AUTOMATIC ACOUSTIC SIREN DETECTION IN TRAFFIC NOISE BY PART-BASED MODELS

Jens Schröder[1], *Stefan Goetze*[1], *Volker Grützmacher*[2], *Jörn Anemüller*[1,3]

[1] Fraunhofer IDMT / Hearing, Speech and Audio Technology, 26129 Oldenburg, Germany
[2] Adam Opel AG, 65423 Rüsselsheim, Germany
[3] University of Oldenburg, Dept. of Physics, 26111 Oldenburg, Germany

{jens.schroeder, s.goetze}@idmt.fraunhofer.de, dr.volker.gruetzmacher@de.opel.com, joern.anemueller@uni-oldenburg.de

ABSTRACT

State-of-the-art classifiers like hidden Markov models (HMMs) in combination with mel-frequency cepstral coefficients (MFCCs) are flexible in time but rigid in the spectral dimension. In contrast, part-based models (PBMs) originally proposed in computer vision consist of parts in a fully deformable configuration. The present contribution proposes to employ PBMs in the spectro-temporal domain for detection of emergency siren sounds in traffic noise, resulting in a classifier that is robust to shifts in frequency induced, e.g., by Doppler-shift effects. Two improvements over standard machine learning techniques for PBM estimation are proposed: (i) Spectro-temporal part ("appearance") extraction is initialized by interest point detection instead of random initialization and (ii) a discriminative training approach in addition to standard generative training is implemented.

Evaluation with self-recorded police sirens and traffic noise gathered on-line demonstrates that PBMs are successful in acoustic siren detection. One hand-labeled and two machine learned PBMs are compared to standard HMMs employing mel-spectrograms and MFCCs in clean and multi condition (multiple SNR) training settings. Results show that in clean condition training, hand-labeled PBMs and HMMs outperform machine-learned PBMs already for test data with moderate additive noise. In multi condition training, the machine learned PBMs outperform HMMs on most SNRs, achieving high accuracies and being nearly optimal up to 5 dB SNR. Thus, our simulation results show that PBMs are a promising approach for acoustic event detection (AED).

Index Terms— acoustic event detection (AED), part-based model (PBM), siren detection

1. INTRODUCTION

In daily traffic, emergency vehicles like ambulances, police cars and fire engines take a special role. While in action, they are prior to other motorists and cyclists and only partly have to stick to speed limits, traffic signs, red traffic lights etc. To alert other road users about their approaching, they use lights and sirens. Unfortunately, these lights and sirens might be missed leading to hazardous and life-threatening situations. Reasons for missing alerts can e.g. be physical deficiencies like hearing impairment or just simple distraction. Thus, technologies are researched which automatically detect emergency vehicles to warn road users and hence prevent accidents.

A couple of developments to detect emergency vehicles deal with special additional devices like infrared [1], ultra sonic [2] or radio [3–5] transmitter/receiver pairs [6]. For these technologies, the emergency vehicle has to be equipped with a transmitter broadcasting a particular signal. The road user to be alerted needs to have a corresponding receiver.

Few technologies use the existing warning signal, namely the acoustic siren. In Germany, the sound of the siren is defined by DIN14610 [7]. A siren has to consist of a low and a high tone with one repetition of this sound sequence. It is supposed to last for (3 ± 0.5) s. A pause between two sound sequences is not supposed to last longer than 0.8 s. The frequencies of the tones range between 360 Hz and 630 Hz with relative ratio of 1.333 between high and low tone. In rural environments, siren signals usually can be switched to lower frequency ranges than in urban areas. Anyhow, due to Doppler effects of approaching and vanishing vehicles, the frequency at a receiver can be outside the mentioned fundamental frequency range. Although siren signals may be slightly different for other countries, usually a clearly recognizable *melody* is defined in respective standards which makes the described methods applicable in general. For this paper, we will restrict the simulations and discussions to German emergency signals for simplicity reasons.

To automatically detect alarms, the tonality of such sounds can be exploited e.g. by searching for dominant peaks in spectrograms [1, 8]. Recent approaches estimate the pitch through the autocorrelation function [9–11].

