Proceedings of the IEEE ITSC 2006
2006 IEEE Intelligent Transportation Systems Conference
Toronto, Canada, September 17-20, 2006

A Multimodal Framework for Vehicle and Traffic Flow Analysis

Jeffrey Ploetner and Mohan M. Trivedi

Computer Vision and Robotics Research Laboratory (CVRR)
University of California, San Diego
La Jolla, CA 92093, USA
ploetner@ucsd.edu, mtrivedi@ucsd.edu

***Abstract*—This paper presents an overview of a novel multimodal system being developed at UC San Diego for vehicle detection and traffic flow analysis. A Distributed Multimodal Array (DiMMA) framework is presented for sensory data acquisition, processing, analysis, fusion, and "active" control mechanisms needed to recognize objects, events, and activities which have multi-modal signatures. Current sensing modalities being researched include video, audio, seismic, magnetic, and passive infrared. Feature extraction and data fusion techniques are being investigated to improve robustness and study the advantages and disadvantages of each sensing modality. Preliminary results of this rapidly deployable system are discussed, along with possible future expansions, including laser range scanners, geophones, pneumatic road tubes, and traditional inductive loops.**

I. Introduction

THERE is a great need for accurate and reliable detection, classification, and tracking of vehicles. At one level, there is the need to collect data and model higher level traffic patterns, which transportation agencies use to ease congestion by better planning and optimization of roadways and construction. At a lower level, the ability to detect and classify individual vehicle types allows for traffic composition analysis, and has applications such as ensuring vehicle class compliance on automated toll roads. If classification and identification are strong enough, systems can be developed for downstream vehicle reidentification, which allows for anonymous tracking for travel time estimates, routing and origin/destination information, as well as dynamic path demands. There is also a strong interest in these technologies from a security standpoint, as detection and classification of approaching vehicles is useful for perimeter security and force protection.

A related field that may be leveraged is structural health monitoring, where sensors are embedded in infrastructure to monitor structural health. Primarily used to monitor long term structural integrity of critical infrastructure after earthquakes or other events, embedded sensors and related technology can sometimes also be used for vehicle detection and traffic flow analysis.

Manuscript received July 3, 2006.

Research reported in this paper is supported by a number of sponsors including US DoD Technical Support Working Group, NSF Information Technology Research Grant for Structural Health Monitoring, and the NSF Graduate Research Fellowship Program.

II. Related Studies

A. Sensor Technology

Many different traffic sensing modalities exist, but the most widely used type of vehicle detectors are inductive loop sensors, which are embedded in the roadway. It is well known that installation of such loops is intrusive and requires shutting down and cutting into the roadway. Furthermore, they do not always work well with motorcycles and bicycles, and they deteriorate with the pavement over time. Despite these drawbacks, inductive loops are a mature technology that is well known, well tested, and widely deployed.

Many alternatives to inductive loops exist, and each type has advantages and disadvantages. For short-term observations from a fixed location, pneumatic road tubes can be temporarily laid across a road to collect basic but important statistics such as vehicle count, speed, number of axles, and vehicle density. More recently, video cameras and computer vision have allowed for rich and non-intrusive traffic analysis over a much larger spatial area. A single video camera can cover multiple lanes for several hundreds of meters, whereas point based detectors require a large deployment array to get wide area coverage. Since many traffic agencies already have video infrastructure deployed for manual field monitoring from a control room, lots of work is being done to leverage this and build video based traffic systems. Though automated traffic analysis from video is starting to gain some traction, it is still an open research area to create robust and reliable real-time systems for quickly changing or extreme environmental conditions.

B. Previous Research

There has been a lot of historical work done on vehicle and traffic flow analysis. More recently, there has been a push to extend systems to use not only multiple sensors, but multiple sensing modalities. Multi-sensor fusion using complementary sensing modalities greatly increases the robustness of any sensing system. It is known that even the

1-4244-0094-5/06/$20.00 ©2006 IEEE

best video detection algorithms are unable to remain robust at all times because of the vast variety of different weather and lighting conditions. Complementing video cameras with sensing modalities that are invariant to these weather conditions is an area that the CVRR is actively pushing towards. Not a lot of work has been done comparing more than a couple types of sensing modalities at once. The following are a few vehicle detection references.

