

# Information Fusion Based Learning for Frugal Traffic State Sensing

Vikas Joshi<sup>1</sup>, Nithya Rajamani<sup>1</sup>, Takayuki K<sup>2</sup>, Naveen Prathapaneni<sup>1</sup>, L. V. Subramaniam<sup>1</sup>

<sup>1</sup>IBM India Research Labs, <sup>2</sup>IBM Tokyo Research Labs,

<sup>1</sup>{vijoshi7, nitrajam, n.prathapaneni, lvsubram}@in.ibm.com, <sup>2</sup>KATS@jp.ibm.com,

## Abstract

Traffic sensing is a key baseline input for sustainable cities to plan and administer demand-supply management through better road networks, public transportation, urban policies etc., Humans sense the environment frugally using a combination of complementary information signals from different sensors. For example, by viewing and/or hearing traffic one could identify the state of traffic on the road. In this paper, we demonstrate a fusion based learning approach to classify the traffic states using low cost audio and image data analysis using real world dataset. Road side collected traffic acoustic signals and traffic image snapshots obtained from fixed camera are used to classify the traffic condition into three broad classes viz., Jam, Medium and Free. The classification is done on {10sec audio, image snapshot in that 10sec} data tuple. We extract traffic relevant features from audio and image data to form a composite feature vector. In particular, we extract the audio features comprising MFCC (Mel-Frequency Cepstral Coefficients) classifier based features, honk events and energy peaks. A simple heuristic based image classifier is used, where vehicular density and number of corner points within the road segment are estimated and are used as features for traffic sensing. Finally the composite vector is tested for its ability to discriminate the traffic classes using Decision tree classifier, SVM classifier, Discriminant classifier and Logistic regression based classifier. Information fusion at multiple levels (audio, image, overall) shows consistently better performance than individual level decision making. Low cost sensor fusion based on complementary weak classifiers and noisy features still generates high quality results with an overall accuracy of 93 - 96%.

## 1 Introduction

Traffic management and engineering is one area which is seeing a lot of AI application in context of urban city transportation systems from data modeling, uncertainty in decision making, planning, scheduling, simulation and impact analysis

perspectives [Biplav and Anand, 2012] [Bielli *et al.*, 1994]. Sensing traffic condition is required in-order to understand the traffic pattern in the city and will also help authorities regulate traffic better. Developed countries employ Intelligent Transportation System (ITS) solutions to detect and manage the traffic condition; magnetic loop detectors [Robertson and David, 1991] being the most popular of sensors. These techniques involve heavy implementation cost and also assume a lane based traffic flow, making it unsuitable in the developing regions which has highly chaotic and non-lane based traffic. RF sensing based approach is also proposed in [Sen *et al.*, 2012] which is applicable under the chaotic traffic conditions with low implementation cost. However, it was shown to work well for congestion detection only, while discriminative capability for three broad classes is not known.

Humans sense traffic by viewing the traffic condition and/or hearing to the traffic noise on the road. Hence audio and image based techniques have been used to sense the traffic [Tyagi *et al.*, 2012][Sen *et al.*, 2010][Coifman, 1998][Li *et al.*, 2008]. These techniques have lower implementation cost and are suitable even for chaotic non-lane based traffic conditions. Vivek *et al.*, [Tyagi *et al.*, 2012] proposed MFCC based classifier which models the frequency spectra of the audio signal to discriminate between the traffic classes. Rijurekha *et al.*, [Sen *et al.*, 2010][Sen *et al.*, 2011] used the honks signals along with two audio sensors to estimate speed of vehicles and thus the traffic state of the road. In another related research [Barbagli *et al.*, 2012], speed of the vehicle is estimated by calculating the time delay for vehicle sound to travel from one sensor to another. Although acoustic based traffic sensing techniques are promising, they have inherent disadvantages. Limitations arise because of inherent "noise" from the environmental factors and hence the ability of current audio processing techniques to sense traffic under such uncertain noisy conditions. For example, interference from other side of the road, from pedestrians, air traffic or public events (prayer calls, festivals), stopping vehicles etc., significantly affects the performance of acoustic classifier. Hence though impressive results were obtained in [Tyagi *et al.*, 2012], we obtained relatively poor results when tested with data collected in more general settings and collected through cheaper smart phones. Approach proposed in [Sen *et al.*, 2010] heavily relies on the honk signals, which may not be available all the time nor even expected in some locations. In acoustic sen-