Beritelli et al. [12] proposed a method using a standard acoustical pattern recognizer to detect sirens. Artificial neural networks (ANN) employed mel-frequency cepstral coefficients (MFCCs) that are well-known from automatic speech recognition (ASR). The ANN output was averaged over 400 ms windows. On their database of self recorded sirens from Italian emergency vehicles and traffic noise, they achieved an accuracy of 99% up to an signal to noise ratio (SNR) of 0 dB. However, it has not clearly been stated if Doppler effect was taken into account, i.e. if vehicle movements were considered.

In [13], alarm detection of unknown alarms was investigated. A common approach from ASR and a sinusoidal modeling were compared to each other. The database consisted of different types of alarms and background noises from the internet. All tests were done at 0 dB signal-to-noise ratio (SNR). Both systems were stated to perform poorly.

Since a German siren signal can be seen as a four-tone melody, the task of siren detection is similar to melody spotting. For melody spotting, Durey and Clements [14, 15] suggested to model each note by a hidden Markov model (HMM) [16]. Besides different mappings of spectrograms also MFCCs were tested. The database consisted of songs played on a keyboard, where each song was played several times to achieve variance in the data. MFCCs turned

out to achieve the highest accuracy (90%) among the tested features.

Siren signals often contain energy only in certain frequency bands (fundamentals and harmonics) and the first tone is lower in frequency than the second one, regardless of Doppler-effects, rural and urban modes etc. Thus, an algorithm that is able to deal with frequency shifts and is flexible in time would be beneficial for siren detection. In computer vision, tasks dealing with objects composed of several elements that are located anywhere in an image are very common and researched widely. Approaches exist that use these technologies for acoustic purposes. Ezzat and Poggio [17] proposed an modified bag-of-features (BOF) approach for word-spotting. Therein, 2-d patterns were extracted from mel-spectrograms. The relative positions within an event were stored. This led to a codebook of 2-d patches. For testing, the test sample was cross-correlated with all codebook patches in a region around the corresponding relative position. The highest scores were collected as feature vector inputs for a support vector machine (SVM) [18]. With only few training data, this approach outperformed classical HMM/MFCC combinations.

Schutte [19] adopted so called part-based models (PBM) [20] from computer vision for classification of isolated phonemes. PBMs consist of single parts that have a defined but deformable configuration. The configuration and part appearances are modeled flexibly by Gaussian distributions. In computer vision, PBMs are often used for face recognition since a face is composed of different parts (e.g. eyes, nose, etc.) but may be anywhere within an image. As features, edge-detectors were applied on spectrograms. On a small and preliminary test set, this led to promising results. Especially in noisy conditions, if only certain unimportant frequency bands were corrupted, higher accuracies were reached than using standard HMMs/MFCCs.

In this paper we propose to use of PBMs for siren detection since they provide the required flexibility in time and frequency. In the following, the PBMs are introduced in Sec. 2. Here, we describe an adjustment to the initialization of the semi-supervised learning algorithm known from the literature. Additionally, we propose a new, discriminative approach. The experimental setup including the database is described in Sec. 3. Results are presented in Sec. 4 and conclusions are drawn in Sec. 5.

2. PART-BASED MODELS

A part-based model (PBM) is a compounding of M parts in a flexible configuration. Thus, a PBM $\Theta = (A, S, E)$ can be defined by its part appearances $A = (a_1, ..., a_M)$, the relative positions of parts $S = \{s_{ij} | i, j = 1 \dots M\}$ and the actual connections between parts E. To estimate the likelihood $p(I|\Theta)$ of an image I belonging to a model Θ, a summation over each possible configuration L of Θ within I has to be calculated. The likelihood $p(I|\Theta)$ can be approximated by the likelihood of the configuration L^* of Θ that fits best into the image

$$p(I|\Theta) \approx \max_L p(L, I|\Theta) = p(L^*, I|\Theta). \tag{1}$$

The likelihood $p(L, I|\Theta)$ can be separated into a contribution of appearances and relative positions

$$p(L, I|\Theta) = p(I|L, A) \cdot p(L|S, E). \tag{2}$$

If the parts are independent of each other, the likelihoods can be factorized into a product of single appearances