Sun et al. [1] used video cameras and inductive loop signatures to build a vehicle reidentification system. The system, while primitive and relying on strong and unlikely assumptions, illustrated the concepts involved in multimodal vehicle reidentification, namely synchronization, correspondence, feature selection, and data fusion.

Perconti et al. [2] is conducting sensor fusion research for vehicle detection using multiple nodes with video and microphone arrays. Source localization from the microphone array can be used to cue the video control to look in the direction of approaching vehicles.

Cheung et al. [3] are using rapidly deployable wireless magnetic sensor networks on roadways for vehicle detection and classification. Preliminary results show very high detection rates and the promise of wireless sensor networks as a more convenient and easily deployable infrastructure.

In addition to multimodal vehicle detection research, here is some recent work in structural health monitoring systems.

Elgamal et al. [4] have developed a high level health monitoring framework for bridges and civil infrastructure, laying the foundation for what requirements health monitoring systems should have. Many of the system components they describe, such as sensor networks and databases, are also required for a traffic analysis system.

Karbhari et al. [5] have developed a web-based structural health monitoring system for a bridge that includes accelerometers, strain gauges, and temperature sensors. Data from the sensors is transmitted wirelessly to a remote server and processed in real time. With appropriate software modifications, the very same accelerometer and strain gauge sensors that monitor the health of the bridge could also be used to count, record, and classify traffic. Lynch [6] gives a recent and comprehensive overview of wireless sensor technologies for structural health monitoring.

III. Integrated Multimodal Systems

The CVRR has been pursuing research directed towards the development of "intelligent" or "smart" environments [7]. A defining characteristic of intelligent spaces is having situational awareness of objects, events and activities taking place in these spaces. In the case of intelligent highways, relevant events and activities would involve those associated with traffic flow, vehicles, incidents, or structural conditions [8]. Sensors are an essential part of an intelligent environment, as they provide inputs which can be analyzed to recognize objects, events, and activities.

A framework which has formed a general basis for the CVRR's intelligent environments is that of Distributed Interactive Video Arrays (DIVA) [9], which provide wide area coverage, from multiple perspectives, to support active exploration of the environment using event triggered mechanisms. The research presented in this paper uses the DIVA framework as its basis and is directed towards extending the sensory modalities beyond video. The two primary sensing modalities, other than video, which are examined in this paper, are those of audio and seismic. In such a distributed multimodal array (DiMMA) framework, new sensory data acquisition, processing, analysis, fusion, and "active" control mechanisms need to be derived to recognize objects, events, and activities which have multi-modal signatures.

The goal of this research is to take the best aspects of multimodal vehicle detection and leverage structural health monitoring technologies to build a more robust and complete system and framework. The CVRR has developed several systems in the past that use DIVAs for vehicle detection, tracking, and event detection [9]. Many issues have been addressed, such as omnidirectional localization and tracking, camera handoff, feature extraction, vehicle reidentification, and 3D tracking. The CVRR has also developed a test bed for vehicle classification using video cameras and strain gauges [10], the lab's first undertaking into multimodal vehicle detection. These systems have motivated the current work of adding many more sensors and sensing modalities.

A. Architectures and Subsystems

In general, traffic monitoring systems need to be modular, scalable, accurate, and reliable in all weather and environmental conditions. There are a number of design issues and tradeoffs that need to be studied in designing multimodal heterogeneous sensor networks and systems. One critical aspect is timing and synchronization—all data needs to be reliably time stamped and synchronized very accurately to establish correspondences between different sensing modalities. Unless data is synchronized and time stamped when it is acquired, it can quickly become a maze to try to manually synchronize and piece together later. This is obviously an absolute necessity for any automated or real-time system. Another issue is the question of centralized versus distributed analysis—how much processing should be done at the sensor source, and how much should be done from a central location. Doing more processing/analysis closer to the sensors allows cuts down on the amount of information that is needed to be transmitted back to the central server and database. However, the central server generally has access to more external data and more processing resources. Depending on the types of sensors used and type of processing and data required, some tasks will be better to do locally, and some will be better or required to do from a central location. Figure 1 shows a

block diagram of a representative framework for the multimodal system that is described next.

