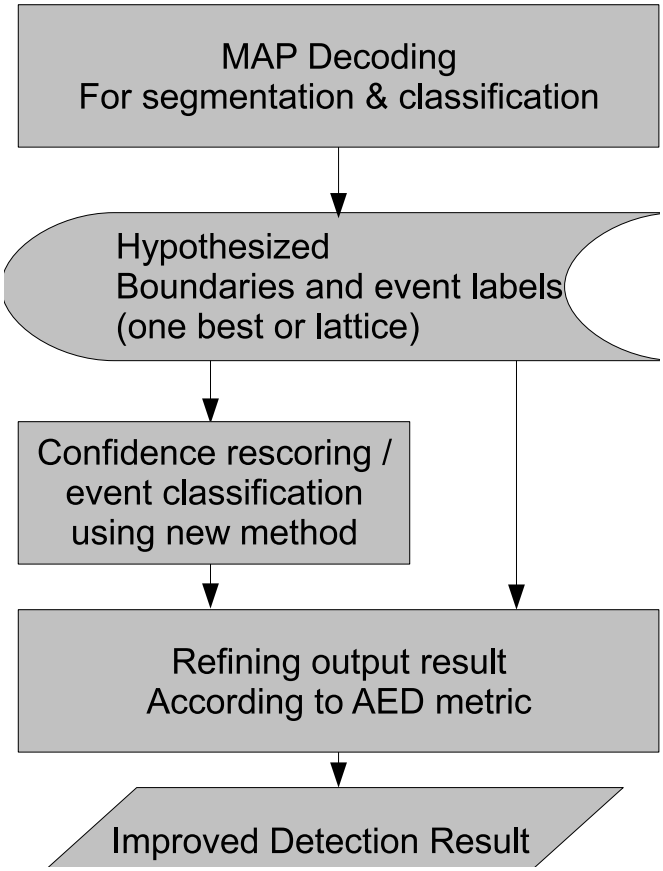

Figure 2.3: Hybrid architecture of AED system.

used as the confidence rescoring module. It models feature frames within all hypothesized audio segments, and proposes event types that might be different from the hypothesis obtained through MAP decoding.

Both hypothesized event types, referred to as the MAP labels and the SVM labels respectively, include the events of concern and a "background" label. Therefore, event label substitutions, each defined by a MAP label and an SVM label, may include substitutions between any pair of events, from an acoustic event to background or from background to an acoustic event. On the held out development data, the performance change is measured when only one particular type of label substitution is allowed. Those label substitution types that lead to the most performance boost on the held out data are chosen as the *valid event label*

substitutions, to be applied in testing. All other types of label substitutions are suppressed in testing, by retaining the MAP label.

We find in practice that the above valid event label substitutions are too specific and sometimes do not carry over well between different data. Therefore, in the experiments we only define valid event label substitutions according to the MAP labels. In fact, the most favorable approach turns out to allow the SVM-GMM-supervector classifier to assign labels to the audio segments labeled as background by the MAP decoding, recovering events that were missed in the first pass, but not to perform any substitutions among MAP-labeled non-background events. Readers interested in more general methods to combine detection results from multiple systems are refered to literature about the Recognition Output Voting Error Reduction (ROVER) [68], particularly its voting search modules.

The hybrid architecture works for two reasons.

First, the SVM-GMM-supervector approach functions complementarily to the MAP decoding as they operate in different hypothesis spaces. In particular, the MAP decoding engages properties such as state transition, varying length and N-gram event sequence statistics in the decision of boundaries and hypothesized event labels. The MAP decoding might suppress proposing short events or events similar to the background given the high variation in the background. By contrast, the SVM-GMM-supervector approach only considers feature distribution within an audio segment locally. The purely local approach of the rescoring module has been shown to outperform HMMs in tasks with loose sequence constraints [69]. Besides, the SVM-GMM-supervector approach does not impose explicit temporal structure within the audio segments, in contrast to left-to-right HMMs.

Second, the objective of MAP decoding differs from that of AED. For the maximum a posteriori hypothesis, each frame in the observation is considered. The detection metric, AED-ACC, only considers the temporal relationship between the hypothesized event boundaries and the reference event boundaries. Furthermore, neither MAP decoding nor the SVM-GMM-supervector classifier

treat background and acoustic events differently, while the AED-ACC measures only the F-score in detection of non-background events. SVM-GMM rescoring aims at the target performance metric by constraining it to allow only label substitutions (changes from the MAP labels) that are believed to improve the AED performance metric.

2.5 Seminar Room AED Experiments

2.5.1 Dataset and metric

The acoustic event detection experiments use the official data for CLEAR 2007 AED Evaluation [1]: about three hours for system development and two hours for system evaluation. All data are realistic seminar style, having both speech and acoustic events with possible overlap. The evaluation data has 1454 instances of target events. The target events included in the AED performance metric are door slam (ds), paper wrapping (pw), footsteps (st), phone ringing (pr), spoon cup jingle (cl), keyboard typing (kt), applause (ap), coughing (co), laughter (la), key jingle (kj), chair moving (cm), and knocking (kn). The counts of these events in the evaluation data are as in Figure 2.4. Many of the events are subtle and have low SNR compared to background noise or speech.

The performances are measured using AED-ACC [1], defined as the F-score (the harmonic mean between precision and recall) comparing system output acoustic event (AE) labels and reference AE labels. In particular, an event detected by the system is correct when there exists at least one matching reference event whose temporal center falls within the time boundaries of the detected event or the temporal center of the detected event is within the boundaries of at least one matching reference event. A reference event is considered correctly detected if its temporal center is within at least one matching system output or if there exist at least one matching system output whose temporal center falls within the boundaries of the reference event. AED-ACC aims to score detection

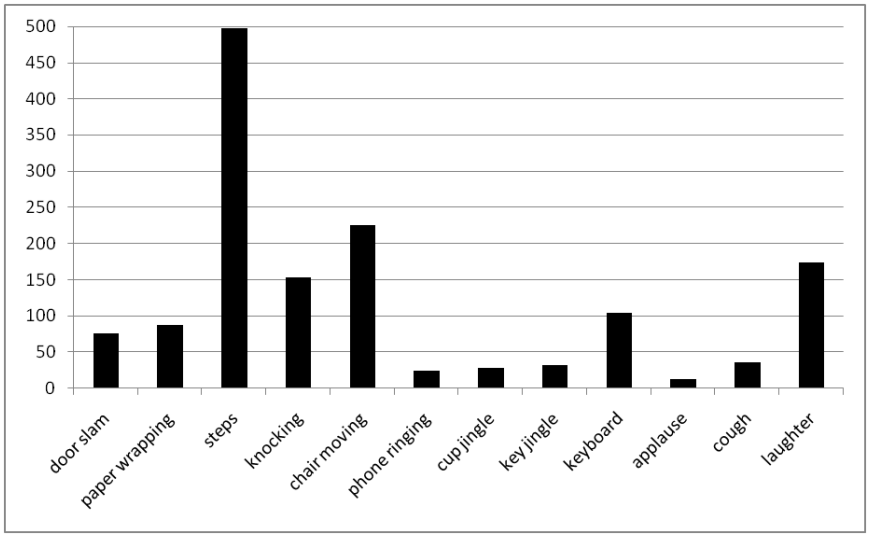

Figure 2.4: Counts of the twelve acoustic events in the evaluation data.

and classification of all acoustic event instances, oriented for applications such as real-time services for smart rooms and audio-based surveillance.

2.5.2 Experiment setup

The audio features used in these experiments are AED feature derived using a modified AdaBoost approach we proposed in [42]. The feature pool consists of two feature sets widely-used in speech recognition as well as other audio applications. The first set consists of 26 MFCCs calculated in the 0 Hz - 11000 Hz band along with their first order regression (delta) coefficients and second order regression (acceleration) coefficients. The second set consists of 26 log frequency filter bank parameters, their delta and acceleration coefficients on the same frequency range. The AED feature set is derived using a boosting approach from the union of the two baseline feature sets. The AED feature set used has 78 feature components.

Two sets of experiments are carried out to demonstrate the performance of the tandem connectionist-HMM approach and the SVM-GMM-supervector ap-

proach for refining event label hypotheses.

The first experiment evaluates the tandem connectionist-HMM approach. The contextual window size (number of input nodes divided by 78) is picked to be five. The number of hidden nodes is chosen as 1200 empirically for best performance on a development dataset. The number of output nodes is set to 14, i.e., the number of acoustic events plus one for frames labeled as unknown sounds and one for background frames. The transformed output of the best-performing ANN is concatenated with the derived AED feature set as the input to the HMM component.

The second experiment presents performance of the SVM-GMM-supervector approach discussed in Subsection 2.3, used in the hybrid architecture discussed in Subsection 2.4. The number of Gaussian mixtures is set to be 128. Two sets of results are reported, obtained by applying the approach on top of either the HMM-based approach or the tandem connectionist-HMM approach.

When training the systems, we hold out one third of the three hour development data to tune some system parameters. Once the parameters are determined, the models are retrained with all the development data.

2.5.3 Results

In Table 2.1, we demonstrate the effectiveness of the tandem HMM-connectionist approach and the SVM-GMM-supervector approach used in the hybrid architecture. We can observe that the average AED-ACC across all twelve events improves from 34% to 35.3% by engaging the tandem approach (denoted as "Tandem"). The SVM-GMM-supervector (denoted as "HMM+S") boosts performance from 34% to 37.5% by relabeling event segments proposed by the HMM-based AED system (denoted as "HMM"), as described in Subsection 2.4. Using this hybrid architecture of both tandem and SVM-GMM-supervector approaches yields the best AED-ACC of 41.2% (denoted as "Tandem+S").

Performance on individual acoustic events is also presented for the different

settings. It is shown that the number of individual acoustic events scoring the highest is the largest for the best setting of "Tandem+S". The single most dramatic performance boost on an individual event is that of "keyboard typing" (kt), achieved by engaging the SVM-GMM-supervector approach. The MAP decoding approaches, i.e., HMM or tandem approaches, could not well distinguish "keyboard typing" from background. In fact, many events that are easily confused with the background in the first pass, e.g., "keyboard typing" and "steps", are recovered for reasons discussed in Subsection 2.4. This highlights that the SVM-GMM-supervector in the hybrid architecture has capability complementary to the MAP decoding approaches. The best setting of "Tandem+S" performs significantly better than the baseline HMM-based system according to the Friedman's test ($p = 0.02$).

All results presented here are improved from our system in the 2007 CLEAR Acoustic Event Detection Evaluation, where we achieved the best performance, similar to the performance of the baseline HMM system in Table 2.1.

2.6 Acoustic Fall Classification and Detection Experiments

Assistance to dependent people, particularly to the elderly living alone at home, has been attracting increasing attention in today's aging societies [3]. Reliable and speedy detection of falls by automatic monitoring of the home is expected to be of benefit to both elderly and caregivers.

We apply the AED methods to automatic fall detection using one unobtrusive far-field microphone. The detection task identifies existence and approximate occurrence time of falls. Segment boundaries of the acoustic input are found by the Viterbi algorithm using single-state HMMs (GMMs) with self-transitions for different falls and other noise events. A bigram model is trained on the fall, noise and background sequences observed in the training data. Each audio

Table 2.1: Effectiveness of different components in the AED system.

AED-ACC (%)	ap	cl	cm	co	ds	kj	kn	kt	la	pr	pw	st	Average
HMM	44.4	25.5	31.3	31.2	57.3	33.2	13.5	1.9	51.3	36.7	17.6	36.8	34.0
Tandem	52.6	21.9	37.2	**51.3**	**63.0**	29.6	11.5	0.0	54.2	**42.7**	25.8	34.6	35.3
HMM+S	44.4	25.0	33.7	31.2	56.6	33.2	**20.9**	35.5	51.3	36.7	19.2	41.3	37.5
Tandem+S	52.6	21.5	**37.4**	47.9	**63.0**	29.6	13.6	**44.8**	**58.6**	**42.7**	**26.7**	**44.4**	**41.2**

Table 2.2: Sound classes for fall classification and detection.

FA	sound resulting from the subject falling
ST	noise when the subject sits down on the chair, possibly leading to a bit of chair movement
CL	noise of clapping hands
GU	noise when the subject gets up from the floor
MP	noise of moving, putting, or catching an object
DO	noise of dropping an object on the floor
DN	noise of opening/closing doors
WK	noise of walking steps
MO	other noise, including speech and non-speech human voices, telephone ringing and other acoustically salient noise
BG	background noise, usually not perceptually salient

segment is classified into fall or various types of noise, either directly using the hypothesis labels obtained in the Viterbi algorithm or after being refined by the SVM-GMM-supervector approach.

To better distinguish fall from all competing noise, we model falls and nine classes of noise in the living environment. These classes, shown in Table 2.2, are adopted with three considerations: Each class should have a sufficient number of instances in the training data. Each class is relatively distinguishable from others. The classes are chosen to better distinguish falls from noise.

2.6.1 Dataset

Our experiments are carried out on the acoustic fall data collected in the European project Netcarity [3, 70]. The dataset [1] is of about 7 hours in 32 sessions, involving 13 different actors as subjects that might fall or perform other activities, and various other people that produce noise in the background. Figure 2.5

[1] We would like to thank the authors of [70] for the Netcarity dataset, and Vit Libal and Larry Sansone for assistance with the dataset.

provides a snapshot. This dataset well simulates an environment that elderly people might encounter at home. We split the dataset into 20 training sessions, 7 testing sessions and 5 held out sessions for tuning the parameters. The subjects in the training and held out sessions do not overlap with those in testing. We map the labels in the Netcarity dataset to the ten classes detailed in Table 2.2 as the ground truth.

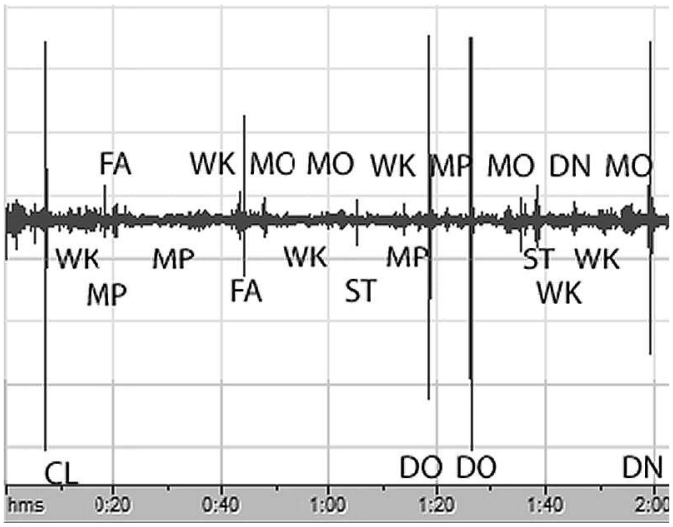

Figure 2.5: Snapshot of Netcarity fall dataset (boundaries omitted for simplicity).

2.6.2 Experiment setup

The first experiment is classification of audio segments whose ground-truth boundaries are provided. Classification accuracy of all the ten classes in Table 2.2 reflects the overall performance of the classifiers. F-score of the fall segments reflects the capability to distinguish falls from all other noise. Both the GMM approach and the SVM-GMM-supervector approach are implemented with 512 Gaussian components for each GMM in this experiment.