$$p(I|L, A) = \prod_i p(I|l_i, a_i) \tag{3}$$

and pairwise positions

$$p(L|S, E) = \prod_{i,j} p(l_i, l_j|s_{ij}). \tag{4}$$

The likelihoods of appearances and relative positions can be modeled by probability functions like Gaussian distributions:

$$p(I|l_i, a_i) = \frac{1}{(2\pi)^{\frac{W}{2}} \cdot |\Sigma_i|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(w(l_i)-\mu_i)^T \Sigma_i^{-1} (w(l_i)-\mu_i)} \tag{5}$$

and

$$p(l_i, l_j|s_{ij}) = \frac{1}{2\pi \cdot |\Sigma_{ij}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\left((l_j-l_i)-\mu_{ij}\right)^T \Sigma_{ij}^{-1} \left((l_j-l_i)-\mu_{ij}\right)}, \tag{6}$$

where $w(l_i) \in I$ depicts an extracted part of size W from the image I at location l_i. μ_i and Σ_i represent the mean and the (diagonal) covariance matrix of appearance a_i and μ_{ij} and Σ_{ij} the mean and the covariance matrix of the relative position s_{ij}, respectively.

Felzenszwalb and Huttenlocher [20] developed an efficient matching algorithm for this kind of PBM. To learn the models, they proposed an expectation-maximization (EM) algorithm with an initial hand-labeled model. Another possibility to develop initial models for PBMs was proposed by Crandall and Huttenlocher [21]. Single part appearances were randomly extracted from the positive training data X that belonged to the created model class. The appearances were generalized by an EM. Single parts i and j were combined to pair models $\Theta_{ij} = (a_i, a_j, s_{ij})$. Therefore, the difference of the best locations l_i^* and l_j^* were averaged to gain the relative positions s_{ij}. The best $M - 1$ pair models with a common part r were combined to form 1-fan models Θ_r [21]. The model with the maximum likelihood on X was used as initial model:

$$\Theta_{\text{pos}} = \arg\max_{\Theta_r} p(X|\Theta_r). \tag{7a}$$

In this contribution, we change the extraction of appearances in the beginning. Instead of extracting at random points, the extraction is done at points of interest. The points of interest are detected by the Foerstner operator [22]. The Foerstner operator locates points where the derivation in an image is high. In a spectrogram, these are regions with high energy fluctuations between adjacent time-frequency units. To reduce the computational load, not all interest points are used. The locations of interest points of each image/spectrogram are clustered by kmeans into five regions. From each cluster, an interest point is selected randomly.

The proposed initial model by Crandall and Huttenlocher [21] is generative, i.e. only positive training samples were considered. Such a model may not be discriminative enough for all kinds of tasks. In contrast, a discriminative model is supposed to produce high scores on positive samples and low ones on negative ones. Thus, we substitute $p(X|\Theta_r)$ in Eq. (7a) by the ratio of the likelihoods of positive and negative data:

$$\Theta_{\text{neg}} = \arg\max_{\Theta_r} \frac{p(X|\Theta_r)}{p(Y|\Theta_r)}. \tag{7b}$$

Hence, by Eq. (7a) a generative initial model can be achieved whereas Eq. (7b) leads to a discriminative one. Both of these approaches will be evaluated in Sec. 4.

In the sense of acoustic processing, I can be any kind of spectro-temporal representation. Accordingly, a PBM for AED consists of spectro-temporal patches in a relative and flexible time-frequency configuration within a spectrogram.

3. EXPERIMENTAL SETUP

3.1. Database

To evaluate the algorithm, a database of police siren signals was collected. Siren signals of eight non-driving German police cars were recorded at 16 kHz sampling frequency. The siren systems comprised three different producers. If possible, both the rural and urban siren was recorded. Each signal was recorded inside the police car as well as outside (about 2 m in front of the car). The environment was quite silent so that the recordings can be regarded as noise free. The siren signals were trimmed to single occurrences of one low and one high tone and embedded in a silent signal of 5 s length. Thus, 378 clean siren signals were available. To consider the Doppler effect, the signals were artificially Doppler shifted using standard audio sofware (Adobe Audition 1.0). A signal source with velocities between $-50\frac{m}{s}$ to $50\frac{m}{s}$ was assumed, that moved straight towards/from a stationary observer. Different samples of traffic noise were downloaded from the internet and trimmed to 5 s-samples.