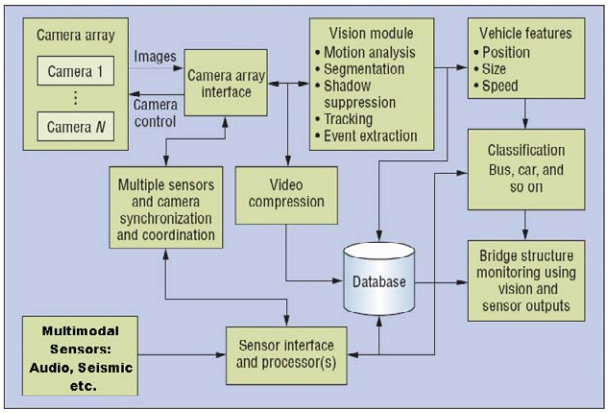

Fig. 1. Block diagram of system framework.

B. *Vehicle Detection and Traffic Flow Analysis*

The CVRR has developed and is continually expanding an integrated multimodal system and framework for vehicle detection, tracking, and event detection. The tradeoffs between various sensing modalities are being tested. The system currently uses video cameras, microphones, seismic accelerometers, passive infrared sensors, and magnetometers. The infrared detectors and magnetometers are deployed in an 8-node wireless sensor network, whereas the rest of the sensors are currently wired to a computer through various data acquisition devices. The whole system is portable and can be deployed and running in approximately one hour. Figure 2 shows an overhead view of a deployment at UCSD on a bridge crossing Interstate 5, and Figure 3 shows an overhead view of a deployment on a busy intercampus loop.

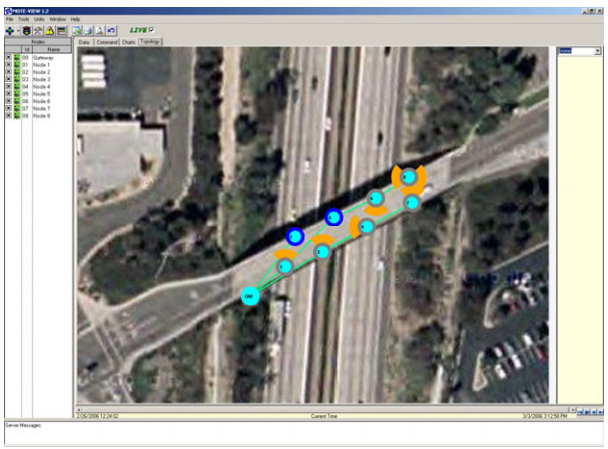

Fig. 2. Test deployment on a bridge.

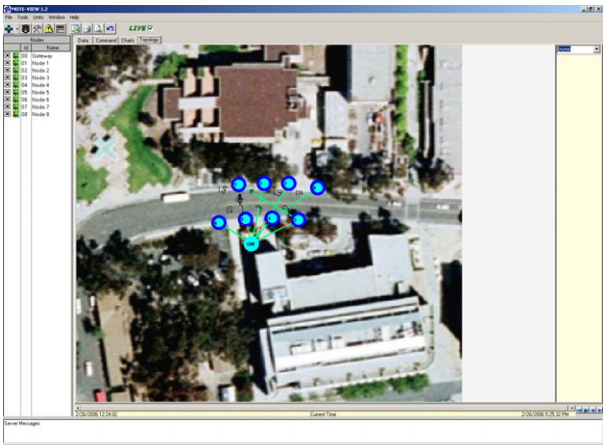

Fig. 3. Test deployment on a roadway.