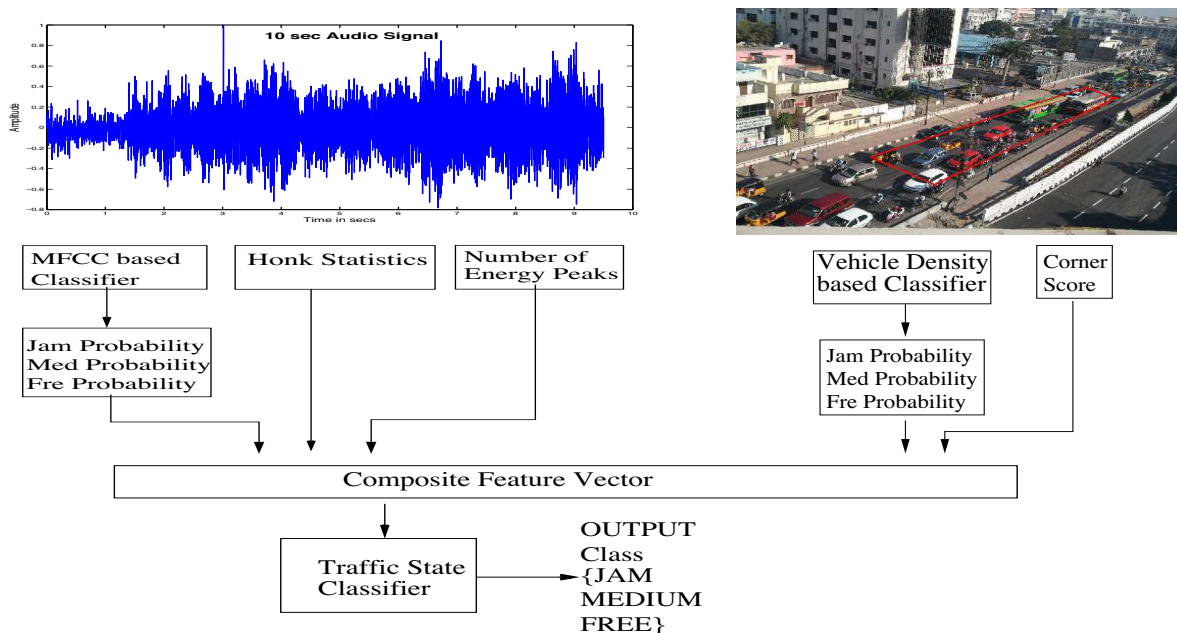


Figure 1: Audio and Image information fusion based traffic classifier. Traffic relevant features are extracted from {10 sec audio, image snapshot in that 10 sec} data tuple and fused to derive the traffic condition

sor networks [Barbagli *et al.*, 2012], the problem of interference of noise signal to vehicular noise is not well addressed and hence may not work well under chaotic traffic conditions.

Image processing techniques have also been proposed for traffic sensing [Coifman, 1998][Li *et al.*, 2008][Jain *et al.*, 2012][Santini, 2000]. In [Li *et al.*, 2008], edge occupation rate is used to detect the congestion. Coifman *et al.*, [Coifman, 1998] used corner features to estimate speed and volume of the vehicles on the road. Image based techniques too have several limitations. Performance of image based algorithms is dependent on the lighting conditions and occlusions. Typically image based features representing vehicle are dependent on height and position of camera w.r.t to vehicle. Also most image algorithms require a still camera and the performance of algorithm is sensitive to small perturbation in the camera position.

Traffic condition in general can be estimated by assessing average speed of vehicles, vehicular density on the road and also by indicative events like amount of honks. Audio based techniques mainly capture and represent the speed of vehicles and other indicative events like honks. While image based techniques could easily provide an estimate of the vehicular density on the road. Hence, although audio and image based traffic sensing methods have limitations, both together provide complementary benefits. Moreover limitations too are complementary, since for example image based sensing is not affected by interference noise from other side of the road and audio based sensing does not depend on the lighting conditions. This motivates us for fusion based approach combining the information from audio and image based traffic sensing techniques to exploit their complementary benefits through AI techniques.

In this paper, we study the efficacy of information fusion from audio and image based features for traffic sensing. Fig. 1 shows information fusion based approach for traffic sensing using audio and image sensors. 10s audio signal and corresponding image snapshot in the 10s segment form the data tuple from which traffic relevant features are extracted based on different audio and image based techniques. Duration of traffic audio sample to estimate traffic condition was fixed to 10s for these fusion experiments. Duration was decided by assessing the trade-off between classification ability, latency involved in transferring and processing the data and bandwidth cost constraints. Our previous experiments together with latency and bandwidth cost estimation with our present prototype formed the basis for our decision to use 10s of audio signal for traffic state classification. Information fusion is done by stacking traffic relevant features obtained to form a composite feature vector. Given a 10s audio segment, individual class probabilities obtained from MFCC based classifier, honk event statistics and number of peaks in the energy contour are added to the feature set. From the image snapshot, individual class probabilities from vehicular density based classifier and corner score form a part of composite feature set. All individual classifiers and features are explained in the next few sections of the paper. Finally traffic state fusion classifier is constructed using the composite feature vector to discriminate between traffic classes. Classification results show that fusion at all levels (audio, image and overall) work consistently better than individual classifiers. Also, overall processing time for 10s audio and image snapshot, and provide classification result is 10s, which is sufficient for real time traffic sensing application. Thus information fusion using frugal, complementary weak classi-

fiers provide impressive overall classification results between 93 – 96%.