The second experiment is detection of falls in acoustic signal of whole ses-

sions. We measure the detection performance using AED-ACC [1], the harmonic mean between precision and recall. In the fall detection experiment, we further require that all proposed fall segments not exceed a maximum length of 5 seconds so that the system output can be used for timely response to falls. Fall segments that exceed 5 seconds, if any, are removed from the output before scoring. We choose detection using the dynamic programming algorithm with the GMM audio segment modeling as our baseline. The SVM-GMM-supervector approach is adopted to re-classify the audio segments with perceptually confusable labels in the baseline output. In this dataset, the perceptually confusable labels are chosen to be falls (FA), dropping objects (DO), getting up (GU) and walking (WK).

The frame-based features are extracted from 25 ms Hamming windows with a step size of 10 ms We calculate 12 perceptual linear predictive (PLP) coefficients and the overall energy. On these 13 dimensions, utterance level cepstral mean subtraction is applied.

2.6.3 Results

Figure 2.6 illustrates the classification accuracy of all the ten fall/noise classes, and the F-score for fall segments. The results show that the SVM-GMM-supervector approach improves from the GMM approach on classifying fall and noise segments.

Figure 2.7 illustrates that using the SVM-GMM-supervector approach to re-classify confusable segments improves AED-ACC measure of the baseline output produced by the Viterbi algorithm using the GMMs.

In these results, we can see that in general the method that performs well in the classification of falls and other noise categories also provides better measures in which we only care about the falls, i.e. the F-score of falls in classification and the AED-ACC in fall detection. This suggests that better modeling of the alternative categories, including background, improves the capability to identify

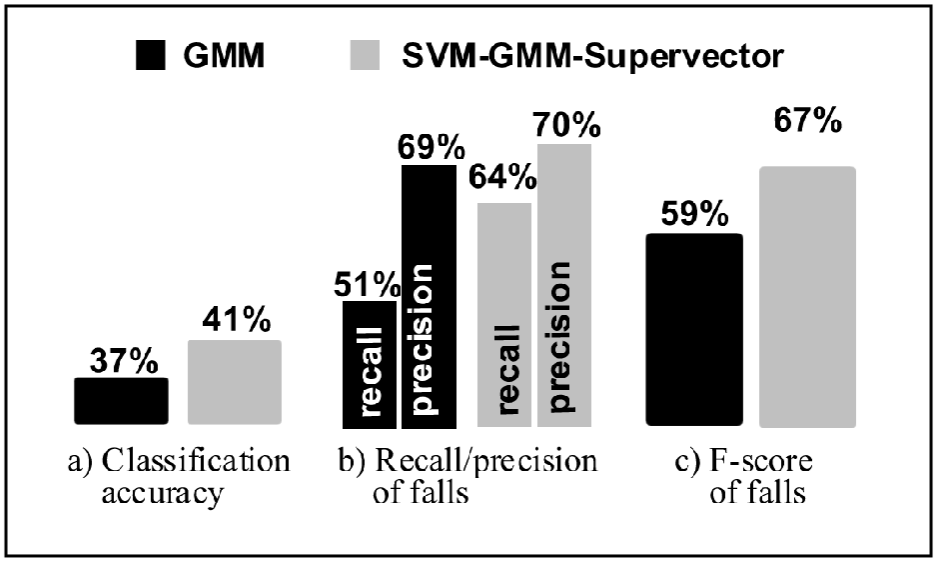

Figure 2.6: Classification of falls/noise.

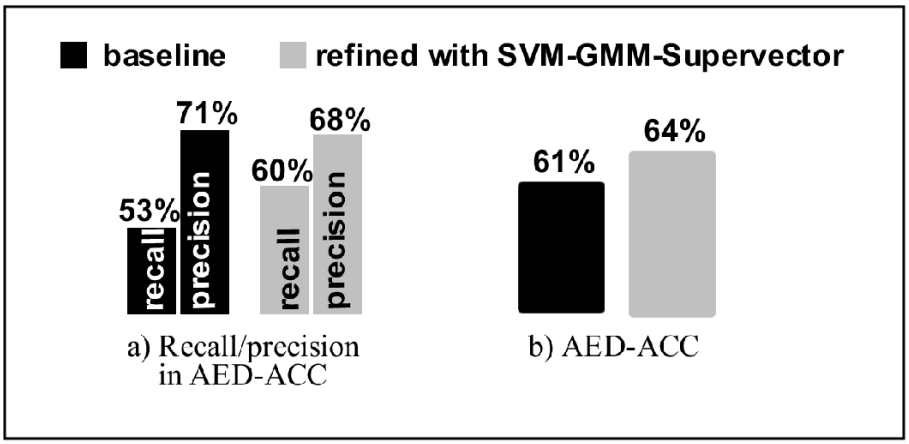

Figure 2.7: Detection of falls.

the target category.

CHAPTER 3

GENERAL IMAGE AND VIDEO MODELING

Real world events present significant variation in the visual cues, even after various computer vision processing, such as motion detection, background subtraction and lighting normalization. Most previous research on video event analysis is limited to video captured by fixed cameras in surveillance applications or greatly constrained live video. Even more challenging is video event recognition in unconstrained domains such as broadcast news, which contains rich information about objects, people, activities, and events [47]. For example, events in broadcast news video may involve small objects, large camera motion, and significant object occlusion, and reliable object tracking becomes very challenging under these scenarios.

Some recent research attempted to provide solutions for event analysis in news video. Ebadollahi et al. [71] proposed to treat each frame in a video clip as an observation and apply HMM to model the temporal patterns of event evolution in news video. Xu and Chang [28] proposed to encode a video clip as a bag of orderless descriptors obtained from mid-level semantic concept classifiers extracted from all of the constituent frames, along with the global features extracted within each video frame, and then apply the Earth Mover's Distance (EMD) [72] to integrate similarities among frames from two video clips. Multi-level temporal pyramid structure was adopted to integrate the information from different sub-clips with integer-value constrained EMD to explicitly align the sub-clips.

Specialized object or semantic concept detectors, such as those for faces, hands, computer screens, books and human figures, have been successfully used

to provide discriminative cues for event detection [34, 35, 28]. Such lower level detectors are believed to provide robust representation for realistic images and video clips. We take an alternative approach, intending not to train ad-hoc and specialized object detectors, which require expensive annotation for training images.

We propose Gaussianized vector representation for realistic image and video modeling. Each image or video clip is expressed as a set of patch-based local descriptors. Such descriptors can be extracted by a feature point detector, such as the SIFT detector [73], or from a dense pixel grid. We use a Gaussian mixture model (GMM) to approximate the distribution of these local descriptors in *each* image or video clip. These Gaussian components are adapted from a global set of Gaussian components according to the maximum a posteriori criterion. This establishes unsupervised correspondence between different images or video clips, and suppresses noise in the distributions. The Gaussianized vector representation is constructed from an image-specific or video-clip-specific GMM by taking properly normalized mean vectors of all the Gaussian components, thus forming a corresponding and uniform-length representation for images or video clips of different sizes and lengths. It is shown that the linear kernel based on such representations approximates the KL divergence between local descriptor distributions of different images or video clips.

Before the kernels are used for categorization or localization problems, a Within-Class Covariance Normalization (WCCN) approach is utilized to depress the kernel components with high-variability for data labeled as the same category. The refined kernel is used as a similarity measurement in the nearest neighbor or nearest centroid classification, as well as in a support vector machine [74] for margin-based classification.

For video events in broadcast news, we successfully demonstrated that the patch-based Gaussianized vector representation achieves the best reported event categorization accuracy, by effective modeling of whole images without annotating the training images [50]. In particular, our results reported in [50] out-

performed the then state-of-the-art [28] based on a set of specialized semantic detectors trained on human-annotated images.

Different from classification or regression problems that work on the whole images, an object localization task involves finding the rectangle bounding boxes that scored the highest according to a particular video model, with varying locations, widths and heights. A natural way to carry out localization is the sliding window approach [30]. However, an exhaustive search in an $n \times n$ image needs to evaluate $O(n^4)$ candidate bounding boxes, and is not affordable for a complicated representation such as the Gaussianized vector representation. Tricky heuristics about possible bounding box locations, widths and heights, or local optimization methods would have to be used, resulting in false estimates. This intrinsic trade-off between performance and efficiency of the sliding window approach is not desirable. Lampert et al. introduced a branch-and-bound search scheme [75], which finds the globally optimal bounding box efficiently without the above problems.

I present an efficient object localization approach based on the Gaussianized vector representation. The branch-and-bound search scheme [75] is adopted to perform a fast hierarchical search for the optimal bounding boxes, leveraging a quality bound for rectangle sets. We demonstrate that the quality function based on the Gaussianized vector representation can be written as the sum of contributions from each feature vector in the bounding box. Moreover, a quality bound can be obtained for any rectangle set in the image, with little computational cost, in addition to calculating the Gaussianized vector representation for the whole image. Experiments on a multi-scale car dataset show that the proposed object localization approach based on the Gaussianized vector representation outperforms previous work using the histogram-of-keywords representation.

3.1 Gaussianized Vector Representation

The Gaussian mixture model (GMM) is widely used in various pattern recognition problems [76, 77]. We propose the Gaussianized vector representation, which encodes an image as a bag of feature vectors, the distribution of which is described by a GMM. Then a GMM supervector is constructed using the means of the GMM, normalized by the covariance matrices and Gaussian component priors. A GMM-supervector-based kernel is designed to approximate Kullback-Leibler divergence between the GMMs for any two images, and is utilized for supervised discriminative learning using an SVM, nearest neighbor or nearest centroid methods.

The Gaussianized vector representation is closely connected to the classic histogram of keywords representation. In the traditional histogram representation, the keywords are chosen by the k-means algorithm on all the features. Each feature is distributed to a particular bin based on its distance to the cluster centroids. The histogram representation obtains rough alignment between feature vectors by assigning each to one of the histogram bins. Such a representation provides a natural similarity measure between two images based on the difference between the corresponding histograms. However, the histogram representation has some intrinsic limitations. In particular, it is sensitive to feature outliers, the choice of bins, and the noise level in the data. Besides, encoding high-dimensional feature vectors by a relatively small codebook results in large quantization errors and loss of discriminability.

Several approaches have been proposed in the literature to overcome these limitations. Soft assignment, which allows each feature vector to belong to multiple histogram bins, has been suggested to capture partial similarity between images [78, 79, 80, 81, 82, 83]. To enhance the discriminating capability of histograms, Farquhar et al. [84] and Perronnin et al. [78] introduced several ways to construct category-specific histograms. Larlus and Jurie [85] and Yang et al. [79] suggested to integrate histogram construction with classifier training, and

Moosmann et al. [86] proposed to use randomized forests to build discriminative histograms.

Gaussianized vector representation enhances the histogram representation in the following ways. First, k-means clustering leverages the Euclidean distance, while the GMM leverages the Mahalanobis distance by means of the component posteriors. Second, k-means clustering assigns one single keyword to each feature vector, while the Guassinized vector representation allows each feature vector to contribute to multiple Gaussian components statistically. Third, histogram-of-keywords only uses the number of feature vectors assigned to the histogram bins, while the Gaussianized vector representation also engages the weighted mean of the features in each component, leading to a more informative representation.

3.1.1 GMM for feature vector distribution

We estimate a GMM for the distribution of all feature vectors in an image. The estimated GMM is a compact description of the single image, less prone to noise compared with the feature vectors. Yet, with increasing number of Gaussian components, the GMM can be arbitrarily accurate in describing the underlying feature vector distribution. The Gaussian components impose an implicit multi-mode structure of the feature vector distribution in the image. When the GMMs for different images are adapted from the same global GMM, the corresponding Gaussian components imply certain correspondence.

In particular, we obtain one GMM for each image in the following way.

First, a global GMM is estimated using feature vectors extracted from all training images, regardless of their labels. Here we denote z as a feature vector, whose distribution is modeled by a GMM, a weighted linear combination of K

unimodal Gaussian components,

$$p(z;\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(z;\mu_k^{global},\Sigma_k).$$

$\Theta = \{w_1, \mu_1^{global}, \Sigma_1, \cdots\}$, w_k, μ_k and Σ_k are the weight, mean, and covariance matrix of the kth Gaussian component,

$$\mathcal{N}(z;\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)}. \tag{3.1}$$

We restrict the covariance matrices Σ_k to be diagonal [87], which proves to be effective and computationally economical.

Second, an image-specific GMM is adapted from the global GMM, using the feature vectors in the particular image. This is preferred to direct separate estimation of image-specific GMMs for the following reasons:

1. It improves robust parameter estimation of the image specialized GMM, using the comparatively small number of feature vectors in the single image.

2. The global GMM learned from all training images may provide useful information for the image specialized GMM.

3. As mentioned earlier, it establishes correspondence between Gaussian components in different images-specific GMMs.

For robust estimation, we only adapt the mean vectors of the global GMM and retain the mixture weights and covariance matrices. In particular, we adapt an image-specific GMM by the maximum a posteriori (MAP) criterion with the weighting all on the adaptation data. The posterior probabilities and the updated means are estimated as

$$Pr(k|z_j) \;=\; \frac{w_k \mathcal{N}(z_j;\mu_k^{global},\Sigma_k)}{\sum_{k=1}^{K} w_k \mathcal{N}(z_j;\mu_k^{global},\Sigma_k)}, \tag{3.2}$$

$$\mu_k = \frac{1}{n_k} \sum_{j=1}^{H} Pr(k|z_j) z_j, \tag{3.3}$$

where n_k is a normalizing term,

$$n_k = \sum_{j=1}^{H} Pr(k|z_j), \tag{3.4}$$

and $Z = \{z_1, \ldots, z_H\}$ are the feature vectors extracted from the particular image.

As shown in Equation 3.2, the image-specific GMMs leverage statistical membership of each feature vector among multiple Gaussian components. This sets the Gaussianized vector representation apart from the histogram of keyword representation which originally requires hard membership in one keyword for each feature vector. In addition, Equation 3.3 shows that the Gaussianized vector representation encodes additional information about the feature vectors statistically assigned to each Gaussian component, via the means of the components.

Given the computational cost concern for many applications, another advantage of using GMM to model feature vector distribution is that efficient approximation exists for GMM that does not significantly degrade its effectiveness. For example, we can prune out Gaussian components with very low weights in the adapted image-specific GMMs. Another possibility is to eliminate the additions in Equation 3.3 that involve very low priors in Equation 3.2. Neither of these approaches significantly degrades GMM's capability to approximate a distribution [76].