1) Video Analysis: Video based traffic flow analysis has been an active research area for over a decade. The goal is to derive vehicle and traffic related parameters using video images. Development of accurate, reliable, and robust algorithms to successfully handle wide variations in the scene composition, illumination conditions, shadows, and occlusions is not a simple task. Most of the previous efforts use single rectilinear CCD cameras, and use simple linear transforms to translate from image to world coordinates. While single sensor views are useful, dependence on a single view severely limits the quantity and quality of data available from the viewable environment. In order to overcome such limitations, the DIVA framework was introduced [8]. The DIVA supports the following capabilities:

a) *Distributed video networks*: to allow complete coverage the sensors must be placed in a wide area.

b) *Active camera systems*: exploitation of redundant sensing is mandatory. For this reason, this framework must have one, or more, central "monitors" able to select the camera with the best view of a given area in response to an event. Focus-of-attention in multiple camera systems is a relevant, and relatively new, research area.

c) *Multiple object tracking and handoff*: to create a model of the environment and interact with it, the objects in the scene must be detected, segmented and tracked not only in each view but also among different views. This problem is usually referenced as the "camera handoff" problem or the "reidentification" problem.

d) *3-D localization*: once that the object has been detected, tracked in different views and re-identified, the system should be able to assert where it is in the 3-D world coordinates. 3-D camera coordination in a multicamera system in an effective way is still a challenging research topic.

e) *Multisensor integration*: how to exploit information from rectilinear CCD cameras, omnidirectional cameras and infrared cameras in an integrated and effective way.

The research reported in this paper fits in the above DIVA framework, and extends its functionality to utilize two other sensing modalities, acoustic and seismic. Unfortunately, the passive infrared motion detectors and magnetometers in the wireless sensor network do not yet provide accurate or reliable enough data for useful analysis, so these modalities have been discarded from analysis.

2) Seismic Analysis: Seismic vibrations are greatly dependent on what type of ground the accelerometers are placed on. One will get very different results placing accelerometers on soil, concrete, or structures. Figure 4 shows an eight second sequence of accelerometer data collected from the bridge using four seismic accelerometers. One can see that the whole bridge generally moves as one unit, making source localization from the accelerometer array difficult, but still partially possible. One the other hand, Figure 5 shows detailed accelerometer data from a roadway on campus. The figure shows a car driving by the seismic array, and one can clearly see the by the progression of colors that the front and rear axles of the vehicle move from one sensor to the next. Much future analysis can be done from the information provided by the accelerometers.

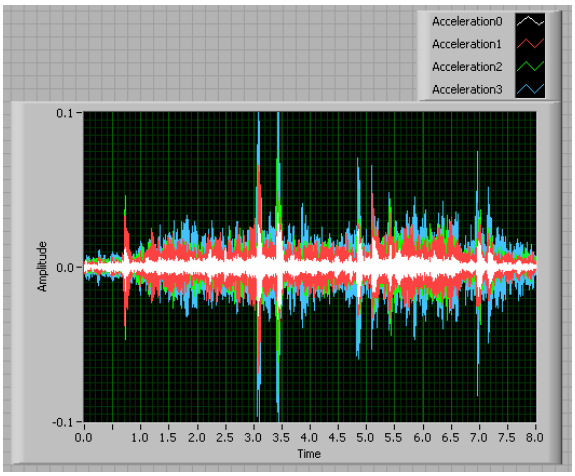

Fig. 4. Seismic accelerometer data from bridge.

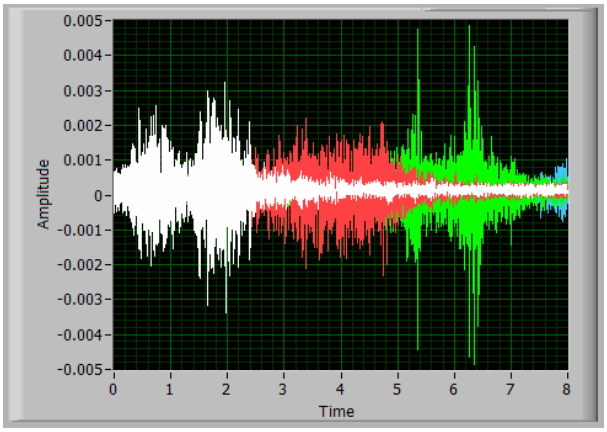

Fig. 5. Seismic accelerometer data from roadway.