For sirens and traffic, three disjunct sets were generated: A training set (250 clean sirens, 240 traffic samples), a development set (64 clean sirens, 233 traffic samples) and a test set (64 clean sirens, 237 traffic samples).

Multi condition training was performed to develop robust models, i.e. the training noise was added to the clean recordings with defined signal-to-noise ratios (SNR). The SNR was calculated on the time interval of the present siren. The SNR conditions were clean, 20 dB, 10 dB, 05 dB, 0 dB, and -5 dB. For testing, additionally -10 dB, -15 dB and -20 dB were investigated.

3.2. Classifiers

Two kinds of classifiers were investigated: PBMs and HMMs. The PBM used mel-spectrograms as input. The signal was windowed by 25 ms Hamming windows with 15 ms overlap. The frequencies ranged from 300 Hz to 4500 Hz resulting in 40 mel bands. The size of the parts were seven mel bands and 30 time frames. In the mel-spectrograms of the clean siren data, a 4-parts PBM was labeled as initial model. This initial model was used for the EM algorithm on the whole training set. For the clean condition training only clean data was used. The multi condition training also utilized noisy samples. The resulting hand-labeled PBM will be denoted by PBM(H).

A traffic model was not developed since a siren signal also includes traffic noise and could easily be misclassified. Instead, a threshold to distinguish the classes was considered. Therefore, the siren model was applied to both classes in the development set. The distributions of the scores were modeled by Gaussians. The threshold was defined as the intersection between both Gaussians.

The machine learned models were initialized randomly. To avoid random outliers in classification results, the initialization was done with ten different random seeds. The results will be shown by means and standard deviations over the seeds. The generative one comprising only positive training data (Eq. (7a)) is denoted by PBM(+), the discriminative one (Eq. (7b)) by PBM(-).

For the HMMs, the htk-framework [23] was utilized. The siren model consisted of six emitting states for the siren. The traffic model comprised only one state. The grammar was either traffic(optional)-siren-traffic(optional) or traffic only. The number of mixtures was estimated on the development set. The smallest number with highest accuracy was used. As features, the described mel-spectrogram (HMM(Mel)) and MFCCs with 20 coefficients based on the mel-spectrogram (HMM(MFCC)) were tested.

4. RESULTS

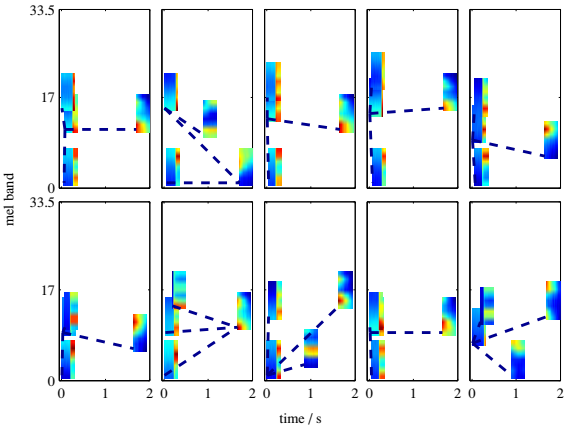

Fig. 1. Models of PBM(+) of the multi condition training. The means μ_i of the part appearances a_i are plotted at the mean relative positions μ_{ij}. The connections between the parts are indicated by dashed lines.

In Fig. 1 and Fig. 2, the 20 machine learned PBMs (of all seeds) of the multi condition training are plotted. The generative models PBM(+) tend to model the on- and offsets of the the fundamental frequencies and harmonics. In contrast, the discriminative model PBM(-) prefers the stationary parts of the tones, that seem to differ more from the traffic noise than the parts considered for the generative models. The parts of PBM(-) are ordered on top of each other. Thus, the full time range of a siren is not as well exploited as it is done for PBM(+).

	clean	multi
PBM(H)	**0.79**	**0.86**
PBM(-)	0.56 ± 0.16	0.85 ± 0.17
PBM(+)	0.59 ± 0.18	0.81 ± 0.17
HMM(MFCC)	0.76	0.80
HMM(Mel)	0.63	0.74

Table 1. Overall accuracies of the five classifiers for the clean and multi condition training.