3.1.2 Kernel function based on Gaussianized vector representation

Suppose we have two images whose ensembles of feature vectors, Z_a and Z_b, are modeled by two adapted GMMs according to Section 3.1.1, denoted as g_a and g_b. A natural similarity measure is the approximated Kullback-Leibler divergence

[56]

$$D(g_a||g_b) \leq \sum_{k=1}^{K} w_k D(\mathcal{N}(z;\mu_k^a,\Sigma_k)||\mathcal{N}(z;\mu_k^b,\Sigma_k)), \tag{3.5}$$

where μ_k^a denotes the adapted mean of the kth component from the image-specific GMM g_a, and likewise for μ_k^b. The right side of the above inequality is equal to

$$d(Z_a, Z_b) = \frac{1}{2}\sum_{k=1}^{K} w_k(\mu_k^a - \mu_k^b)^T\Sigma_k^{-1}(\mu_k^a - \mu_k^b). \tag{3.6}$$

The term $d(Z_a, Z_b)^{\frac{1}{2}}$ can be considered as the Euclidean distance in another high-dimensional feature space,

$$\begin{aligned} d(Z_a, Z_b) &= \|\phi(Z_a) - \phi(Z_b)\|^2 \\ \phi(Z_a) &= [\sqrt{\frac{w_1}{2}}\Sigma_1^{-\frac{1}{2}}\mu_1^a; \cdots ; \sqrt{\frac{w_K}{2}}\Sigma_K^{-\frac{1}{2}}\mu_K^a]. \end{aligned} \tag{3.7}$$

Thus, we obtain the corresponding kernel function

$$k(Z_a, Z_b) = \phi(Z_a) \bullet \phi(Z_b). \tag{3.8}$$

3.2 Robustness to Within-Class Variation

The variation of the object class and the background adds to the difficulty of the localization problem. The Gaussianized vector representation is based on Gaussian mixtures adapted from the global model. To further enhance the discriminating power between objects and the background, we propose incorporating a normalization approach, which depresses the kernel components with high-variation within each class. This method was first proposed in the speaker recognition problem [88] as Within-Class Covariance Normalization (WCCN).

We assume the Gaussianized vector representation kernels in Equation 3.8 are characterized by a subspace spanned by the projection matrix V^{all}. The desired normalization suppresses the subspace, V, that has the maximum inter-image

distance d_V for images (or image regions for the localization application) of the same category (or either the object or the background):

$$d_V^{ab} = \|V^T\phi(Z_a) - V^T\phi(Z_b)\|^2. \tag{3.9}$$

Since V identifies the subspace in which feature similarity and label similarity are most out of sync, this subspace can be suppressed by calculating the kernel function as in Equation 3.10, where C is a diagonal matrix, indicating the extent of such asynchrony for each dimension in the subspace.

$$k(Z_a, Z_b) = \phi(Z_a)^T(I - VCV^T)\phi(Z_b). \tag{3.10}$$

We can find the subspace V by solving the following:

$$V = \arg\max_{V^TV=I} \sum_{a\neq b} d_V^{ab} W_{ab}, \tag{3.11}$$

where W_{ab}=1 when Z_a and Z_b both belong to the object class or the background class, otherwise $W_{ab} = 0$.

Denote $\hat{Z} = [\phi(Z_1), \phi(Z_2), \cdots, \phi(Z_N)]$, where N is the total number of training images; it can be shown that the optimal V consists of the eigenvectors corresponding to the largest eigenvalues Λ of the matrix $\hat{Z}(D - W)\hat{Z}^T$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} W_{ij}, \forall i$.

The eigenvalues Λ indicate the extent to which the corresponding dimensions vary within the same class. In order to ensure the diagonal elements of C remain in the range of $[0, 1]$, we apply a monotonic mapping $C = 1 - max(I, \Lambda)^{-1}$.

3.3 Categorization with Gaussianized Vector Representation

3.3.1 Nearest neighbor or nearest centroid

The video event recognition, as a categorization problem, can be conducted directly based on the kernel similarity and the nearest neighbor or nearest centroid approach. Here we use the kernel similarity between a testing video clip and the centroid of an event for similarity metric, where the centroid of an event is defined in the Gaussianized vector space: namely, the centroid, $\bar{Z}^s$, of the s-th event is

$$\phi(\bar{Z}^s) = \frac{1}{N^s} \sum_{i \in \pi^s} \phi(Z_i), \tag{3.12}$$

where Z_i is the set of patch-based descriptors extracted from the i-th training video clip, N^s is the number of video clips belonging to the s-th event, and π^s denotes the index set of the samples belonging to the s-th event. Then, the final video event recognition is based on normalized similarity vector as

$$C_1(Z) = [\frac{K(Z, \bar{Z}^1)}{\sum_s K(Z, \bar{Z}^s)}, \frac{K(Z, \bar{Z}^2)}{\sum_s K(Z, \bar{Z}^s)}, \cdots, \frac{K(Z, \bar{Z}^S)}{\sum_s K(Z, \bar{Z}^s)}],$$

where S is the total number of predefined event categories, and Z is the set of patch-based descriptors extracted from a test video clip.

3.3.2 Support vector machine

Alternatively, a support vector machine (SVM) is used with the above kernel to distinguish between categories, or between objects and backgrounds. The binary classification score for a test image can be formulated as

$$g(Z) = \sum_t \alpha_t k(Z, Z_t) - b, \tag{3.13}$$

where α_t is the learned weight of the t^{th} training sample Z_t and b is a threshold parameter. $k(Z, Z_i)$ is the value of a kernel function for the t^{th} training Gaussianized vector representation Z_i and the test Gaussianized vector representation Z.

Similarly, the multi-class SVM can also output a confidence vector, denoted as

$$C_2(Z) = [p_1(Z), p_2(Z), \cdots, p_S(Z)], \tag{3.14}$$

where $p_s(Z)$ can be roughly considered as the probability of the video clip or image belonging to the s-th category. Then, the classification can be conducted based on the output values in $C_2(Z)$.

The support vectors and their corresponding weights are learned using the standard quadratic programming optimization process. We use the SVM training tools implemented in Libsvm [66] for both binary classification and multi-class classification.

3.3.3 Combining different classifiers

The motivations of centroid-based video event recognition and margin-based video event recognition are essentially different. Our preliminary experiments show that the outputs from these two classifiers are often complementary to each other; therefore, we can optionally fuse the outputs from these two classifiers. The vectors $C_1(Z)$ and $C_2(Z)$ both roughly measure the probabilities that a test video clip belongs to different video events, and hence we can average them for a more robust output as

$$C(Z) = \frac{C_1(Z) + C_2(Z)}{2}. \tag{3.15}$$

The classification can be done based on the averaged probability vector $C(Z)$.

3.3.4 Visualizing the Gaussianized vector representation

We visualize the Gaussianized vector representation to demonstrate that soft correspondence across different video clips is established and much more information than the histogram-of-keywords is represented.

Each video clip is first represented as a set of patch-based local descriptors. We project these local descriptors into a 2D feature space using a dimensionality reduction technique, Locality Preserving Projection [89]. All the component means of the global GMM are mapped to this 2D space. For local descriptor, its coordinates in this 2-D space are the sums of the coordinates of the component means of the global GMM, weighted by the posteriors of the components for the given descriptor.

Figure 3.1 shows the 2D distributions of the patch-based descriptors from three video clips, two of which belong to the same video event category of *Election Campaign Greeting*, and the other to the video event of *Running*. We can see that the distributions in the 2D space are characterized by distribution near different components of the global GMM, as indicated by the different colors in Figure 3.1. These components implicitly establish the correspondence between patch-based descriptors in different video clips, which shows that the Gaussianized vector representation offers the capability to match the patches from two video clips, similar in content yet different in spatial positions, scales, and temporal positions. For the video clips from the same event category we can see that the feature vector distributions near the corresponding components tend to share a similar structure, while they are relatively more different for those from different categories.

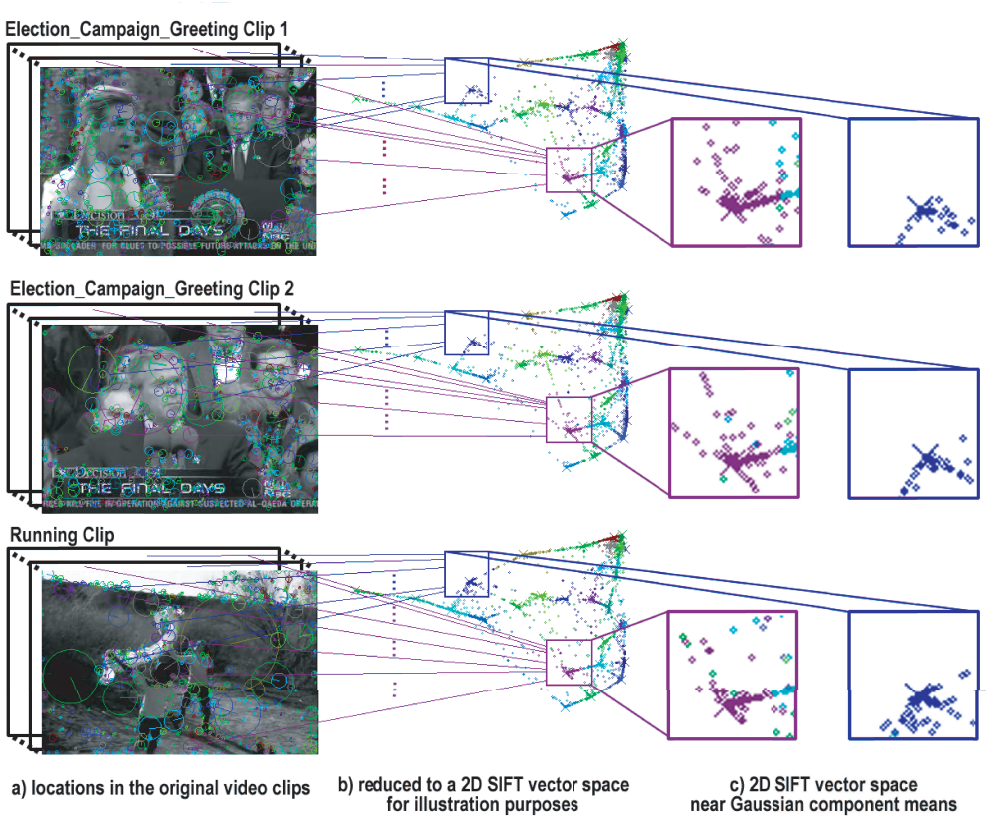

Figure 3.1: Visualization of the Gaussianized vector representation and its capability of matching local visual cues different in spatial positions, scales, and temporal positions.

3.4 Localization with Gaussianized Vector Representation

Object localization predicts the bounding box of a specific object class within the image. Effective object localization relies on an efficient and effective searching method, and robust image representation and learning method. The task remains challenging due to within-class variations and the large search space for candidate bounding boxes.

Robust image representation and learning is critical to the success of various computer vision applications. Some of the successful features are histogram of oriented gradients [90] and Haar-like features [91]. Patch-based histogram-of-keywords image representation methods represent an image as an ensemble

of local features discretized into a set of keywords. These methods have been successfully applied in object localization [75] and image categorization [92].

In this section, I present an object localization approach combining the efficient branch-and-bound searching method with the robust Gaussianized vector representation. The branch-and-bound search scheme [75] is adopted to perform a fast hierarchical search for the optimal bounding boxes, leveraging a quality bound for rectangle sets. We demonstrate that the quality function based on the Gaussianized vector representation can be written as the sum of contributions from each feature vector within the bounding box. Moreover, a quality bound can be obtained for any rectangle set in the image, with little extra computational cost, in addition to calculating the Gaussianized vector representation for the whole image.

To achieve improved robustness to variation within the object class and the background, we propose incorporating the normalization approach in Section 3.2 that suppresses the within-class covariance of the Gaussianized vector representation kernels in the binary support vector machine (SVM) and the branch-and-bound searching scheme.

I first present the efficient search scheme based on branch-and-bound in Subsection 3.4.1. Then I detail the quality function and quality bound for the Gaussianized vector representation in Subsections 3.4.2 and 3.4.3 respectively. In Subsection 3.4.4, the variation-normalization approach is incorporated in the localization framework.

3.4.1 Branch-and-bound search

Localization of an object is essentially to find the subarea in the image on which a quality function f achieves its maximum, over all possible subareas. One way to define these subareas is the bounding box, which encodes the location, width and height of an object with four parameters, i.e., the top, bottom, left and right coordinates (t, b, l, r).

The sliding window approach is most widely used in object localization with bounding boxes [30, 93]. To find the bounding box where the quality function f reaches its global maximum, we need to evaluate the function on all possible rectangles in the image, whose number is on the order of $O(n^4)$ for an $n \times n$ image. To reduce the computational cost, usually only rectangles at a coarse location grid and of a small number of possible widths and heights are considered. On the other hand, different approaches can be adopted to use a local optimum to approximate the global one, when the quality function f has certain properties, such as smoothness. All these approaches make detection tractable at the risk of missing the global optimum, and with demand for well informed heuristics about the possible location and sizes of the object.

In recent years, the most popular technique in the sliding window approach is the cascade [91]. The cascade technique decomposes a strong object/non-object classifier into a series of simpler classifiers. These classifiers are arranged in a cascade, so that the simpler and weaker classifiers will eliminate most of the candidate bounding boxes, before the more powerful and complicated classifiers will make finer selection. However, the cascade of classifiers is slow to train. Moreover, it unfortunately involves many empirical decisions, e.g., choosing the false alarm rate and missed-detection rate at each stage of the cascade. The cascade technique always reduces the performance compared with the original strong classifier.

The branch-and-bound search scheme was recently introduced [75] to find the globally optimal bounding box without the heuristics and assumptions about the property of the quality function. It hierarchically splits the parameter space of all the rectangles in an image, and gives priority to the parts with higher quality bounds.

For localization based on bounding boxes, a set of rectangles is encoded with $[T, B, L, R]$, each indicating a continual interval for the corresponding parameter in (t, b, l, r). The approach starts with a rectangle set containing all the rectangles in the image, and terminates when one rectangle is found that has a

quality function no worse than the bounds $\hat{f}$ of any other rectangle set.

At every iteration, the parameter space $[T, B, L, R]$ is split along the largest of the four dimensions, resulting in two rectangle sets both pushed into a queue together with their upper bounds. The rectangle set with the highest upper bound is retrieved from the queue for the next iteration.

The steps of the branch-and-bound search scheme can be summarized as follows:

1. Initialize an empty queue Q of rectangle sets. Initialize a rectangle set $\mathbf{R}$ to be all the rectangles: T and B are both set to be the complete span from zero to the height of the image. L and R are both set to be the complete span from zero to the width of the image.

2. Obtain two rectangle sets by splitting the parameter space $[T, B, L, R]$ along the dimension with the largest range.

3. Push the two rectangle sets in Step 2 into queue Q with their respective quality bound.

4. Update R with the rectangle set with the highest quality bound in Q.

5. Stop and return R if $\mathbf{R}$ contains only one rectangle R. Otherwise go to Step 2.

The quality bound $\hat{f}$ for a rectangle set $\mathbf{R}$ should satisfy the following conditions:

1. $\hat{f}(\mathbf{R}) \geq max_{R \in \mathbf{R}} f(R)$

2. $\hat{f}(\mathbf{R}) = f(R)$, if R is the only element in $\mathbf{R}$

Critical for the branch-and-bound scheme is to find the quality bound $\hat{f}$. Given the proven performance of the Gaussianized vector representation in classification tasks shown in previous work [94, 50, 95, 96], we are motivated to design a quality bound based on this representation for efficient localization.

3.4.2 Quality function

For the Gaussianized vector representation, the binary classification score in Equation 3.13 informs the confidence that the evaluated image subarea contains the object instead of pure background. Therefore, we can use this score as the quality function for the Gaussianized vector representation.