3) Audio Analysis: Figure 6 shows a typical three minute sequence of audio, along with a spectrogram. By taking the normalized power of the signal and averaging it, as shown in Figure 7, one can clearly see a series of large and small blips. The large spikes correspond to large vehicles crossing the bridge, while the small spikes correspond to cars. By doing adaptive thresholding and training relative to the background noise, one can do a simple classification into groups of small and large vehicles. More rigorous audio processing will be done in future experiments, with arrays of microphones for source localization, and directional noise cancellation using beam-forming to suppress off-road noise.

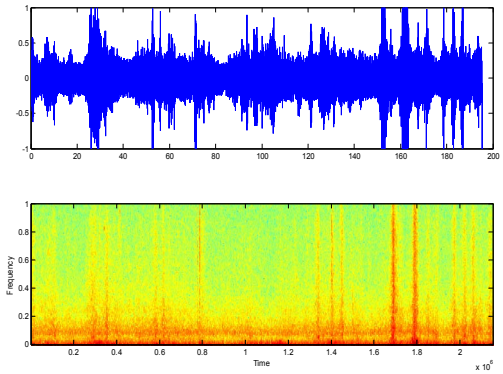

Fig. 6. Audio waveform and spectrogram.

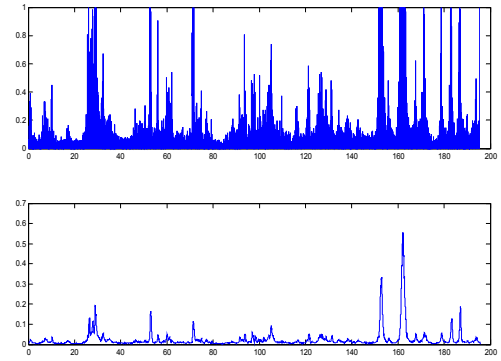
Fig. 7. Audio waveform power and averaged power.

C. Event-based Triggering

Because of the cost, power, and computational requirements of video sensors, it is expensive to have a large scale dense deployment of cameras. One way to increase the effectiveness of available video sensors is to build an event-based triggering system, where events can be detected by some method and commands can be issued to the cameras to zoom into a particular area. One could deploy a sensor network of relatively cheap sensor nodes in a more dense configuration, and use events detected by the sensor network to cue the more valuable and sparse cameras to zoom into or pay particular attention to a particular area. The same concept can be applied to other types of sensors as well.

IV. Preliminary Experimental Studies

Ongoing experimentation with the system is being performed, and a preliminary vehicle classification system is now presented based on audio data only. Just two classes have been initially chosen—large and small vehicles, and the admittedly strong assumption is made that large vehicles are generally louder than small vehicles. Table 1 shows a confusion matrix for this vehicle classification system using only audio, run on a sequence of 35 minutes of data. The system does an acceptable job of classification considering that only audio data is used. After looking at where the system has false positives or misses, there is strong confidence that these can be greatly reduced if the audio data is fused with video data. Table 2 shows a predicted confusion matrix for a combined audio and video classification system. When using video in addition to audio, the number of false detections is greatly reduced because both audio and visual detection are needed to make the detection. Since many of the false detections for audio were due to other vehicle noises, the video will generally not also have a simultaneous false detection. Likewise, incorrect large vehicle predictions are due to small vehicles being loud and sounding like large vehicles. Using video data can correct most of these incorrect classifications, because it can see the actual size of vehicles. Were there to be any large vehicles that sounded like small vehicles, though there was none in the test data, video analysis would be able to pick that up as well.

		Predicted Vehicle Size		
		Large	Small	Missed Small
Actual Vehicle Size	Large	29	0	0
	Small	14	169	28
	False Detect	0	45	

Table 1. Confusion Matrix for Audio Only

		Predicted Vehicle Size		
		Large	Small	Missed Small
Actual Vehicle Size	Large	29	0	0
	Small	3	203	5
	False Detect	0	2	

Table 2. Predicted Confusion Matrix for Audio plus Video

Figure 8 shows how seismic data was manually synchronized by some jumping impulses, along with some seismic events. The larger Figure 9 on the next page shows video snapshots corresponding to the last half of the seismic data in Figure 8. The multimodal sensor data provides a very rich picture of what is happening and opens the whole field up for new analysis and fusion techniques, which will be studied in the future.