The accuracies of the classifiers over all SNRs are given in Tab. 1. The accuracies are defined as the mean of the correct recognition rate of the sirens and the traffic noise. The hand-labeled PBM(H) performs best in each training condition. The accuracies for the clean condition training over the SNRs are plotted in Fig. 3. All tested classifiers are capable of classifying sirens with high accuracy on the trained clean SNR condition. Up to 0 dB SNR, the hand-labeled PBM(H) and the HMM(MFCC) achieve high accuracies over 95 % with slightly better performance of HMM(MFCC). Below 0 dB SNR, their performances decrease down to chance level of 50 % accuracy. The performance of the machine learned PBMs decreases to chance level right after the clean condition. PBM(+) shows high standard deviations meaning that the developed models

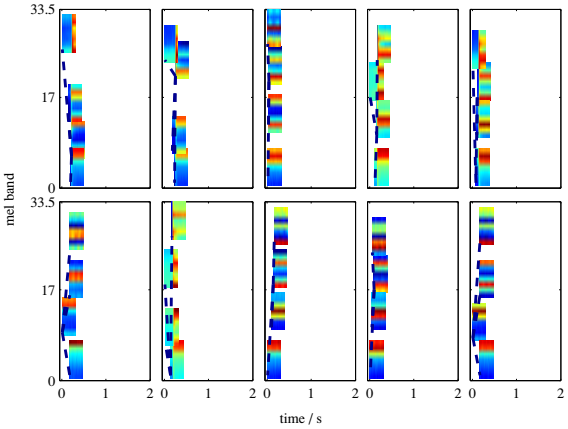

Fig. 2. Models of PBM(-) of the multi condition training. The means μ_i of the part appearances a_i are plotted at the mean relative positions μ_{ij}. The connections between the parts are indicated by dashed lines.

of the seeds perform rather differently. HMM(Mel) achieves better results than the machined learned PBMs. Up to 20 dB SNR it classifies nearly optimal before the performance degrades to chance level.

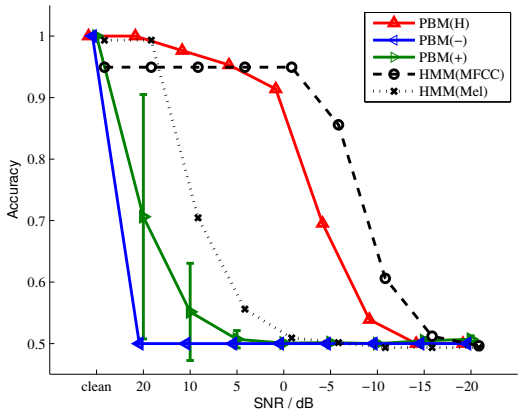

Fig. 3. Accuracies for clean condition training and different SNRs.

For the multi condition training (c.f. Fig. 4), the accuracies up to the lower trained SNR limit of -5 dB are stable on high level. For HMM approaches, the accuracies up to 0 dB are constant since the loss in accuracy to the optimum results from the misclassified traffic samples that are independent of the SNR. The most accurate classifier up to -10 dB SNR is the hand-labeled PBM(H). The second best is the discriminative PBM(-). The machine learned PBMs perform better than the HMMs up to 5 dB SNR. In this SNR range, HMM(Mel) is considerably less accurate than all other classifiers.

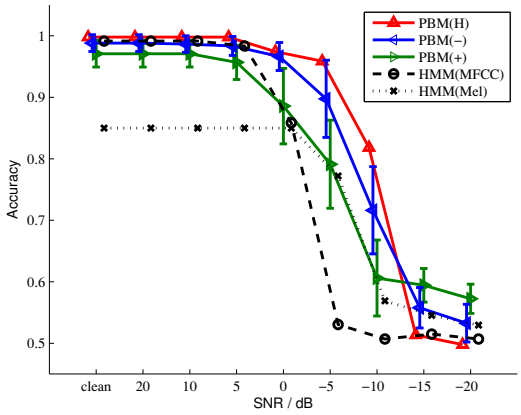

Fig. 4. Accuracies for multi condition training and different SNRs. The training comprised SNRs up to -5 dB.