The rest of paper is organized as follows. In section 2, we explain audio based feature extraction methods followed by image based feature extraction methods in section 3. Data collection procedure is explained in section 4. Fusion based classifier and experimental results are explained in section 5. Finally conclusions are made in section 6.

## 2 Audio Features

Audio signal collected from road side is used to identify traffic state. Road side cumulative acoustic signal is a mixture of vehicular noise mainly consisting of tire noise, engine noise, air turbulence, and honks [Amman and Das, 2001]. Jam condition is mainly dominated by engine idling noise and honks, while acoustic signal collected in free condition mainly contains air turbulence and tire noise. Spectral content of the acoustic signal is seen to be distinctly different for three different traffic conditions [Tyagi *et al.*, 2012]. Honks are also indicative of traffic condition and hence honk information is also used. The number of peaks in the energy contour provides some measure of vehicles crossing the sensor and hence can be used for traffic sensing additionally. In the following subsections we describe three different algorithms to extract traffic relevant features from audio signal.

### 2.1 MFCC Classifier based features

Vivek *et al.*, proposed MFCC based features to discriminate the traffic states. MFCC features capture the spectral shape of acoustic signal and hence were used as basic parameterization. MFCC features are popularly used for speech and speaker recognition. Steps involved in obtaining the MFCC features and building the classifier are explained in detail in [Tyagi *et al.*, 2012]. Short Time Fourier Transform (STFT) of acoustic signal is obtained by windowing the signal with a window size of 100ms. Mel-warping and smoothing of the frequency spectrum is done by passing the Fourier coefficients through a mel-scaled filter-bank followed by log compression. Finally discrete cosine transform is applied on log filter-bank coefficients to obtain 13 dimensional coefficients. Delta and delta-delta coefficients are appended to obtain final feature vector of 39 dimensions. Thus a 100ms audio frame is characterized by 39 dimensional MFCC features. The next frame is obtained by shifting the window with a shift size of 50ms. Thus within 10s of audio signal there are total of 200 frames (10s/(50ms)).

Using MFCC features directly in fusion framework is difficult since the number of features will be large (39 \* 200) for the 10s audio segment. Instead the individual class probabilities are obtained using above MFCC features. Class probabilities can be obtained by modeling MFCC features using generative models as done in [Tyagi *et al.*, 2012]. MFCC features representing particular class are modeled using Gaussian Mixture Models (GMMs) [Tyagi *et al.*, 2012]. Three GMMs are built, one for each class. GMMs are trained using entire training audio dataset (details of the dataset are explained in section 4). For a given audio segment, average frame likelihoods are obtained from Jam, Medium and Free

models. Finally, individual class probabilities are obtained from likelihoods according to Eqn. 1.

$$P(X \in e_i) = \frac{L(X \in e_i | \lambda_{e_i})}{\sum_i L(X \in e_i | \lambda_{e_i})} \quad (1)$$

where  $P(X \in e_i)$  is the probability of feature vector  $X$  belonging to event class  $e_i \in \text{Jam, Medium, Free}$ .  $L(X \in e_i | \lambda_{e_i})$  is the likelihood of the event  $e_i$  w.r.t to model  $\lambda_{e_i}$  for event  $e_i$ . Thus from a given 10s, individual class probabilities are obtained from MFCC classifier, which forms 3 dimensions of composite feature vector.

### 2.2 Honk based feature

Number of honks at a particular location could provide useful information about the traffic state. In general, more number of honks would correspond to Jam condition. Although honking would depend on the attributes of the driver and location of driving, a more chaotic condition would naturally provoke a tendency and the need to honk. In this section we describe a honk statistics based feature for traffic state estimation. The term *Honk Statistics* mentioned in the paper correspond to percentage of honk frames within 10s audio signal. Honks have been previously used in [Sen *et al.*, 2011] [Mohan *et al.*, ] as one of the feature vector in their discriminative classifiers.