In particular, according to Equation 3.8 and Equation 3.13, the quality function f can be defined as follows:

$$f(Z) = g(Z) = \sum_t \alpha_t \phi(Z) \bullet \phi(Z_t) - b, \tag{3.16}$$

which can be expanded using Equation 3.7,

$$\begin{aligned} f(Z) &= \sum_t \alpha_t \sum_{k=1}^{K} \sqrt{\frac{w_c}{2}} \Sigma_c^{-\frac{1}{2}} \mu_k \\ &\quad \bullet \sqrt{\frac{w_k}{2}} \Sigma_c^{-\frac{1}{2}} \mu_k^t - b \\ &= \sum_t \alpha_t \sum_{k=1}^{K} \frac{w_k}{2} \Sigma_k^{-1} \mu_k \bullet \mu_k^t - b. \end{aligned} \tag{3.17}$$

According to Equation 3.3, the adapted mean of an image-specific GMM is the sum of the feature vectors in the image, weighted by the corresponding posterior. Therefore,

$$\begin{aligned} f(Z) &= \sum_t \alpha_t \sum_{k=1}^{K} \frac{w_k}{2} \Sigma_k^{-1} \frac{1}{n_k} \sum_{j=1}^{H} Pr(k|z_j) z_j \bullet \mu_k^t - b. \\ &= \sum_{j=1}^{H} \left\{ \sum_{k=1}^{K} \frac{1}{n_k} Pr(k|z_j) z_j \bullet \frac{w_k}{2} \Sigma_k^{-1} \sum_t \alpha_t \mu_k^t \right\} - b. \end{aligned} \tag{3.18}$$

3.4.3 Quality bound

We define the "per feature vector contribution" as the contribution of each feature vector in a subarea to the confidence that this subarea is the concerned object. In particular, the "per feature vector contribution" is defined as in Equation 3.19.

$$W_j = \sum_{k=1}^{K} \frac{1}{n_k} Pr(k|z_j) z_j \bullet \frac{w_k}{2} \Sigma_k^{-1} \sum_t \alpha_t \mu_k^t. \quad (3.19)$$

Therefore, Equation 3.18 can be rewritten as Equation 3.20, showing that the quality function can be viewed as the sum of contributions from all involved feature vectors.

$$f(Z) = \sum_j W_j - b. \quad (3.20)$$

Given a test image, if we approximate the terms n_k with their values calculated on the whole image, the per feature vector contributions $W_j, j \in 1, ..., H$ are independent from the bounding box within the test image. This means that we can precompute W_j and evaluate the quality function on different rectangles by summing up those W_j that fall into the concerned rectangle.

We design a quality bound for the Gaussianized vector representation in a way similar to the quality bound for the histogram of keywords proposed in [75]. For a set of rectangles, the quality bound is the sum of all positive contributions from the feature vectors in the largest rectangle and all negative contributions from the feature vectors in the smallest rectangle. This can be formulated as

$$\begin{aligned} \hat{f}(\mathbf{R}) = & \sum_{W_{j_1} \in R_{max}} W_{j_1} \times (W_{j_1} > 0) \\ & + \sum_{W_{j_2} \in R_{min}} W_{j_2} \times (W_{j_2} < 0). \end{aligned} \quad (3.21)$$

where R_{max} and R_{min} are the largest and the smallest rectangles.

We demonstrate that Equation 3.21 satisfies the conditions of a quality bound for the branch-and-bound search scheme defined in Section 3.4.1.

First, the proposed $\hat{f}(\mathbf{R})$ is an upper bound for all rectangles in the set $\mathbf{R}$. In particular, the quality function evaluated on any rectangle R can be written as the sum of postive contributions and negative contributions from feature vectors in this rectangle,

$$\begin{aligned} f(R) &= \sum_{W_{j_1} \in R} W_{j_1} \times (W_{j_1} > 0) \\ &\quad + \sum_{W_{j_2} \in R} W_{j_2} \times (W_{j_2} < 0). \end{aligned} \tag{3.22}$$

Obviously, given a rectangle set $\mathbf{R}$, the first term in Equation 3.22 is maximized by taking all the positive contributions from the largest rectangle in the set. The second term in Equation 3.22 is negative and its absolute value can be minimized by taking all the negative contributions in the smallest rectangle.

Second, when the rectangle set $\mathbf{R}$ contains only one rectangle, $R_{min} = R_{max} = R$. Equation 3.21 equals Equation 3.22,

$$\hat{f}(\mathbf{R}) = f(R).$$

This quality bound defined by Equation 3.21 is used in the branch-and-bound scheme discussed in Section 3.4.1 to achieve fast and effective detection and localization. Note that since the bound is based on sum of per feature vector contributions, the approach can be repeated to find multiple bounding boxes in an image, after removing those features claimed by the previously found boxes. This avoids the problem of finding multiple non-optimal boxes near a previously found box.

Note that estimating W_j in Equation 3.19 involves no more computation than the calculation in a binary classifier using the Gaussianized vector representation of the whole image. To further expedite the localization, we can use two integral

images [91] to speed up the two summations in Equation 3.21 respectively. This makes the calculation of $\hat{f}(\mathbf{R})$ independent from the number of rectangles in the set $\mathbf{R}$.

3.4.4 Incorporating variation-normalization

To further improve the discriminating capability of the Gaussianized vector representation in the localization problem, we incorporate the normalization approach in Section 3.2. In particular, this involves the following modifications of the proposed efficient localization system.

First, the SVM is trained using kernels with normalization against within-class variation. In particular, Equation 3.10 is used instead of Equation 3.8.

Second, Equation 3.16 is replaced by Equation 3.23 to suppress the subspace that corresponds to the most within-class variation when evaluating the quality of the candidate regions.

$$f(Z) = g(Z) = \sum_t \alpha_t \phi(Z)^T (I - VCV^T)\phi(Z_t) - b. \tag{3.23}$$

Third, the per feature vector contribution function in Equation 3.19 needs to be revised accordingly.

Let us denote

$$P = \begin{bmatrix} \sqrt{\frac{w_1}{2}}\Sigma_1^{-1/2} & & \\ & \ddots & 0 \\ 0 & & \sqrt{\frac{w_K}{2}}\Sigma_K^{-1/2} \end{bmatrix} \tag{3.24}$$

$$H^t = [H_1^t; \cdots; H_K^t] \tag{3.25}$$

$$= P(I - VCV^T)\phi(Z_t), \tag{3.26}$$

$$\tag{3.27}$$

where H^t summarizes information from the t^{th} training image.

With Equations 3.20, 3.23 and 3.24, it can be shown that the per feature vector contribution function can be written as in Equation 3.28.

$$W_j = \sum_{k=1}^{K} \sum_{t} \alpha_t H_k^t \bullet \frac{1}{n_k} Pr(k|z_j) z_j. \quad (3.28)$$

3.5 Video Categorization Experiments

Our video event detection experiments are conducted over the large TRECVID 2005 video corpus as in [47], with shot boundaries provided.

3.5.1 Dataset and metric

As in [28], the following ten events are chosen from the LSCOM lexicon [97, 98, 47, 99]: *Car Crash, Demonstration Or Protest, Election Campaign Greeting, Exiting Car, Ground Combat, People Marching, Riot, Running, Shooting*, and *Walking*. They are chosen because these events are relatively frequent in the TRECVID data set [98] and are intuitively recognizable from visual cues. The number of video clips for each event class ranges from 54 to 877. When training the SVM for each event, we use the video clips from the other nine events as the negative samples. We randomly choose 60% of the data for training and use the remaining 40% for testing, with the same configurations as in [28, 47].

We use non-interpolated average precision (AP) [100, 2] as the performance metric, which is the official performance metric in TRECVID. It reflects the performance on multiple average precision values along a precision-recall curve. The effect of recall is also incorporated when AP is computed over the entire classification result set. Mean average precision (MAP) is defined as the mean of APs over all ten events.

3.5.2 Results

Temporally Aligned Pyramid Matching (TAPM) is the best reported algorithm for video event recognition in unconstrained news video [47]. We also got the result by histogram-of-keywords representation with SVM classification. Table 3.1 summarizes the experiment results for different algorithms. Note that: 1) TAPM-1 is the TAPM algorithm with same weights for all the three levels; 2) TAPM-2 refers to the TAPM algorithm with different weights for the three levels; 3) Hist+SVM refers to histogram-of-keywords representation with SVM classification; 4) Kernel+NN is the algorithm based on the Gaussianized vector and the nearest neighbor classifier; 5) Kernel+SVM means the Gaussianized vector kernel with SVM classification; 6) Kernel+WCCN refers to the nearest centroid algorithm using the Gaussianized vector with WCCN; and 7) WCCN&SVM refers to the algorithm based on the fusion of two classifiers based on the Gaussianized vectors, as presented in Section 3.3.3. The last row, referred to as mean AP, is the mean of APs over ten events. From all these results, we can have a set of interesting observations:

1. The mean average precision is boosted from the best reported 38.2% in [47] to 60.4% based on our new framework with straightforward classifier fusion.

2. For the video event of *Election Campaign Greeting*, the average precision is dramatically increased from the 13.9% to 94.8%.

3. The fusion of the two classifiers can generally further improve the average precision compared with the single classifier individually.

4. Our proposed framework is outperformed by the TAPM algorithm on detecting the video event of *Exiting Car*. A possible explanation is that our framework does not explicitly model temporal information, and the video event of *Exiting Car* heavily depends on the temporal contextual information.

5. The components of the Gaussianized vector representation (compared with histogram-of-keywords), suppressing within-class variance (WCCN), and SVM all contribute to the whole system, and the best result is achieved based on the integration of all.

6. The best setting of "GV+WCCN+SVM" performs significantly better than either "TAPM-1" or "TAPM-2" according to the Friedman's test ($p = 0.01$).

More details of the performance are presented using confusion matrices as in Figure 3.2. The mean average precision and the overall recognition accuracy are also presented in the titles in this figure.

From these confusion matrices, we observe that: 1) when evaluated by the confusion matrices, the fusion of classifiers again improves the recognition accuracy; and 2) the better the overall recognition accuracy, the greater the possibility that the video event of *Shooting* is mis-recognized; and a possible explanation is that the event of *Shooting* is visually very similar to the event of *Ground Combat*, and cannot benefit from the improved discriminating capability that dramatically improves the accuracy of most other events.

For video event recognition, the boundaries of the video clip are often ambiguous, and also the frame rate of the video clip may vary. A good algorithm should be robust to these factors, and hence a set of experiments are presented to evaluate the algorithmic robustness to these factors. In these experiments only a random portion of the frames within each video clip are used to construct the Gaussianized vector, with other aspects of the video event recognition framework unchanged.

The detailed experimental results are shown in Figure 3.3, with nine configurations using percentages of frames as 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% respectively. From these results, we can see that our system is robust to the variation of boundaries and the frame rates of video clips. In particular, even when only 20% of the frames are used, our result (55.3%) still outperforms

Table 3.1: Average precision (%) of video events by different algorithms.

Event Name	TAPM-1 [47]	TAPM-2 [47]	Hist+SVM	Kernel+NN	Kernel+SVM	Kernel+WCCN	WCCN&SVM
Car Crash	51.1	51.0	33.0	33.5	39.7	46.5	**53.3**
Demonstration	23.6	23.6	38.2	38.3	49.3	48.5	**50.1**
Election Campaign	13.9	13.7	82.5	79.2	92.6	**94.8**	94.4
Exiting Car	**50.7**	50.1	22.1	31.5	35.2	33.9	38.1
Ground Combat	44.2	44.1	68.1	58.2	71.4	72.8	**73.4**
People Marching	25.8	25.8	70.0	67.7	75.8	76.9	**78.7**
Riot	22.7	22.9	16.9	**30.9**	24.9	25.4	27.7
Running	86.7	86.6	88.1	89.3	91.4	89.9	**91.9**
Shooting	10.4	9.9	18.0	20.0	21.9	22.7	**23.1**
Walking	52.4	52.8	52.6	59.3	73.3	66.5	**73.8**
Mean AP	38.2	38.1	49.0	50.8	57.6	57.8	**60.4**

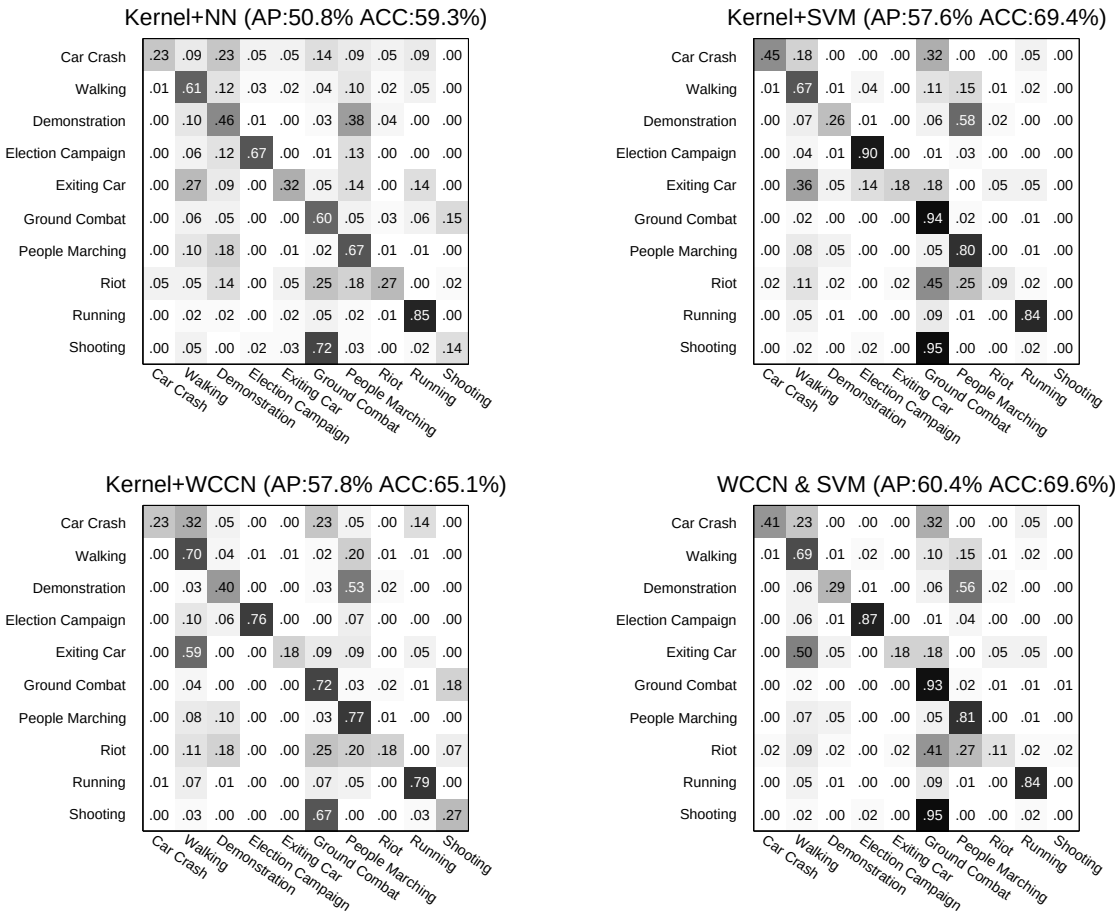

Figure 3.2: Confusion matrices for different methods based on the Gaussianized vector kernel.

the best result (38.2%) reported in [47]. We do point out that these frames are randomly sampled.