5. CONCLUSIONS

The performance of PBMs for siren detection in traffic noise was investigated and compared to standard HMM approaches. It was shown that, for clean condition training, clean test samples could be classified with high accuracies by all approaches. The proposed machine learned approaches fail for all other SNRs than the learned one. Both hand-labeled PBM(H) and HMM(MFCC) perform comparably well for this task.

The machine learned PBMs benefit most from multi condition training compared to the clean training. Here, the the hand-labeled PBM(H) and the discriminative, machine learned PBM(-) perform better than the HMM approaches for most SNR conditions.

6. ACKNOWLEDGMENTS

The authors like to thank the police of Lower Saxony, especially the Autobahn police department in Ahlhorn, for their kind support in siren acquisition. Furthermore, the authors also like to thank *Adam Opel AG* for their commitment and enthusiasm during this study. Partial support by DFG (FOR 1732) and EC (Project EAR-IT, 318381) is acknowledged.

7. REFERENCES

[1] N. Vidyasagar, A. Jain, and M. Bianchi, “Emergency vehicle detection system,” University of Illinois, Tech. Rep., December 2010, eCE 445 Senior Design Project Fall 2010, Project 15.

[2] W. E. Brill, “Emergency vehicle detection system,” US Patent 6362749, 3 26, 2002.

[3] B. King and D. A. Yancey, “Gps-based vehicle warning and location system and method,” US Patent 6895332, 5 17, 2005.

[4] R. Lawson, “Emergency vehicle warning system,” US Patent 7061402, 6 13, 2006.

[5] R. T. Halischak, “Multiple emergency vehicle alert system,” US Patent 7236101, 6 26, 2007.

[6] C. Bygrave, "Early warning system for approaching emergency vehicles," US Patent 7515065, 4 7, 2009.

[7] DIN14610, "Sound warning devices for authorized emergency vehicles," January 2009, Deutsches Institut für Normung.

[8] X. Xiao and H. Yao, "Automatic detection of alarm sounds in cockpit voice recordings," in *Proceedings of the IITA International Conference on Control, Automation and Systems Engineering (CASE)*, Zhangjiajie, China, 2009, pp. 599–602.

[9] F. Meucci and L. Pierucci, "A real-time siren detector to improve safety of guide in traffic environment," in *Proceedings of the European Conference on Signal Processing 2008*, Lausanne, Switzerland, 2008.

[10] M. Mielke, A. Schäfer, M. Wahl, and R. Brück, "A mixed signal asic for detection of acoustic emergency signals inroad traffic," *International Journal of Microelectronics and Computer Science*, vol. 1, no. 2, 2010.

[11] R. A. Lutfi and I. Heo, "Automated detection of alarm sounds," *Journal of the Acoustical Society of America*, vol. 132, no. 2, September 2012.

[12] F. Beritelli, S. Casale, A. Russo, and S. Serrano, "An automatic emergency signal recognition system for the hearing impaired," in *Proceedings of the 12th Digital Signal Processing Workshop*, Wyoming, USA, September 2006.

[13] D. P. W. Ellis, "Detecting alarm sounds," in *Proceedings of the Recognition of real-world sounds: Workshop on consistent and reliable acoustic cues*, Aalborg, Denmark, 2001, pp. 59–62.

[14] A. Durey and M. A. Clements, "Melody spotting using hidden Markov models," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Bloomington, IN, USA, October 2001, pp. 109–117.

[15] ——, "Features for melody spotting using hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Orlando, FL, USA, May 2002, pp. II–1765 – II–1768.

[16] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[17] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proceedings of the Workshops on Statistical and Perceptual Audition*, Brisbane, Australia, 2008.

[18] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[19] K. T. Schutte, "Parts-based models and local features for automatic speech recognition," Ph.D. dissertation, MIT Department of Electrical Engineering and Computer Science, Cambridge, Massachusetts, USA, June 2009.

[20] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[21] D. J. Crandall and D. P. Huttenlocher, "Weakly supervised learning of part-based spatial models for visual object recognition," in *Proceedings of the European Conference of Computer Vision*, Graz, Austria, 2006, pp. 16–29.

[22] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centers of circular features," in *ISPRS Intercommission Workshop on Fast Processing of Photogrammetric Data*, 1987, pp. 281–305.

[23] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, 2006.