Steps to obtain honk statistics from given 10s audio signal are explained as below:

1. **Short Time Fourier Transform (STFT):** Audio signal is divided into frames with window size of 100ms and shift size of 50ms as used for MFCC classifier. FFT is then applied on the windowed signal.
2. **Honk detection:** Honk frames are detected from the STFT of the audio signal. Honk frames are typically characterized by number of harmonic peaks in the frequency range of 2kHz to 4kHz, referred to as honk frequency range. Since there are multiple peaks, variance of the squared magnitude values of the frequency spectrum is high. Hence the *variance of the amplitudes of squared magnitude frequency spectrum* within the honk frequency range is compared with a threshold to detect the honk frames. There are a few other approaches in literature to detect the honks [Sen *et al.*, 2010] [Mohan *et al.*, ]. Our experiments (not discussed in this paper for brevity) showed that variance based approach we used is more robust to spurious peaks compared to other approaches.
3. **Calculating the honk statistics:** Percentage of the honks frames (honk statistics) within a 10s audio segment is calculated next. Within 10s audio segment, there are 200 frames (with frame shift size of 50ms). Honk detection is done for each frame. Percentage of honk frames is then calculated from 200 frames.

Thus honk percentage obtained is used in the composite feature vector for traffic sensing.

### 2.3 Energy peaks based feature

Energy contour of the audio signal also contains some discriminative information about the traffic classes. Free traffic condition typically has fewer number of vehicles moving

at higher speed. Hence energy contour for free condition is characterized by sparse (because of fewer vehicles) and sharp peaks (since vehicle speeds are high). On contrary, energy contour of Jam condition has more number of peaks, since there are more number of vehicles passing the audio sensor within a given time. Hence number of peaks within 10s signal is also used in the composite feature vector and is used to discriminate the traffic classes.

Since traffic acoustic signal collected is noisy, energy contour has lot of spurious peaks, which could result in false detection of peaks. We attempt to reduce the false detection of peaks. The steps involved in obtaining number of peaks from 10s audio segment is described below:

1. Short time energy contour of the audio signal is obtained using window size of 100ms and shift size of 50ms as used in MFCC and honk statistics. Thus 10s audio segment has 200 frames and hence 200 energy points.
2. Energy signal is smoothened by performing low-pass filtering, using a simple averaging filter with order 20. This removes high frequency noise and reduces spurious peaks present in the energy contour. The bandwidth and order of the filter is chosen such that peaks due to vehicles are still retained.
3. Energy signal is divided into bins corresponding to 1s of audio signal. A maximum of one peak is allowed in a particular bin. The assumption behind this rule is that there could be at maximum one vehicle contributing to the energy signal during the 1s audio bin. Limitation with above constraint is that, it could also miss the true peaks if many vehicles are passing sensor point at higher speeds. However it is not a common scenario and hence though the constraint could miss a few true peaks, it is seen to significantly reduce the false peaks.
4. Each energy bin is then checked for presence of a valid energy peak. Only peaks above a defined threshold are considered to be valid energy peaks. Threshold is set empirically by examining the training data.
5. Finally the number of energy peaks detected in the 10s energy contour is used as a feature contributing the composite feature vector.

Energy peak based feature alone does not perform acceptably well as seen by the results in the Table 3. However, it is still discriminative enough that inclusion of energy peaks in the final composite vector is seen to have importance in the classification results, especially in improving the results for Jam class (note that recall of NPeaks for Jam is very good for most of the classifiers in the fusion experiments).

### 3 Image Features

Image snapshot of traffic condition can provide information about vehicular density on the road. Since qualitative vehicular speed information is captured by MFCC based audio classifier, our focus primarily is to get more information about vehicular density alone from image data. Images are taken from fixed overhead camera as explained in section 4. Region of interest (also termed as Focus Area) within the image

is defined manually as shown in Fig. 2 (a). Focus area needs to be defined manually once for each location to cover enough ground in the image to represent the approximate traffic density in the road. The algorithms described in the paper, extract the traffic relevant information from the defined focus area.

#### 3.1 Vehicular density based features

##### Estimation of Vehicular Density

A simple heuristic based approach for traffic sensing is used by estimating the vehicular density on the road. Vehicular density in a given image within focus area is obtained according to Eqn. 2.

$$VehDen = \frac{\text{number of pixels covered by vehicles}}{\text{total number of pixels}} \quad (2)$$

Pixels within the chosen focus area are classified into 1) Road segment or 2) Non-road segment. Non-road segment is assumed to be occupied by the vehicles. Road segment is the gray area that is visible on the roads that is not covered by vehicles or other occlusions. Classification into road and non-road segment is done by a simple thresholding based approach. A section of the road segment within a sample image (with 33000 pixels) is taken to obtain mean ( $\bar{\mu}_{road}$ ) and variance ( $\bar{\sigma}_{road}^2$ ) of the pixel distribution belonging to the road segment.  $\bar{\mu}_{road}$  and  $\bar{\sigma}_{road}^2$  are 3 dimensional vectors representing Red (r), Blue (b) and Green (g) pixel statistics. Classification of the pixel  $p(i, j)$  is done according to Eqn. 3