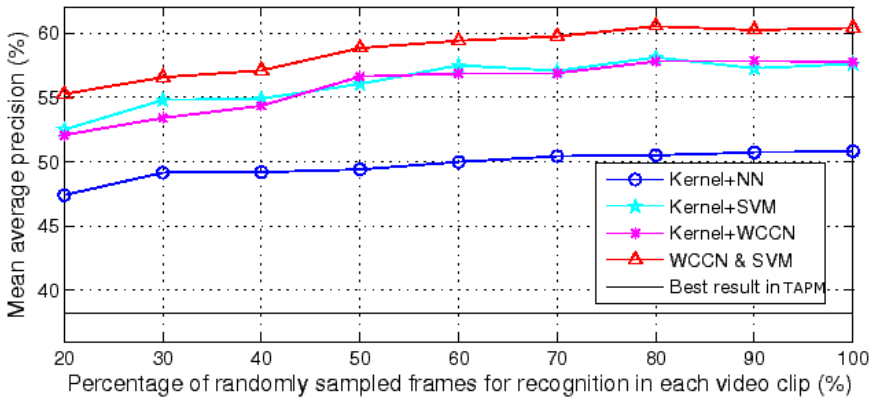

Figure 3.3: Mean average precision by different algorithms using randomly sampled subsets of the video frames.

3.6 Video Localization Experiments

We carry out object localization experiments using the proposed efficient object localization approach based on the Gaussianized vector representation. We compare the detection performance with a similar object localization system based on the generic histogram of keywords. In addition, we demonstrate that the proposed within-class variance normalizing approach can be effectively incorporated in object localization based on the Gaussianized vector representation.

3.6.1 Dataset

We use a multi-scale car dataset [101] for the localization experiment. There are 1050 training images of fixed size 100×40 pixels, half of which exactly show a car while the other half show other scenes or objects. Since the proposed localization approach has the benefit of requiring no heuristics about the possible locations and sizes of the bounding boxes, we use a test set consisting of 107

Figure 3.4: Sample images in the multi-scale car dataset.

images with varying resolution containing 139 cars in sizes between 89×36 and 212×85. This dataset also includes ground truth annotation for the test images in the form of bounding rectangles for all the cars. The training set and the multi-scale test set are consistent with the setup used in [75].

A few sample test images of the dataset are shown in Figure 3.4. Note that some test images contain multiple cars and partial occlusion may exist between different cars as well as between a car and a "noise" object, such as a bicyclist, a pedestrian or a tree.

3.6.2 Metric

The localization performance is measured by recall and precision, the same way as in [101] and [75]. A hypothesized bounding box is counted as a correct detection if its location coordinates and size lie within an ellipsoid centered at

the true coordinates and size. The axes of the ellipsoid are 25% of the true object dimensions in each direction. For multiple detected bounding boxes satisfying the above criteria for the same object, only one is counted as correct and the others are counted as false detections.

3.6.3 Gaussianized vectors

The feature vectors for each image are extracted as follows. First, square patches randomly sized between 4×4 and 12×12 are extracted on a dense pixel grid. Second, an 128-dimensional SIFT vector is extracted from each of these square patches. Third, each SIFT vector is reduced to 64 dimensions by principal component analysis. Therefore, each image is converted to a set of 64-dimensional feature vectors.

These feature vectors are further transformed into Gaussianized vector representations as described in Section 3.1. Each image is therefore represented as a Gaussianized vector. In particular, we carry out the experiment with 32, 64, 128 Gaussian components in the GMMs respectively.

3.6.4 Robustness to within-class variation

We identify the subspace that contains the undesirable within-class variation using the eigen analysis method in Section 3.2. In particular, the subspace consists of the top 100 dimensions, out of all the dimensions of the Gaussianized vectors, that are to be suppressed in the calculation of the kernels.

3.6.5 Results

To keep the setting the same as in [75], we search each test image for the three best bounding boxes, each affiliated with the quality function score. In particular, the branch-and-bound search scheme is applied to each test image three

times. After each search, those features claimed by the found boxes are removed as discussed in Section 3.4.1.

The precision-recall curves are obtained by changing the threshold on the quality function score for the found boxes. The equal error rate (EER) equals $1 -$ F-measure when precision equals recall. As the threshold is lowered, more detections out of the top three bounding boxes in each image are accepted.

The precision-recall curves and the EER are presented in Figure 3.5 and Figure 3.6 respectively. "G-n" denotes the result using n components in the Gaussianized vector representation. The suffix "N" means the within-class normalization. "Histogram" denotes the performance using the generic histogram-of-keywords approach by Lampert et al. We compare the results with a localization system using the same banch-and-bound scheme, but based on the generic histogram of keywords with 1000 entry codebook generated from SURF descriptors at different scales on a dense pixel grid [75].

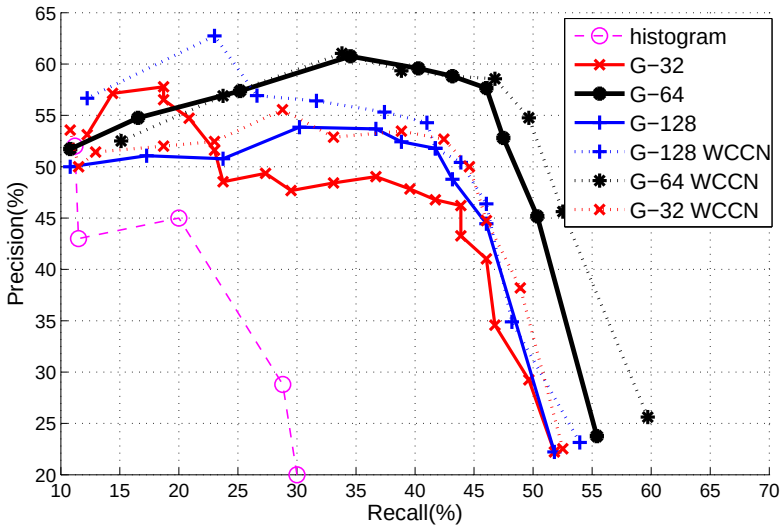

Figure 3.5: Precision-recall curves for multi-scale car detection.

We can see that the Gaussianized vector representation outperforms the histogram of keywords in this multi-scale object detection task. In particular, using 64 Gaussian components gives the best performance. In general, normalizing

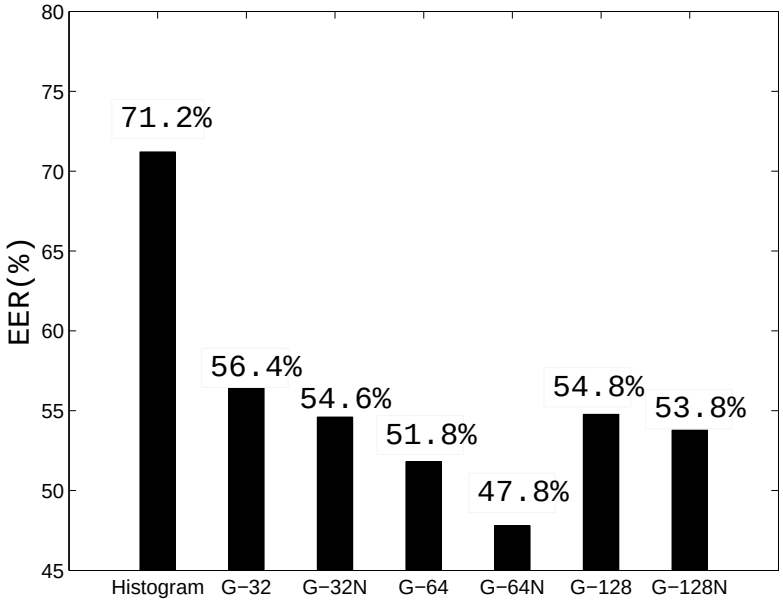

Figure 3.6: Equal error rates for multi-scale car detection.

against within-class variation further improves the system.

Figure 3.7 presents a few examples of correct detection and erroneous detection using 64 Gaussian components. Each test image is accompanied by a "per-feature-contribution" map. Negative and positive contributions are denoted by blue and red, with the color saturation reflecting absolute values. The quality function evaluated on a bounding box is the sum of all the per-feature-contributions, as discussed in Section 3.4.

The examples of correct detection demonstrate that the system can effectively localize one or multiple objects in complex backgrounds.

The three examples of erroneous detection probably occur for different reasons: 1) The car is a bit atypical, resulting in fewer features with highly positive contributions. 2) The two cars and some ground texture form one rectangle area with highly positive contributions, different from the two separate bounding boxes in the ground truth. 3) The car is highly confusable with the background, resulting in too many highly negative contributions everywhere, preventing any

rectangle to yield a high value for the quality function.

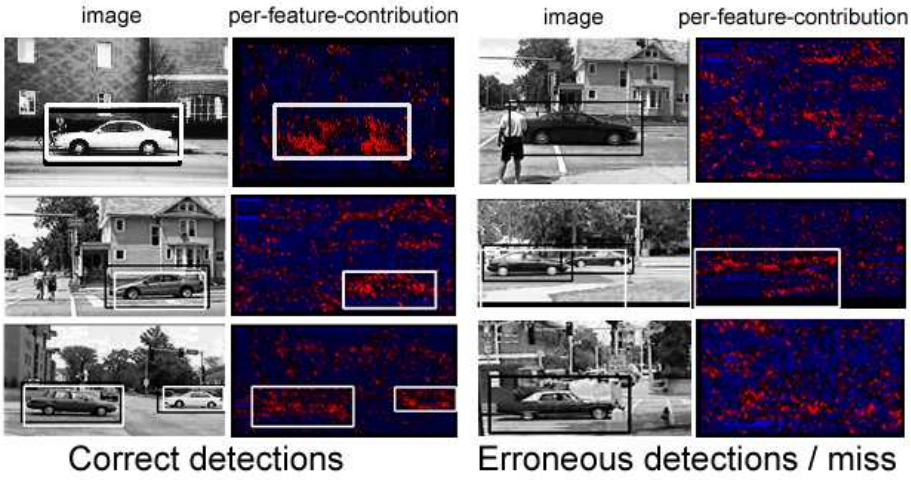

Figure 3.7: Examples of good and bad localization based on the Gaussianized vector representation. (The black and the white bounding boxes in the images are the ground truth and the hypotheses respectively.)

CHAPTER 4

IMPROVING ACOUSTIC EVENT DETECTION USING VISUAL CUES

Various audio-visual integration strategies have been proposed. In particular, [36] classifies them into three categories. The first is early integration which extracts feature vectors from both audio and visual observations and concatenates them into one feature vector sequence for use in one model with the same structure as for one modality. The second is late integration, which extracts feature vector sets separately and uses two sets of models generating reliability weights to be combined across modalities. This is also referred to as decision fusion or separate identification. The third is intermediate integration, e.g., product HMM, coupled HMM.

Recently, incorporating both audio and visual information for AED has been demonstrated as an effective approach to improve the performance and robustness over the audio-only systems [43, 12, 44]. However, these works either leverage on specific visual object detectors, usually requiring hand-labeled training data, or expect dominance or strong prior of the visual cues in the recorded video, sometimes impossible for realistic applications.

Leveraging additional visual cues for audio signal analysis has been explored in other applications, such as speech recognition [45] and person identification [46]. In particular, the multi-stream HMM and the couple HMM (CHMM) are two effective models for audio-visual fusion. While audio-visual event detection shares a lot of challenges with audio-visual speech recognition, they differ in multiple ways: First, the visual cues for general acoustic event detection can be much less constrained: there is no consistent visual region, such as the mouth in audio-visual speech processing, in which all the event information is embed-

ded. Second, the synchrony and asynchrony between the two modalities is not governed by a well constrained mechanism, such as human speech articulation. For example, key jingling presents mostly simultaneous audio and visual footprints, but we can observe a person move before or after s/he makes the foot-step sound, or a door start moving before making a slamming sound, the asynchrony being more arbitrary than what is observed in audio-visual speech. It is not yet studied whether the audio-visual models in speech processing can be effectively applied in audio-visual event modeling to improve acoustic event detection.

In this chapter, we study using a generalizable visual representation to improve acoustic event detection, via different audio-visual synchrony and asynchrony modeling. In particular, a combination of optical flow and overlapping spatial pyramid histograms characterizes the visual cues, which can be non-dominant in the recorded video. Compared with more task-specific alternatives [43], the proposed visual features have the merit of requiring minimum labeling efforts: No extra labels are required other than the event onset/offset timestamps used for audio-only modeling. We propose applying multi-stream HMMs for synchronized audio-visual event modeling and coupled hidden Markov models [21, 102] for more flexible modeling allowing asynchrony.

Acoustic event detection and classification experiments are performed on meeting room data with eleven general non-speech acoustic events. With the proposed visual representation and multi-modal modeling, the visual cues, often local and subtle in the images, are shown to consistently improve both classification and detection accuracy of the concerned events. All the experiments use the video associated with the audio as the only extra data resource, requiring no additional labeling.

The organization of this chapter is as follows. Section 4.1 presents the generalizable visual features adopted in this work, in particular the overlapping spatial pyramid histograms based on optical flow. Section 4.2 discusses the audio-visual modeling methods, in particular the multi-stream HMM and the coupled HMM. Section 4.3 presents the experimental results on audio-visual event classification

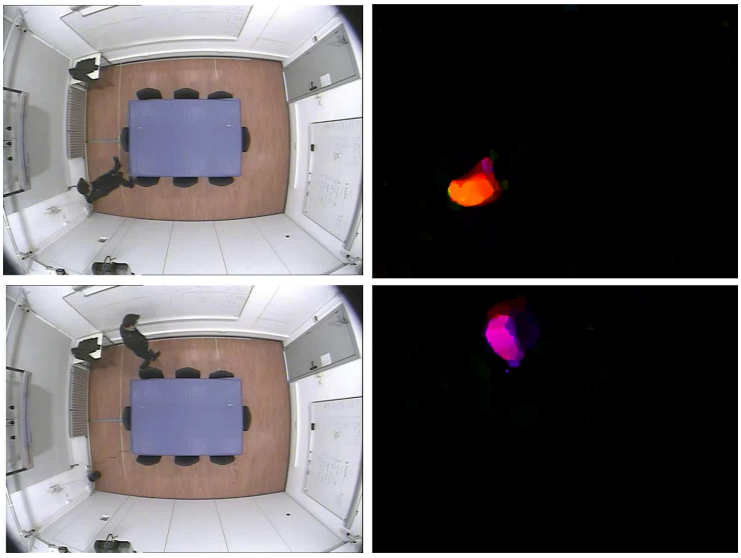

Figure 4.1: (Left) An example of “foot step” in the overhead camera; (Right) the corresponding optical flow for each image, where hue and intensity indicate direction and magnitude.

and detection.