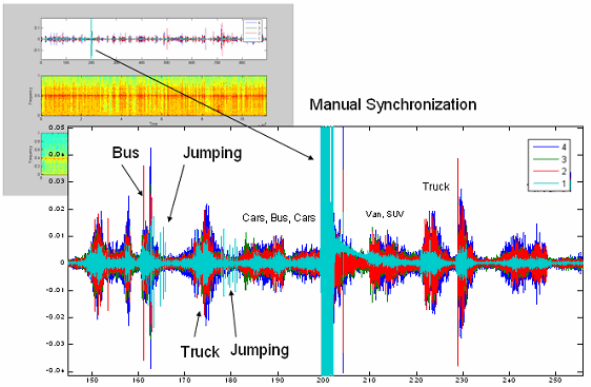

Fig. 8. Events, Manual synchronization of seismic data.

V. Conclusions and Future Work

An overview has been presented of a novel multimodal system under development for vehicle detection, classification, and traffic flow analysis.

In the short term, much more data will be collected under the current system and subjected to a much more thorough analysis. Models will be built and classifiers will be trained to detect and classify vehicles for audio, video, and seismic sensors. Data fusion techniques will be employed to enable more robust tracking and event detection.

In the medium term, more sensing modalities will be added to the system. Laser range scanners will be added, and road tubes could be added. A larger array of microphones will also be added. Geophones versus seismic accelerometers will be studied. A classical inductive loop would be nice to compare things to, because it is currently the gold standard for vehicle detection. The goal is to get as much data as possible so that one can make better design decisions about what is really needed, what is not, and in what configuration.

In the long term, once a critical mass of sensors is discovered that is required to adequately monitor an area to a desired level of detail and robustness, a more long-term deployment-oriented system will be developed, with the eventual goal of real-time processing and dissemination.

Acknowledgment

We acknowledge the assistance and cooperation of our colleagues from the Computer Vision and Robotics Research Laboratory. We also thank Dr. Rajesh Hegde for his valuable advice in audio signal processing.

REFERENCES

[1] Sun, C.C.; Arr, G.S.; Ramachandran, R.P.; Ritchie, S.G. "Vehicle Reidentification Using Multidetector Fusion" IEEE Transactions on Intelligent Transportation Systems, Volume 5, Issue 3, Sept. 2004 Page(s):155 – 164.

[2] Perconti, P. Loew, M. Hilger, J. "Overview of sensor fusion research at RDECOM - NVESD & recent results on vehicle detection using multiple sensor nodes," Proceedings of the Sixth International Conference of Information Fusion, 2003. Volume: 1, 492- 498.

[3] S.-Y. Cheung, S. Coleri Ergen and P. Varaiya. "Traffic surveillance with wireless magnetic sensors," Proc. 12th ITS World Congress, San Francisco, Nov. 2005.

[4] A. Elgamal, J. P. Conte, S. Masri, M. Fraser, T. Fountain, A. Gupta, M. M. Trivedi, M. El Zarki, "Health Monitoring Framework for Bridges and Civil Infrastructure," 4th International Workshop on Structural Health Monitoring, Stanford University, Stanford, CA, September 15 - 17, 2003.

[5] Hong Guan, Vistasp M. Karbhari, Charles S. Sikorsky, "Web-Based Structural Health Monitoring of an FRP Composite Bridge," Computer–Aided Civil and Infrastructure Engineering, Volume 21, Number 1, January 2006, pp. 39-56(18).

[6] Lynch, Loh. "A Summary Review of Wireless Sensors and Sensor Networks for Structural Health Monitoring," The Shock and Vibration Digest, Vol. 38, No. 2, March 2006, 91–128.