$$\bar{p}(i, j) \rightarrow \begin{cases} \text{Vehicle} & |p^c(i, j) - \mu_{road}^c| \geq T(\sigma_{road}^c) \forall c \in \{r, g, b\} \\ \text{Road} & \text{otherwise} \end{cases} \quad (3)$$

where  $c$  represents  $r, g, b$  dimension. Condition in Eqn. 3 need to be meet for all 3 dimensions in-order to associate a pixel to vehicle.  $T$  is the threshold and is chosen to be 6 (6-sigma limit) in our experiments so that road segment has reasonable margin. Our goal is to have features which are indicative of the qualitative traffic density even if they might not be the most accurate when used alone. During night conditions, vehicles are not visible and hence this approach does not work well. Vehicle lights are more prominent indicators of presence of vehicle during night conditions and are also easily separable from the background. Hence the algorithm is modified for night conditions, where vehicular density is defined according to Eqn. 4.

$$VehDenNgt = \frac{\text{number of pixels covered by vehicle lights}}{\text{total number of pixels}} \quad (4)$$

Binary classification of pixels from night image is done to check if a pixel represents vehicle light. Threshold based classification is done as in day algorithm, where mean ( $\bar{\mu}_{light}$ ) and variance ( $\bar{\sigma}_{light}^2$ ) of vehicle light pixels are obtained and threshold is applied as in Eqn. 5.

$$\bar{p}(i, j) \rightarrow \begin{cases} \text{light} & |p^c(i, j) - \mu_{light}^c| \leq T(\sigma_{light}^c) \forall c \in \{r, g, b\} \\ \text{Non light} & \text{otherwise} \end{cases} \quad (5)$$

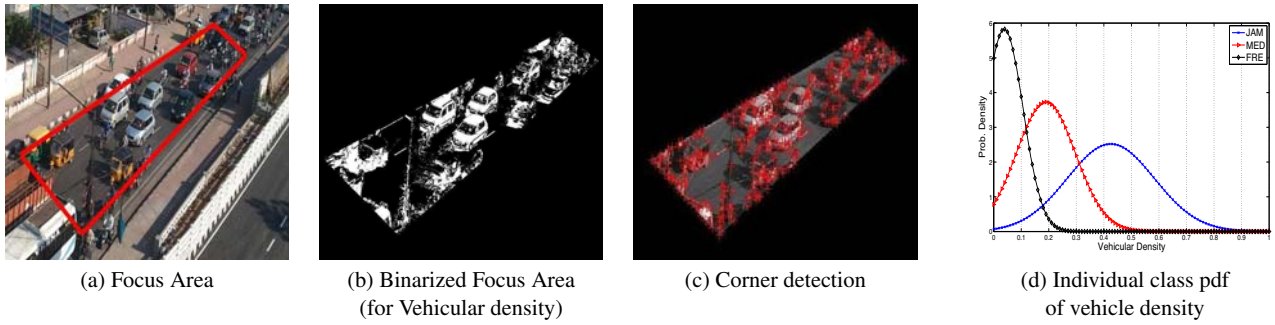


Figure 2: Figure showing (a) Sample focus area (b) Corresponding binarized focus area (c) Corners detected (d) Individual class pdfs of vehicular density traffic image classifier

Thus the vehicular density obtained is used as discriminative feature to classify traffic into Jam, Medium and Free classes. This approach to estimate vehicular density is very similar as in [Jain *et al.*, 2012]. Presented algorithm does have limitations since it does not detect constant obstruction present in the image and could falsely classify those into vehicular area. This problem can be addressed by observing the statistics of pixels over certain time duration. Constant obstruction within vehicular pixel class would typically have fixed pixel values along time, unlike other pixels. We reserve this enhanced approach for future studies, when we also leverage image for speed estimation. Range and statistics of vehicular density for each class is different for day and night conditions because of difference in definition of vehicular density as shown by Eqns. 2 and 4. Hence vehicular density cannot be directly included into composite vector. In order to have a normalized representation across day and night, each class probabilities are obtained and are used as representative features (instead of the actual vehicular density score). Approach to obtain the individual class probabilities, given the vehicular density and condition of the day (day or night), is explained in the following subsection.

### Modeling Vehicular density

Vehicular density obtained for each class is modeled using Gaussian distributions, since histogram for each class was mostly Gaussian in nature. Separate models are built for day and night conditions. Parameters of distribution are learnt using respective training data. Each class pdf's are plotted as shown in Fig. 2(d). Given a sample traffic image, vehicular density is estimated. Then each class likelihoods are obtained using the individual class models. Individual class probabilities are then obtained according to Eqn. 1. Finally class probabilities are used as representative features (irrespective of day or night condition), which will be a part of composite feature vector.