4.1 Generalizable Visual Features for AED

Previous literature [43] reported using ad-hoc visual detectors to generate visual features for the purpose of improving event detection. However, training these detectors requires expensive labeling efforts, usually at least bounding boxes of the concerned objects. Moreover, these detectors are task-specific. Alternatively, we explore using visual features that do not require such training and data labeling, and are not task-specific, i.e. generalizable.

In this work, we propose using a combination of optical flow and overlapping spatial pyramid histograms to characterize the visual cues in the acoustic events.

The visual cues of the non-speech audio-visual events are mostly related to

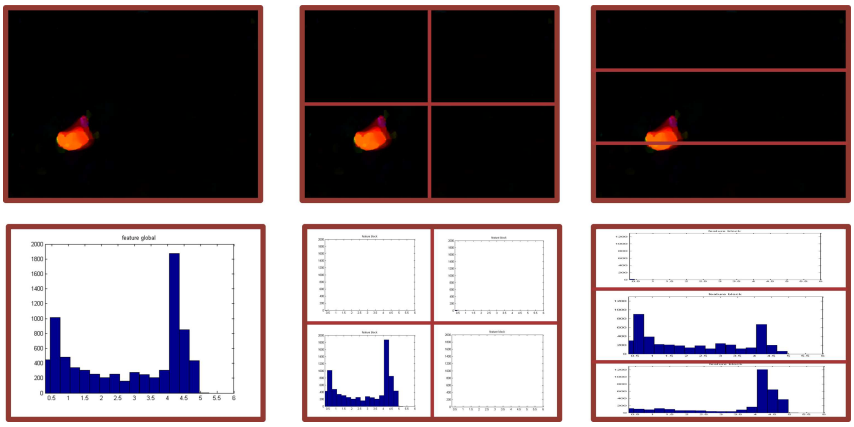

Figure 4.2: Optical flow based overlapping spatial pyramid histograms for a footstep event: (first row) spatial pyramid arrangement and optical flow magnitude; (second row) optical flow magnitude histogram in each corresponding block.

motion. We propose using visual features based on optical flow between consecutive frames to capture the movement information. We utilize a highly efficient algorithm on variational methods utilizing a GPU [103] to calculate the optical flow, i.e. the horizontal and vertical movement for each pixel. Fig. 4.1 illustrates the extracted optical flow for a "foot step" event.

The visual cues of the acoustic events have their spatial correlates: the spatial distribution sometimes, but not always, differs between the different events and the background. Therefore, we define eight overlapping blocks from the whole image, including both the complete image and seven spatially local regions. The histograms of motion vector magnitude within all the blocks are employed as the video features [104]. We refer to this representation as the *overlapping spatial pyramid histograms*. Similar representation was successfully used for kernel estimation in general image scene categorization [105], which shares the property that the visual cues are highly variant and sometimes localized.

An example of the proposed visual representation for a "foot step" event is illustrated in Fig. 4.2.

4.2 Multi-Modality Fusion for AED

We propose using multi-stream HMMs for synchronized audio-visual event modeling, and coupled hidden Markov models [21, 102] for more flexible modeling allowing asynchrony.

Different fusion methods have been explored for the audio and visual modalities [45]. First, feature fusion techniques include plain feature concatenation [106], feature weighting [107] and a data-to-data mapping of either one modality into the space of another or both modalities into a new common space [37]. Second, decision fusion provides a mechanism for capturing reliabilities of each modality by classifier combination. Third, intermediate fusion performs multimodal integration at a level between decision fusion and feature fusion. Intermediate integration strategies have been shown to outperform the early and late integration strategies in various applications [36].

Multi-stream HMMs and coupled HMMs are used as two intermediate fusion methods . The synchrony and asynchrony between the modalities are modeled by the hidden state transitions. Though both models have been successfully applied in audio-visual speech recognition [45], they have not been applied in improving general non-speech acoustic event detection.

4.2.1 Multi-stream hidden Markov models

In a two-stream HMM, the state-dependent emission of the audiovisual observation $o_{av,t}$ is governed by $P(o_{av,t}|S_t) = P(o_{a,t}|S_t)^{\lambda_{a,S_t,t}} P(o_{v,t}|S_t)^{\lambda_{v,S_t,t}}$ for all HMM states S_t, where $\lambda_{s,S_t,t}$ denotes the nonnegative stream weights, which models the stream reliabilities as a function of modality s, HMM state S_t and time t.

Multistream HMMs assume the state synchrony between audio cues and visual cues. Because of the simple topology, it is relatively easy to obtain robust estimation of the parameters.

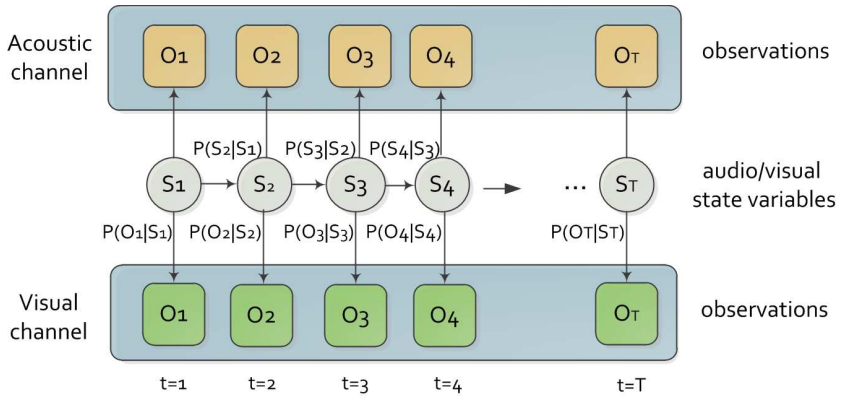

Figure 4.3: Hidden Markov model encoded as a dynamic Bayesian network.

Fig. 4.3 illustrates a two-stream HMM, where the transition probabilities are referred to as $P(S_t|S_{t-1})$. State observation distributions are referred to as $P(o_{av,t}|S_t)$. S_t is a multinomial random variable representing the state of the CHMM system variable at time t. Note, both the streams progress in a synchronous fashion.

4.2.2 Coupled hidden Markov models

The assumption of audio-visual state synchrony is not always satisfied. For example, in an object dropping event, the acoustic sound may not exist when the object is in motion, but only when the object stops dropping. Similarly, a door slamming sound occurs at the end of the door movement. Though the asynchrony between modalities can be alleviated by a larger local time window for each frame, a more flexible statistical model allowing asynchrony between the hidden state sequences for the two modalities is desired. In particular, the coupled HMM [21] has been introduced to address this issue for other applications.

This work uses coupled HMM to model modality asynchrony in audio-visual events. We select the transition-only coupled hidden Markov model (CHMM), in which different modalities are coupled through state transitions. The CHMM

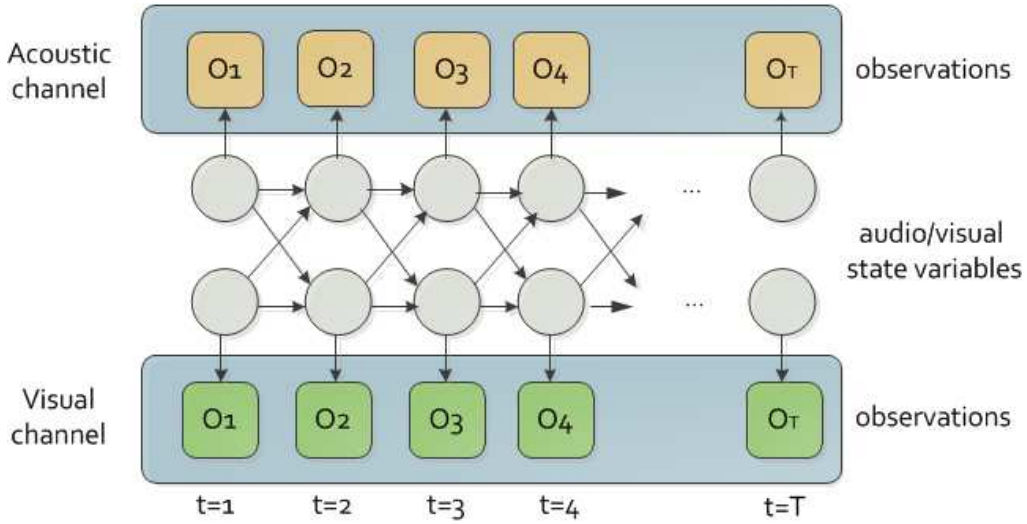

Figure 4.4: Audio-visual fusion using CHMM.

is capable of capturing both the synchronous and asynchronous inter-modality dependencies. CHMM has been shown to be an effective method in audio-visual speech recognition [102].

A CHMM can be viewed as two parallel rolled-out HMM chains coupled through cross-time and cross-chain conditional state transition probabilities. An n-chain CHMM has n hidden nodes in a time slice, each connected to itself and its nearest neighbors in the next time slice. In our task, we use a 2-chain CHMM for audio-visual modeling, as shown in Fig. 4.4, where circular nodes in each slice are the multinomial state variables, square nodes in each slice represent the observation variables, and the directed links represent conditional dependence between nodes.

The state of the CHMM system in each time slice is jointly determined by the two multinomial state variable, each depending on its two parent states in the previous time slice. The configuration permits unsynchronized progression of the two chains while keeping the Markov property that a future state variable is conditionally independent of the past given the present state variables. Note that CHMM can be seen as a generalized multi-stream HMM.

Following a transformation strategy based on state-space mapping and param-

eter tying [102], we can convert a CHMM to an equivalent HMM, whose hidden states each correspond to the state of the system described by the CHMM. The number of hidden states in the equivalent HMM equals the number of possible combinations of states from both modalities. Fig. 4.5 illustrates a 2-chain CHMM with $Q_a = 3$ and $Q_v = 2$, where Q_a and Q_v are the numbers of audio and visual states respectively. For example, state 3 in the equivalent HMM corresponds to the CHMM state defined by audio state $q_a = 2$ and visual state $q_v = 1$. The modality-dependent observation probabilities corresponding to the same observation distribution in the original CHMM are tied and coded using the same tag. For example, the output densities modeling the visual stream in states 1, 3, 5 are tied and tagged as "V_1", because they correspond to $P(O_v|q_v = 1)$ in the CHMM.

In this work, we use a left-to-right non-skip HMM for each of the two modalities in the CHMM. The allowed state transitions in the equivalent HMM are derived from state space mapping. In particular, the audio and visual state progressions are allowed asynchrony of up to one state. For example, in the state diagram in Fig. 4.5, given state 1 ($q_a = 1, q_v = 1$) at present, in next time slice, q_a can either transit to $q_a = 2$ or stay in $q_a = 1$, and q_v can either transit to $q_v = 2$ or stay in $q_v = 1$. Hence, state 1 can either stay in itself or transit to CHMM state 2 ($q_a = 1, q_v = 2$) or state 3 ($q_a = 2, q_v = 1$) ,or state 4 ($q_a = 2, q_v = 2$).

For robust estimation of the CHMMs, we perform the CHMM training in two stages. In the first stage, the observation distributions for both modalities are initialized using simpler models. The initial simpler models can be a two-stream audio-visual HMM, which requires strict state synchrony between audio and visual modalities, or one audio-only HMM and one video-only HMM, which impose no explicit state correspondence between the two modalities. In the second stage, the audio and visual observation distributions from the multi-stream HMM or two single-modality HMMs are used to construct the CHMM-equivalent HMM. Additional parameter estimation iterations using the Balm-Welch algorithm are performed with this equivalent HMM.

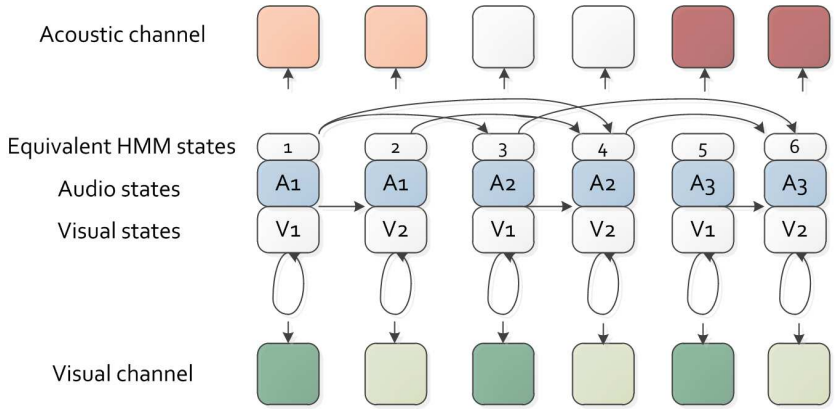

Figure 4.5: Converting a CHMM to an equivalent HMM by state-space mapping and parameter tying.

4.3 Audio-Visual Experiments

4.3.1 Dataset and setup

We use the audio-visual dataset collected by the Department of Signal Theory and Communications and the TALP Research Center of the Universitat Politecnica de Catalunya [35]. The database contains multimodal recordings of acoustic events (AEs) in a meeting room environment. The target events in this dataset include: knock (door, table), door slam, steps, chair moving, spoon (cup jingle), paper work (listing, warping), key jingle, keyboard typing, phone ringing/music, applause and cough. There are approximately 90 instances per event class for the whole dataset of six sessions (S01-S06). Among S01-S04, we use three sessions for training, and one for testing. All reported measures are averaged from four-fold cross validation. Additional two sessions (S05, S06) are used as the development set. We use the observations from a far field microphone and an overhead camera.

The audio in this dataset is quite clean. To make the task more realistic we add different levels of Gaussian white noise to the recorded audio, to illustrate the

performance of the different approaches at different noise levels. Perceptual linear prediction coding (PLP) coefficients are used as the audio features because of their effectiveness demonstrated in [5]. In particular, PLP coefficients, including 12 coefficients and the 0^{th} cepstral coefficient, are extracted from 30 ms Hamming windows with a temporal step of 20 ms. The delta and acceleration coefficients are computed and appended to the static PLP coefficients. Cepstral mean normalization is performed on each recorded session.

The visual features are obtained according to Section 4.1 using 20 bins for each histogram of optical flow magnitude. The concatenation of histograms from all blocks is projected into 40 dimensions using principal component analysis, retaining 98% of the total energy. These visual features are interpolated to match the 20 ms frame period of the audio features.

In this experiment, each multistream HMM or CHMM has 4 audio and 4 video states with stream weights tuned on the development data using coarse-to-fine grid search. For simplicity, the stream weights are time-invariant. The different methods are evaluated using classification accuracy and detection accuracy AED-ACC [1, 35].

4.3.2 CHMM training schemes

Initialization of the observation distributions in the CHMM is important, because of the high degree of freedom in the CHMM topology. As discussed in Section 4.2, we explore two different initialization schemes for CHMM, referred to as *CHMM*$_m$ and *CHMM*$_s$, in which the observation distributions of the CHMMs are initialized using multistream HMMs, or pairs of audio-only and video-only HMMs respectively.

The CHMMs parameters (the Gaussian means, covariance, mixtures weights, and the state transition probabilities) are further estimated with a few iterations using the Balm-Welch algorithm. We found in our pilot experiments that allowing estimation of all the CHMM parameters above is better than estimating any

subset of parameters above and using the initialized parameters for the rest.