[7] M. M. Trivedi, K. S. Huang, I. Mikic, "Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces", IEEE Trans. on Systems, Man and Cybernetics, Part A, Volume: 35, Issue: 1, Jan 2005. Pages: 145-163.

[8] M. M. Trivedi, A. Prati, G. Kogut, "Distributed Interactive Video Arrays for Event Based Analysis of Incidents," 5th International IEEE Conference on Intelligent Transportation Systems, Singapore, September 3-6, 2002, pp. 950–956.

[9] M. M. Trivedi, T. L. Gandhi, K. S. Huang, "Distributed Interactive Video Arrays for Event Capture and Enhanced Situational Awareness," IEEE Intelligent Systems, Special Issue on AI in Homeland Security, Volume 20, Issue 5, Sept.-Oct. 2005 Page(s):58 - 66.

[10] R. Chang, T. Gandhi, M. Trivedi, "Computer Vision for Multi-Sensory Structural Health Monitoring System," 7th IEEE Conf. on Intelligent Transportation Systems, October 2004.

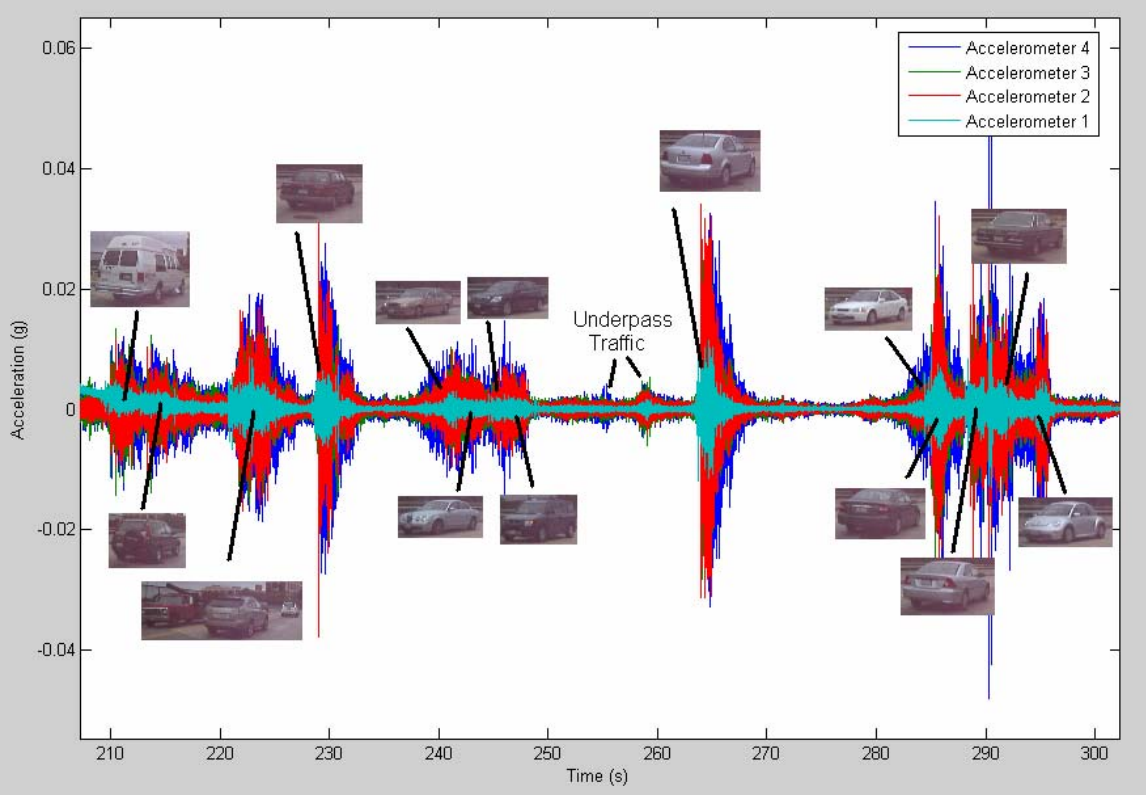

Fig. 9. Seismic and Video Correspondence