### 3.2 Corner based features

Corner features are extensively used by Cofiman *et al.*, [Coifman, 1998] to track vehicles, estimate the speed and volume of vehicles. We use corner based features to discriminate between the traffic classes. Since vehicular density is expected to be more in Jam condition, number of corners within focus area is also expected to be more in comparison with Medium

Condition (Day/Night)		City		Total	
		DEL	HYD		
Day	Ground Truth	Jam	270	275	545
		Med	0	387	387
	Fre	337	228	565	
	Total	607	890	1497	
Night	Ground Truth	Jam	0	115	115
		Med	0	115	115
	Fre	35	155	190	
	Total	35	385	420	
Total	Ground Truth	Jam	270	390	660
		Med	0	502	502
	Fre	372	383	755	
	Total	642	1275	1917	

Table 1: Table describing the number of data samples from Delhi and Hyderabad cities under Day and Night conditions.

and Free condition. Corners within focus area are detected by standard Harris corner detection algorithm. Normalization is done according to Eqn. 6 to obtain a measure of corner density and is termed as corner score (*CrnScore*).

$$CrnScore = \frac{\text{number of corner pixels}}{\text{total number of pixels}} \quad (6)$$

It is observed that statistics of corner score did not vary much with day and night condition (in night corners are automatically detected around vehicle lights). Hence corner score was directly used as feature in the composite feature vector.

## 4 Data Collection Process

We own an end-to-end research prototype system which does:

1. Crowdsourced audio/image/text samples from smart phones and/or receives data from fixed sensors
2. Process audio samples in backend to determine traffic state
3. Upload real-time status to dashboard and raise alerts to events management system in case of severe congestion

Since we do not own any fixed sensor deployments, we designed a data collection process to evaluate the combined sensor fusion with audio and image. We collected tandem audio and image data from two cities with different traffic patterns and vehicular composition - DEL (Delhi) and HYD (Hyderabad) from 6 different locations altogether. Audio and images were collected using Samsung Galaxy model phones. Images were collected from road-side buildings at least 3 to 5

		Train	Test	Total
Ground Truth	Jam	325	335	660
	Med	271	231	502
	Fre	366	389	755
Total		962	955	1917

Table 2: Table describing number of train and test samples

floors above the ground level and smart phone was fixed on tripod for the entire session of capture.

The image was captured with settings of  $2MP$  and snapshot capture at every  $10s$ . This meant each image snapshot was  $550 - 600KB$  in size. Assuming a fixed sensor sampling  $10s$  audio + image snapshot, it would be  $< 1MB$  in size for each sample being processed for a decision. About 10 hours of audio/image data were collected from HYD and DEL (6 and 4 respectively) which were synced up and cleaned up for any corrupted sessions/data. Table 1 presents the data collection finally used for experiments (Night-on or after  $6pm$ ). For the experiments, we split the data into training and test set for audio, image and fusion classifiers. Some data capture sessions in HYD were marked as test along with the ground truth for each record in that test set. For rest of the data, we split the entire data into first contiguous half in test and next half in train (both audio and image). Table presents information on training/test data. Audio and image, though captured by different phones, were synced and correlated. So each image time-stamp is correlated and synced to the audio 10 secs time frame that subsumes it. Thus we built a dataset of 1917 records of analysis with the ground truth for all the experiments where both audio/image features can be analyzed/fused in tandem for each  $10s$  (5.5 hours in total).

## 5 Fusion and experimental results

Information fusion approaches have been researched extensively in multimedia literature [Atrey *et al.*, 2010][Dasarathy, 1997][Lewis and Powers, 2004]. We mainly focus on fusion at feature level, by stacking the features obtained from audio and image data. Final composite feature vector is of 9 dimensions (dim) constituting a) Class probabilities from MFCC based classifier -  $3dim$  b) Honk Percentage -  $1dim$  c) Number of energy peaks -  $1dim$  d) Class probabilities from Vehicular density based classifier -  $3dim$  e) Corner Score -  $1dim$ . Traffic state classifier is then built with composite feature vector. Here, input variables are continuous in nature and output variables are categorical in nature (Jam, Medium or Free). Hence classification algorithms are chosen to suit the given kind of input and output variables. Four classification approaches were tried namely a) SVM classifier b) Decision trees based classifier (c5) c) Logistic regression based classifier and d) Discriminant Analysis based classifier. All the above classifiers are trained using 962 samples and tested on 955 samples as explained in detail in section 4. All the fusion models are built and tested using SPSS Modeler using default settings except c5 was configured to use boosting with 10 trials and logistic regression used a base category of Free class. Radial Basis Function kernel is used in SVM classifier.

## 5.1 Discussion of experimental results

Table 3 shows the precision (P), recall (R) and overall accuracy (A) for the different classifiers. Standard definitions for precision, recall and accuracy are used as given by Eqn. 7.