4.3.3 Results

Table 4.1 and Table 4.2 present the classification and detection results using the proposed visual representation coupled with different audio-visual modeling methods as well as the audio-only and video-only models. In both detection and classification, the multistream HMM system consistently improves from the audio-only system as well as the video-only system for all SNR conditions studied in this work. Further, CHMM-based systems (CHMM$_s$ and CHMM$_m$) outperform the multistream HMM system in event detection for all SNR conditions. "CHMM$_m$" denotes the CHMM-based system initialized using multistream HMMs, while "CHMM$_s$" refers to the CHMM-based system initialized using audio-only and video-only HMMs.

We also performed event detection using original clean audio, the same condition studied in [35]. The proposed visual features and audio-visual modeling perform favorably compared to the best systems reported in [35]. These reference systems [35] leverage a person tracker, a laptop detector, a face detector, and a door activity estimator to capture the visual cues and optional localization information obtained from multiple microphones (denoted as "AV" and "AVL" in Table 4.2 respectively).

Table 4.1: Audio-visual event classification accuracy with different audio SNRs (% mean$\pm$standard error).

SNR	Audio-only	Video-only	Multistream	CHMM$_m$	CHMM$_s$
10dB	28.05$\pm$4.40	61.57$\pm$3.18	64.35$\pm$4.35	67.22$\pm$3.76	65.76$\pm$4.36
20dB	51.54$\pm$5.21	61.57$\pm$3.18	72.33$\pm$6.15	76.40$\pm$5.87	76.92$\pm$5.09
30dB	77.45$\pm$6.96	61.57$\pm$3.18	89.07$\pm$4.13	89.12$\pm$3.51	87.10$\pm$4.36

Fig. 4.6 shows the confusion matrices of event classification using the audio-only HMM, audio-visual multistream HMM, CHMM$_m$ and CHMM$_s$ systems,

Table 4.2: Audio-visual event detection accuracy with different audio SNRs (% mean$\pm$ standard error).

SNR	Audio	Video	Multistream	CHMM$_m$	CHMM$_s$
10dB	26.73$\pm$6.99	45.22$\pm$2.22	45.45$\pm$3.04	50.47$\pm$2.97	48.35$\pm$2.33
20dB	47.96$\pm$6.03	45.22$\pm$2.22	63.74$\pm$3.78	65.89$\pm$3.98	66.28$\pm$3.95
30dB	69.35$\pm$5.26	45.22$\pm$2.22	78.55$\pm$4.13	79.50$\pm$2.71	79.54$\pm$2.27
clean	87.54$\pm$2.99	45.22$\pm$2.22	90.57$\pm$2.07	91.85$\pm$2.11	90.79$\pm$2.97
clean	"AV" [35]	85	"AVL" [35]	86	

averaged across audio SNRs 10 dB, 20 dB and 30 dB. Using the proposed generalizable visual features with the multistream HMM or the CHMM boosts classification accuracy for most event classes compared to the audio-only system. The more flexible CHMM-based systems (CHMM$_s$ and CHMM$_m$) further improve classification of some events, such as kn: knock (door, table) and co: cough from the multistream HMM system.

To verify that the audio-visual state asynchrony allowed by the CHMM systems is utilized, we examine the state sequences found by the Viterbi decoding. The percentages of observation frames claimed by the CHMM states defined by an asynchronous pair of audio and video states are 65.9% for CHMM$_s$, and 65.8% for CHMM$_m$ respectively. Note that the multistream HMM system assigns all frames to states that are defined by synchronous audio and visual states. We do notice that the difference between the multistream HMM system and the CHMM systems is not very large. We believe part of the reason is that there is much asynchrony between the two modalities that exists beyond the one state asynchrony allowed in the current model. For example, for some asynchrony, the audio-visual cues might not overlap temporally at all.

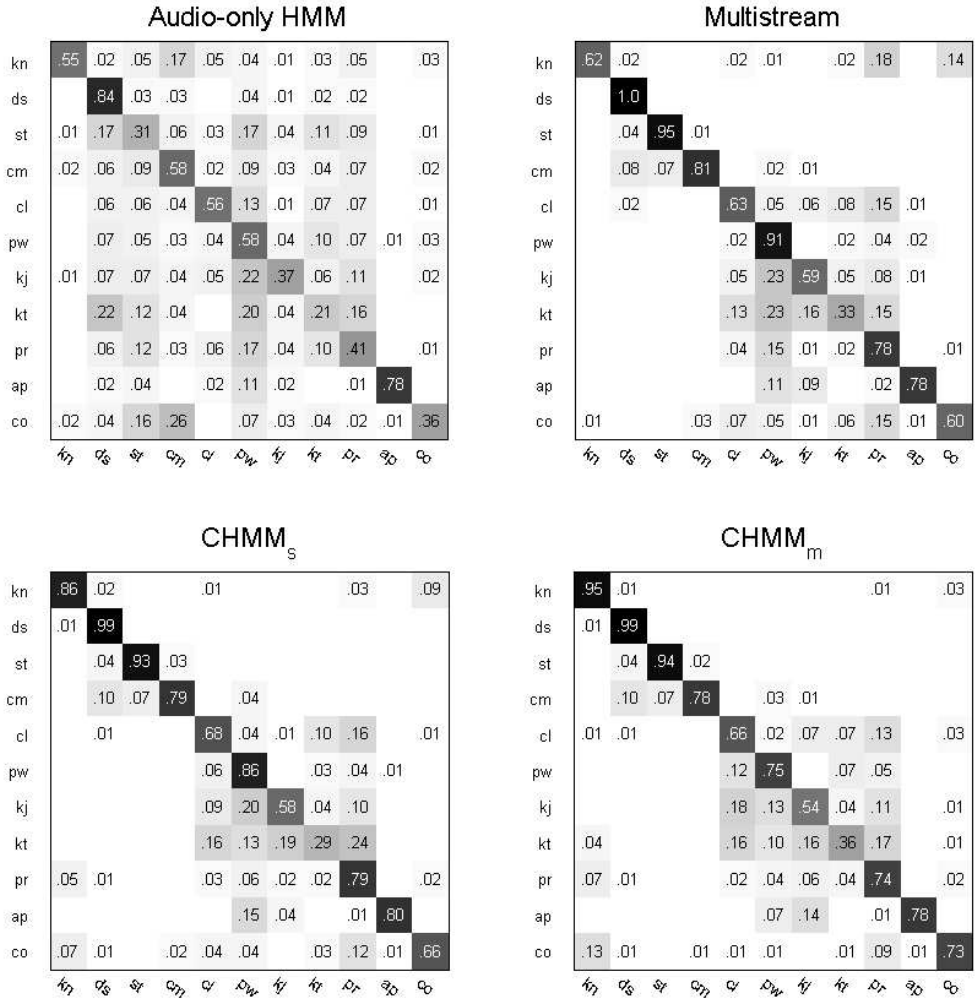

Figure 4.6: Confusion pattern for event classification based on audio-only HMM, audio-visual multistream HMM, CHMM$_m$ and CHMM$_s$.

CHAPTER 5

CONCLUSION AND DISCUSSION

This dissertation focuses on audio-visual modeling methods that can be easily adapted to related tasks. It is shown that these methods can effectively model less-constrained real-world data and deliver state-of-the-art results in acoustic event detection using CLEAR 2007 AED Evaluation data and video event detection using Trecvid video data.

Some related approaches not studied in this dissertation include: modeling through lower-level semantic concept detectors, pinpointed problem-specific methods, explicit alignment modeling between different samples. While these unexplored methods have their merits, this dissertation shows that for some applications, it is possible to deliver comparable, even superior, performance using our methods that usually require less training labeling efforts.

In this chapter, I summarize the studies in Chapters 2, 3 and 4 with discussion, and present possible future work following this dissertation.

5.1 Audio Modeling

On acoustic event detection, I present system architectures improved from our state-of-the-art HMM-based baseline system, designed for better acoustic event detection. Inspired by advances in speech recognition, a tandem connectionist-HMM approach for AED is proposed to combine the sequence modeling capabilities of the HMM with the high-accuracy context-dependent discriminative capabilities of an artificial neural network trained using the minimum cross entropy criterion. An SVM-GMM-supervector approach is designed using noise-

adaptive kernels to approximate the KL divergence between feature distributions in different audio segments, providing complimentary information that helps refine the Viterbi decoding output of the tandem models.

The interaction between speech and non-speech is an important topic not studied in this dissertation. One application for effective non-speech acoustic event detection is to improve speech recognition performance in realistic environments. The acoustic event models can be used to improve the limited background/noise models used in most speech recognition systems. Particularly, with more ubiquitous deployment of speech recognition systems, the capability of identifying non-speech events as noise will be essential to the effectiveness of many processes involved in real-word speech recognition applications, including speech transcription as well as speaker/channel/environment adaptation. On the other hand, effective modeling of human speech, as a major kind of background noise to acoustic event detection, can lead to more accurate detection of non-speech events. Future implementation of systems to study these issues will help answer the question to what extent realistic applications, i.e. speech recognition or non-speech event detection, can harness the benefit of explicit modeling of their interaction.

5.2 Image and Video Modeling

On visual cue modeling, I present the Gaussianized vector representation that works effectively for video event detection in realistic broadcast news data. Our system outperforms the best system in the previous literature using lower-level semantic concept detectors, which are not needed in this work. The Gaussianized vector representation establishes unsupervised correspondence between images or video clips of varying sizes, lengths and layouts. A normalization approach suppresses the within-class variation, by de-emphasizing the undesirable subspace in the Gaussianized vector representation kernels. An efficient object

localization approach is also developed for the Gaussianized vector representation, where the quality bound used in a branch-and-bound search scheme requires little extra computational cost, in addition to calculating the Gaussianized vector representation for the whole image, as in the classification problem.

One motivation of the Gaussianized vector representation is to effectively model realistic data that has hard-to-find and complicated correspondence between different samples. It is shown that for detecting video events in the broadcast news data, the approach in this dissertation outperforms the previous state-of-the-art that uses a set of lower-level concept detectors and explicitly temporal alignment modeling. However, it is plausible that many of the methods explored here can be combined with the above approaches for further improvement. In fact, some of the particular implementations in the experiments in this dissertation can be viewed as simple examples of such combination. For example, we can interpret the Gaussianized vector representation for video events based on SIFT detector and descriptors as a naive semantic concept detector (SIFT) combined with a robust video clip summarization approach (the Gaussianized vector). The intended lack of more explicit alignment modeling in this representation may also change to adapt to more structured image data. In particular, hidden states can be used to partition subparts of face images, as in our extension to the Gaussianized vector representation [95] in a face age estimation problem, beyond the scope of this dissertation. The Gaussianized vector representation has also been used for image segmentation, where each coherent region is modeled by this representation [108].

5.3 Audio-Visual Fusion

Given the challenges in acoustic event detection, I study using generalizable visual features to improve event detection via audio-visual intermediate integration. Optical flow based spatial pyramid histograms are used to represent

the highly variant visual cues for the acoustic events. This representation is demonstrated to significantly improve audio-only event classification and detection performance in systems based on multistream HMMs or coupled HMMs. Our systems perform favorably compared to previously reported systems [35] leveraging ad-hoc visual cue detectors and localization information obtained from multiple microphones.

The multistream HMMs assume strict temporal synchrony between the two modalities. The coupled HMMs allow hidden state asynchrony, but such asynchrony is usually limited to a few adjacent states. There are other techniques that have been used to integrate information from asynchronous data stream. In particular, canonical correlation analysis is an effective feature transform learning method that can be used to project features in different modalities such that their correlation is maximized in the projected spaces [109]. This learning method has been used to estimate a uniform shift or delay between two modalities. The asynchrony in the real-world events is however non-uniform. A recent extention to the above method, called Weakly-paired Maximum Covariance Analysis [110], introduces an explicit temporal alignment matrix that matches temporally local features from one modality with those from the other modality. This method iteratively updates this alignment matrix and the two projection matrices for maximized covariance between the aligned projections for both modalities. The Weakly-paired Maximum Covariance Analysis has been used to project single modality data into a subspace where it has maximum covariance with originally unaligned data from another modality only available in training. This method can be directly applied to improve acoustic event detection by projecting the acoustic features to a subspace where the projection has maximized covariance with the visual feature with learned alignment. We may further adapt the method to learn the alignment matrix during testing without changing the projections, and use both modalities that are projected and aligned in the multistream or coupled HMMs. I expect this will better model the audio-visual asynchrony in real-world events, and regard that as future work extended from this dissertation.

5.4 Human Performance

Human performance is far superior to machine performance in many pattern recognition problems. Most notably, humans easily outperform machines in speech recognition [111, 112], at different noise levels, vocabulary sizes, and with various availability of high-level grammatical information [113]. A recent paper [114] reviews the effective speech recognition by humans, particularly when overlapping with other sounds. It is also known that for recognition of any sound in a natural environment, humans can perceive a number of separate sound sources and identify their locations, pitch and timbre, even when they co-occur with other sounds. There has been continuing contention whether speech perception is special or shares the same mechanism as general sound perception [114]. Similarly, computer vision tasks such as face detection and recognition [115] find humans to excel in conditions most challenging to automatic algorithms, including various kinds of degration such as blur and noise.

In the pursuit of designing automatic machines that perform pattern recognition tasks, most researchers currently take the approach that machines are pure thinking devices that interface with the world during learning only in a very specific way: the machine is provided with data annotated by humans or an automatic labeler, e.g., the audio and video recordings and the corresponding event labels with onset and offset timestamps used in this dissertation. One of the human advantages, besides robust audio/visual signal processing, is the capability to actively interact with the world while comprehensively sensing the environment and applying our previous knowledge. Such capability applies to almost all kinds of human perception, particularly in tasks that humans perform regularly. Some artificial intelligence researchers believe that exploiting this in machine learning is a more promising strategy to push the performance closer to that of humans [116, 117].

These advantages exist in human perception of real-world audio-visual events as well, however, to a lesser extent as the tasks are more arbitrary and less in-

tegrated into human life. It is possible to train humans by presenting them examples of different recorded events. But that differs from the way humans learn how to perform familiar tasks such as speech recognition and face recognition, which we learn through interacting with the world. In practical applications such as surveillance and information retrieval, the input information is captured by specific sensors: should the task have never been exposed to a human being in prior interactive experiences, the limitation of the specific sensors and the lack of interaction might hinder the effectiveness of human learning.

Humans and machines differ in the way they tackle the pattern recognition problems fundamentally. The human level of semantic understanding is not achieved by even the state-of-the-art automatic pattern recognizers. Many automatic pattern recognition models either do not attempt to model the semantics in the data other than what is provided in the annotation of the training data, or do so to only a very limited extent. Even when some of them do try to harness the intrinsic semantics of the sensory data, the performance is often unsatisfactory, sometimes even worse than statistical methods without explicit consideration of such semantics. For example, lower-level visual semantic concept detectors [118] have been developed, and simulated results show that there need to be several thousand concept detectors for broadcast news video retrieval at a performance level comparable to modern text retrieval systems, which is far from human accuracy. Another example is the video event detection task studied in Chapter 3, where better semantic understanding through concept detectors and temporal alignment does not deliver superior performance. Between the sensory data, e.g., the observed acoustic signal, and the target semantic pattern, e.g., the events and their onsets and offsets, is the so-called semantic gap, which poses a major challenge to automatic pattern recognizers because of their limited capability to model previous knowledge and to apply that to new observations. This also relates to the question: for better pattern recognition performance, should machines mimic the human perception process at all, given the capabilities that we can currently build into the machines?

Though this dissertation does not further the understanding of human performance in real-world audio-visual event detection, relevant work in our research group provides insight into the challenge of the task to humans, if they are presented information in the form of recorded audio from a single microphone and video from a single camera. Particularly, two humans are asked to label the time and types of non-speech events in the AMI meeting room corpus [119]. Either with only audio recording available to them or with both audio and video available, each human transcriber disagrees on at least 50% of those events labeled by the other transcriber.

Human cognition has been and will continue to be inspiring advances in machine pattern recognition. It is also important to understand the human advantages and their applicability to different problems.

Besides, for real deployment of pattern recognition, there are practical considerations, such as privacy concerns, long operation hours and high operational cost, that make the "human" option less desirable. This adds to the utility of the technology studied in this dissertation, or automatic pattern recognition in general, even if they cannot match or surpass human performance.

5.5 Final Comments

With the emphasis on robust and generalizable modeling of realistic audio cues and visual cues, this work focuses on methods that can be readily applicable to other real-world audio visual modeling problems. There is much space for these methods to be further tailored for new specific problems, but I hope this work provides a good starting point, particularly when the following resources are limited: the expensive detailed annotation for training data, such as those needed to train ad-hoc lower-level detectors, the efforts of pinpointing specific cues for different events, or the capability of effective explicit modeling of the alignment between highly variant events in the spatial, temporal and feature domains.

REFERENCES

[1] A. Temko, "CLEAR 2007 AED evaluation plan and workshop, 2007." [Online]. Available: http://isl.ira.uka.de/clear07

[2] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[3] "Netcarity – ambient technology to support older people at home." [Online]. Available: http://www.netcarity.org

[4] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia & Expo*, 2005, pp. 1306–1309.

[5] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems," *IV Jornadas en Tecnologia del Habla*, November 2006.

[6] A. Martin and L. Mauuary, "Voicing parameter and energy based speech/non-speech detection for speech recognition in adverse conditions," in *Interspeech03*, 2003, pp. I: 3069–3072.

[7] F. Beaufays, D. Boies, M. Weintraub, and Q. Zhu, "Using speech/non-speech detection to bias recognition search on noisy data," in *International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. I: 424–427.

[8] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[9] D. Ellis, *Prediction-driven computational auditory scene analysis*. Ph.D. thesis, MIT, 1996.

[10] J. Pinquier, "Robust speech / music classification in audio document," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. III: 2005–2008.

[11] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[12] M. Naphade, A. Garg, and T. Huang, "Duration dependent input output Markov models for audio-visual event detection," in *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, August 2001, pp. 253 – 256.

[13] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 3, Jul. 2003, pp. III – 37–40.

[14] J. A. Smith, J. E. Earis, and A. A. Woodcock, "Establishing a gold standard for manual cough counting: video versus digital audio recordings," *Cough*, vol. 2, pp. 6:1–6, 2006.

[15] M. Baillie and J. Jose, "Audio-based event detection for sports video," *Lecture Notes in Computer Science*, vol. 2728, pp. 61–65, 2003.

[16] C. Muller, J. I. Biel, E. Kim, and D. Rosario, "Speech-overlapped acoustic event detection for automotive applications," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 2590–2593.

[17] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Superhuman multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, pp. 45 – 66, 2010.

[18] D. Zhang, D. Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005, pp. 611–618.

[19] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *IEEE International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007.

[20] P. Peursum, G. W. S. Venkatesh, and H. Bui, "Object labelling from human action recognition," in *Proceedings of IEEE International Conference on Pervasive Computing and Communications*, 2003, pp. 399–406.

[21] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997 IEEE Computer Society Conference on*, 1997, pp. 994–999.

[22] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005, pp. 166–173.

[23] J. Niebles, H. Wang, and L. Feifei, "Unsupervised learning of human action categories using spatial temporal words," in *British Machine Vision Conference*, 2006.

[24] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[25] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings. Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[26] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 9, pp. 1575 –1589, Sept. 2007.

[27] A. Bagdanov, L. Ballan, M. Bertini, and A. D. Bimbo, "Trademark matching and retrieval in sports video databases," in *Proceedings of the International Workshop on Multimedia Information Retrieval*, 2007, pp. 1575–1589.

[28] D. Xu and S. Chang, "Visual event recognition in news video using kernel methods with multi-level temporal alignment," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[29] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411 – 418, 2010.

[30] H. A. Rowley, S. Baluja, and T. Kanade, "Human face detection in visual scenes," in *NIPS 8*, 1996, pp. 875–811.

[31] S. M. Chu and T. S. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *ICSLP-2000*, 2000, pp. 747–750.

[32] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko, "Audiovisual speech recognition with articulator positions as hidden variables," in *Phonetic Sciences. International Congress of*, 2007, p. 297.

[33] D. A. Sadlier and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1225–1233, 2005.

[34] C. Canton-Ferrer, T. Butko, C. Segura, X. Giro, C. Nadeu, J. Hernando, and J. Casas, "Audiovisual event detection towards scene understanding," in *Computer Vision and Pattern Recognition Workshops, 2009. IEEE Computer Society Conference on*, June 2009, pp. 81 –88.

[35] C. Butko, C. Canton-Ferrer, C. Segura, X. Giro, C. Nadeu, J. Hernando, and J. Casas, "Improving detection of acoustic events using audiovisual data and feature level fusion," in *Proc. Interspeech*, 2009, pp. 1147–1150.

[36] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 854–866, July 2002.

[37] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 129–136.

[38] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *Proc. ASRU*, 2009.

[39] M. E. Sargin, Y. Yemez, E. Erzin, and M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[40] A. Temko and C. Nadeu, "Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. V: 505–508.

[41] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *Proceedings of the 1st International Evaluation Conference on Classification of Events, Activities and Relationships*, ser. CLEAR'06, 2007, pp. 311–322.

[42] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with AdaBoost feature selection," in *Classification of Events, Activities and Relationships Evaluation and Workshop*, 2007, pp. 345–353.

[43] T. Butko, A. Temko, C. Nadeu, and C. Canton, "Fusion of audio and video modalities for detection of acoustic events," in *Proc. Interspeech*, 2008, pp. 123–126.

[44] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang, "Audio-visual event recognition with application in sports video," in *Intelligent Multimedia Processing with Soft Computing*, ser. Studies in Fuzziness and Soft Computing, 2005, vol. 168, pp. 129–149.

[45] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306 – 1326, Sept. 2003.

[46] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, no. 5, pp. 449 – 480, 2004.

[47] D. Xu and S. Chang, "Video event recognition using kernel methods with multi-level temporal alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1985 – 1997, 2008.

[48] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543 – 1551, 2010.

[49] X. Zhuang, J. Huang, V. Libal, and G. Potamianos, "Acoustic fall detection using gaussian mixture models and gmm supervectors," in *International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 69–72.

[50] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. S. Huang, "Sift-bag kernel for video event analysis," in *Multimedia, 16th ACM International Conference on*, 2008, pp. 229–238.

[51] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Efficient object localization with gaussianized vector representation," in *Proceedings of the ACM Multimedia 1st International Workshop on Interactive Multimedia for Consumer Electronics*, 2009, pp. 89–96.

[52] P. Huang, X. Zhuang, and M. Hasegawa-Johnson, "Improving acoustic event detection using generalizable visual features and multi-modality modeling," 2011, accepted to *International Conference on Acoustics, Speech, and Signal Processing* 2011.

[53] B. Schòlkopf and A. Smola, *Learning with Kernels*. Cambridge, MA, US: MIT Press, 2002.

[54] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Transactions on Information Theory*, vol. 18, no. 3, pp. 363–378, 1972.

[55] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional hmm systems," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. III, 2000, pp. 1635–1638.

[56] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2006, pp. 97–100.

[57] G. Ratsch, T. Onoda, and K.-R. Muller, "Soft margins for adaboost," *IEEE Trans. on Signal Processing*, vol. 42, pp. 287–320, 2001.

[58] Y. Freund and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.

[59] D. Ellis and M. R. Gomez, "Investigations into tandem acoustic modeling for the aurora task," in *Proc. Eurospeech*. ISCA, 2001, pp. 189–192.

[60] D. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, 2001, pp. 517–520.

[61] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "The entropy of the articulatory phonological code: Recognizing gestures from tract variables," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1489–1492.

[62] X. Zhuang, H. Nam, M. Hasegawa-Johnson, L. Goldstein, and E. Saltzman, "Articulatory phonological code for word classification," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 2763–2766.

[63] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[64] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *Signal Processing, IEEE Transactions on*, vol. 39, no. 4, pp. 806 –814, Apr. 1991.

[65] X. Zhou, J. Navrátil, J. W. Pelecanos, G. N. Ramaswamy, and T. S. Huang, "Intersession variability compensation for language detection," in *International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4157 –4160.

[66] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[67] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[68] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, Dec. 1997, pp. 347 –354.

[69] J. Huang, X. Zhuang, V. Libal, and G. Potamianos, "Long-time span acoustic activity analysis from far-field sensors in smart homes," in *International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 4173–4176.

[70] M. Grassi, A. Lombardi, G. Rescio, P. Malcovati, A. Leone, G. Diraco, C. Distante, P. Siciliano, M. Malfatti, L. Gonzo, V. Libal, J. Huang, and G. Potamianos, "A hardware-software framework for high-reliability people fall detection," in *Proc: IEEE Sensors 2008*, 2008, pp. 1328 –1331.

[71] S. Ebadollahi, L. Xie, S. Chang, and J. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2006, pp. 881–884.

[72] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, no. 2, pp. 99–121, 2000.

[73] D. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Proceedings of IEEE International Conference on Computer Vision*, 1999, pp. 1150–1157.

[74] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions, a local svm approach," in *Proceedings of IEEE International Conference on Pattern Recognition*, 2004, pp. 32–36.

[75] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2008, pp. 1–8.

[76] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[77] H. Permuter, J. Francos, and I. Jermyn, "Gaussian mixture models of texture and colour for image database retrieval," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 3, April 2003, pp. III–569–72.

[78] F. Perronnin, C. Dance, G. Csurka, and M. Bressian, "Adapted vocabularies for generic visual categorization," in *European Conference on Computer Vision*, 2006.

[79] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying Discriminative Visual Codebook Generation with Classifier Training for Object Category Recognition," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2008.

[80] J. C. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders, "Kernel Codebooks for Scene Categorization," in *European Conference on Computer Vision*, 2008.

[81] A. Agarwal and B. Triggs, "Hyperfeatures-multilevel local coding for visual recognition," in *European Conference on Computer Vision*, 2006.

[82] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Oct. 2007, pp. 1 –8.

[83] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2008.

[84] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor, "Improving bag-of-keypoints image categorisation: Generative models and pdf-kernels." Department of Electronics and Computer Science, University of Southampton, Tech. Rep., 2005.

[85] D. Larlus and F. Jurie, "Latent mixture vocabularies for object categorization," in *British Machine Vision Conference*, 2006.

[86] F. Moosmann, B. Triggs, and F. Jurie, "Randomized clustering forests for building fast and discriminative visual vocabularies," in *NIPS*, 2007.

[87] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[88] A. O. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, May 2006, p. V.

[89] X. He and P. Niyogi, "Locality preserving projections," in *Proceedings of the Conference on Advances in Nerual Information Processing Systems 16*, 2003.

[90] Q. Zhu, S. Avidan, M. chen Yeh, and K. ting Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2006, pp. 1491–1498.

[91] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2001.

[92] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2005.

[93] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2005, pp. 886–893.

[94] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Regression from patch-kernel," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2008.

[95] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. Huang, "Face age estimation using patch-based hidden Markov model supervectors," in *Pattern Recognition, 2008 International Conference on*, 2008, pp. 1–4.

[96] X. Zhou, X. Zhuang, H. Tang, M. Hasegawa-Johnson, and T. S. Huang, "A Novel Gaussianized Vector Representation for Natural Scene Categorization," in *Pattern Recognition, 2008 International Conference on*, 2008, pp. 1–4.

[97] "LSCOM Lexicon Definitions and Annotations," 2006. [Online]. Available: http://www.ee.columbia.edu/dvmm/lscom/

[98] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, L.Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia Magazine*, vol. 13, no. 3, pp. 86–91, 2006.

[99] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 lscom semantic visual concepts," Columbia University ADVENT, Tech. Rep. 222-2006-8, March 2007.

[100] A. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.

[101] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

[102] S. M. Chu and T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[103] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in *Pattern Recognition (Proc. DAGM)*, Heidelberg, Germany, 2007, pp. 214–223.

[104] N. Ikizler, R. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms," in *Pattern Recognition, 2008. 19th International Conference on*, Dec. 2008, pp. 1 –4.

[105] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2169 – 2178.

[106] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," *D. G. Stork and M. E. Hennecke (Eds.), Speechreading by Humans and Machines. Berlin: Springer-Verlag*, pp. 461–471, 1996.

[107] P. Teissier, J. Robert-Ribes, and J. L. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 629–642, 1999.

[108] X. Zhou, J. Wang, F. Lv, and K. Yu, "Hierarchical image parsing for joint classification, detection and segmentation," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2010.

[109] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis; an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[110] C. H. Lampert and O. Krömer, "Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning," in *Proceedings of the 11th European conference on Computer vision: Part II*, 2010, pp. 566–579.

[111] M. Padmanabhan and M. Picheny, "Towards super-human speech recognition," in *ASR-2000*, 2000, pp. 189–194.

[112] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763 – 786, 2007.

[113] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[114] C. J. Darwin, "Listening to speech in the presence of other sounds," *Phil. Trans. R. Soc. B*, vol. 363, pp. 1011–1021, 2008.

[115] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948 –1962, Nov. 2006.

[116] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, no. 5504, pp. 599–600, Jan 2001.

[117] S. Levinson, W. Zhu, D. Li, K. Squire, R. Lin, M. Kleffner, M. McClain, and J. Lee, "Automatic language acquisition by autonomous robot," in *Int. Joint Conference on Neural Networks*, 2003.

[118] A. Hauptmann, R. Yan, and W.-H. Lin, "How many high-level concepts will fill the semantic gap in news video retrieval?" in *CIVR 2007 ACM International Conference on Image and Video Retrieval*, July 2007, pp. 9–11.

[119] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

AUTHOR'S BIOGRAPHY

Xiaodan Zhuang grew up in Shanghai, China. After receiving his B.S. degree from the Electronic Engineering Department in Tsinghua University, Beijing, in 2005, he joined the graduate program in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign (UIUC), where he received his M.S. degree in 2007.

Xiaodan has broad research interests and experiences in statistical pattern recognition, including speech recognition, non-speech audio event modeling, and image and video modeling. He has coordinated or contributed to the award-winning UIUC teams led by Prof. Thomas S. Huang and Professor Mark A. Hasegawa-Johnson in international evaluations related to non-speech acoustic event detection, multimedia retrieval, and audio-visual biometrics. He has also collaborated with linguists to incorporate prosodic and phonological theories into computational models for improved speech recognition or advanced understanding of speech pronunciation.