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad A = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

MFCC classifier alone performs consistently well in comparison to other individual classifiers. Performance of honk and number of energy peaks based features (represented by Npeaks) is low, while, feature level fusion of MFCC, honk and Npeaks based classifier, represented by Audio\_All, performs consistently better than individual audio feature based classifiers. VehDen and CrnScore represents classification results for vehicular density based classifier and corner score based classifier respectively. Again fusion of image features, represented by Image\_All, performs considerably better than individual classifiers. Decision level fusion (represented by Decision\_All) also performs better than individual classifiers, where the decisions from MFCC classifier and Vehicular Density based classifier are used instead of 3-class probability features for those components. Finally the fusion classifier using composite feature vector, represented by Fusion\_All, consistently out-performs all individual classifiers.

Some of dimensions could have high correlation and hence a smaller sub-space within  $9dim$  can be chosen which have maximum discrimination between the classes. Predictor importance scores for different dimensions showed that atleast 7 out of 9 feature dimensions were important in most of classifiers. Thus even though accuracies for some of features individually are low, their complementary benefits help in fusion.

## 6 Conclusion

We propose an information fusion based learning approach for traffic sensing based on frugal audio and image sensor feeds motivated by their complementary strengths. Information fusion is done by stacking the individual features obtained from the different processing techniques and using the combined set for traffic state estimation. Experiment results demonstrate that composite feature vector based fusion classifier in 9 dimensional space, had more discriminative capabilities than any of individual classifiers. Also, overall processing time for  $10s$  audio and image snapshot, including all feature extraction and final classification is  $10s$ , which is sufficient for traffic sensing application. Thus, fusion of simple, fast and weak classifiers with complementary strengths can provide good quality results in realtime. Since this allows for traffic state to be detected in 3 classes accurately, when collected over a period of time, it could help higher level AI applications to understand and plan better towards reducing traffic congestion in cities that deal with chaotic traffic patterns.

## References

[Amman and Das, 2001] S.A. Amman and M. Das. An efficient technique for modeling and synthesis of automotive engine sounds. *IEEE Transactions on Industrial Electronics*, 48(1):225–234, 2001.

Algorithm	Precision			Recall			Overall Accuracy
	Jam	Med	Fre	Jam	Med	Fre	
MFCC	90.5	77.9	84.8	87.76	73.2	90	85.13
HONK	83.51	48.50	76.65	92.23	56.27	62.46	71.41
Npeaks	61.85	0	70.82	93.4	0	81.74	66.07
Audio_All	95.38	84.68	85.98	92.53	76.62	93.05	88.90
Vehicle Area	85.05	58.08	83.8	93.5	49.78	90.05	78.95
CrnScore	68.11	38.9	86.8	70.2	27.27	99.7	71.83
Image_All	90.26	63.29	95.39	71.94	81.38	95.88	83.98
Decision_All	95.28	89.76	94.26	96.41	83.54	97.17	93.61
Fusion_All	<b>96.70</b>	<b>94.57</b>	<b>96.75</b>	<b>96.41</b>	<b>90.47</b>	<b>99.48</b>	<b>96.23</b>

a) Classification Accuracies (%) using c5 Decision trees

Algorithm	Precision			Recall			Overall Accuracy
	Jam	Med	Fre	Jam	Med	Fre	
MFCC	91.8	77.53	83.57	83.83	76.52	90.46	84.5
HONK	87.82	42.42	61.17	90.44	6.06	90.74	69.93
Npeaks	61.85	0	70.82	93.4	0	81.74	66.07
Audio_All	96.70	85.71	88.97	96.70	78.26	93.55	90.68
VehDen	87.42	61.07	78.28	91.3	39.39	91.77	78.95
CrnScore	65.59	32.39	87.07	72.83	19.91	98.71	70.58
Image_All	88.03	77.40	87.82	92.23	59.30	96.4	85.97
Decision_All	95.65	93.71	91.16	98.50	77.48	98.20	93.00
Fusion_All	<b>97.58</b>	<b>91.26</b>	<b>91.80</b>	<b>96.70</b>	<b>81.73</b>	<b>98.19</b>	<b>93.40</b>

b) Classification accuracies (%) using Logistic regression

Algorithm	Precision			Recall			Overall Accuracy
	Jam	Med	Fre	Jam	Med	Fre	
MFCC	91.83	78.28	83.05	84.13	75.21	90.28	84.5
HONK	87.82	46.42	60.99	90.44	5.62	91.25	70.04
Npeaks	61.85	0	70.82	93.4	0	81.74	66.07
Audio_All	97.29	85.71	88.99	97.00	78.26	93.81	90.89
VehDen	85.36	61.83	78.24	94.02	35.06	91.51	78.74
CrnScore	65.69	32.84	87.10	73.73	19.4	98.97	70.89
Image_All	86.33	80.50	87.44	94.32	55.41	96.65	85.86
Decision_All	95.05	92.10	90.73	97.61	75.75	98.20	92.27
Fusion_All	<b>98.17</b>	<b>91.07</b>	<b>92.68</b>	<b>96.70</b>	<b>84.34</b>	<b>97.93</b>	<b>93.92</b>

c) Classification accuracies (%) using SVM Classifier

Algorithm	Precision			Recall			Overall Accuracy
	Jam	Med	Fre	Jam	Med	Fre	
MFCC	92.33	69.74	87.87	86.26	81.81	83.8	84.19
HONK	93.95	40.93	67.28	83.58	30.30	84.06	70.89
Npeaks	72.61	35.88	70.82	72.83	26.40	81.74	65.24
Audio_All	96.71	78.71	92.45	96.71	84.84	88.17	90.36
VehDen	91.24	53.53	77.82	80.89	45.88	92.03	76.96
CrnScore	64.24	31.39	85.49	53.13	30.30	100	66.7
Image_All	91.08	64.08	80.65	85.37	50.21	95.37	80.94
Decision_All	92.19	79.78	87.09	91.64	64.93	97.17	87.43
Fusion_All	<b>98.48</b>	<b>89.25</b>	<b>91.48</b>	<b>97.01</b>	<b>82.68</b>	<b>96.65</b>	<b>93.40</b>

d) Classification accuracies (%) using Discriminant classifier

Table 3: Classification accuracies (%) for traffic classes with different machine learning approaches

- [Atrey *et al.*, 2010] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. *Multi-modal fusion for multimedia analysis : a survey*. 2010.
- [Barbagli *et al.*, 2012] Barbara Barbagli, Gianfranco Manes, Rodolfo Facchini, Santa Marta, and Antonio Manes. Acoustic Sensor Network for Vehicle Traffic Monitoring. pages 1–6, 2012.
- [Bielli *et al.*, 1994] M. Bielli, G. Ambersino, and M. Boreo. *Artificial Intelligence Applications to Traffic Engineering*. 1994.
- [Biplav and Anand, 2012] Srivastava. Biplav and Ranganathan. Anand. Traffic management and AI. In *AAAI 2012 Tutorial*, 2012.
- [Coifman, 1998] Benjamin Coifman. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research*, pages 271–288, 1998.
- [Dasarathy, 1997] Belur V Dasarathy. Sensor Fusion Potential Exploitation Innovative Architectures and Illustrative Applications. *Proceedings of IEEE*, 85(1), 1997.
- [Jain *et al.*, 2012] Vipin Jain, Ashlesh Sharma, and Lakshminarayanan Subramanian. Road traffic congestion in the developing world. *Proceedings of the 2nd ACM Symposium on Computing for Development - ACM DEV '12*, page 1, 2012.
- [Lewis and Powers, 2004] T Lewis and D Powers. Sensor fusion weighting measures in audio-visual speech recognition. In *Proceedings of the Conference on Australasian Computer Science*, pages 305–314, 2004.
- [Li *et al.*, 2008] Li Li, Chen Long, Huang Xiaofei, and A Jian Huang. Traffic congestion estimation approach

from video using time-spatial imagery. In *International Conference on Intelligent Networks and Intelligent Systems*, 2008.

- [Mohan *et al.*, ] Prashanth Mohan, Venkata N, and Ramachandran Ramjee. Nericell : Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones. *ACM Sensys*.
- [Robertson and David, 1991] D Robertson and Bretherton David. Optimizing networks of traffic signals in real time the scoot method. *IEEE Transactions on Vehicle Technology*, 40, 1991.
- [Santini, 2000] S. Santini. Analysis of traffic flow in urban areas using web cameras. In *Fifth IEEE Workshop on Applications of Computer Vision*, pages 140–145, 2000.
- [Sen *et al.*, 2010] Rijurekha Sen, Bhaskaran Raman, and Prashima Sharma. Horn-Ok-Please. In *ACM Mobisys*, San Fransico, USA, 2010.
- [Sen *et al.*, 2011] Rijurekha Sen, Pankaj Siriah, and Bhaskaran Raman. RoadSoundSense : Acoustic Sensing based Road Congestion Monitoring in Developing Regions. *SECON*, pages 125–133, 2011.
- [Sen *et al.*, 2012] Rijurekha Sen, Abhinav Maurya, Bhaskaran Raman, and Rupesh Mehta. Kyun queue : A sensor network system to monitor road traffic queues. *SenSys*, 2012.
- [Tyagi *et al.*, 2012] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram. Vehicular Traffic Density State Estimation Based on Cumulative Road Acoustics. *IEEE Transactions on Intelligent Transportation Systems*, (Sept), 2012.