JOINT ACOUSTIC-VIDEO FINGERPRINTING OF VEHICLES, PART II

V. Cevher**, F. Guo, A. C. Sankaranarayanan, and R. Chellappa*

Center for Automation Research,
University of Maryland,
College Park, MD 20742
{volkan, fguo, aswch, rama}@cfar.umd.edu

ABSTRACT

In this second paper, we first show how to estimate the wheelbase length of a vehicle using line metrology in video. We then address the vehicle fingerprinting problem using vehicle silhouettes and color invariants. We combine the acoustic metrology and classification results discussed in Part I with the video results to improve estimation performance and robustness. The acoustic video fusion is achieved in a Bayesian framework by assuming conditional independence of the observations of each modality. For the metrology density functions, Laplacian approximations are used for computational efficiency. Experimental results are given using field data.

***Index Terms*—** Object recognition, pattern recognition, acoustic applications, acoustic signal processing, intelligent sensors

1. INTRODUCTION

In object recognition problems, video cameras are preferred because they bring in an array of rich information encoding the object identity. Unfortunately, the video information is not easily amenable for automatic inference due to the nature of the video observations, which may present the object in varying and unknown illumination and pose, background clutter, and occlusion. To achieve recognition, the methods in the literature concentrate on the object appearance [1], shape [2], or a combination of the two [3], by using intrinsic properties of the object that are invariant with respect to the nature of the video observations.

The recognition of vehicles using video first requires robust vehicle segmentation. Statistical and systematical models alleviate this problem by (i) learning the background [4, 5], (ii) handling shadows [6], and (iii) discriminating non-vehicles such as humans [7]. Once the object is segmented, deformable models [2, 8], silhouettes [9, 10], and realistic 3D object models parameterized by appearance [3, 11] are used to extract the defining vehicle characteristics.

In this paper, we focus on vehicle video features that are complementary to the acoustic vehicle profile vector in Part I [12] to improve and, at the same time, to validate the acoustics-only results. We first show how to estimate vehicle wheelbase length and vehicle aspect ratio using video sequences, using minimal camera calibration. We then discuss extraction of vehicle shape and appearance features to emphasize the intra-class variations that cannot be achieved based on vehicle size or vehicle engine type. We focus on vehicle silhouettes since they are easy to determine in our field data.

2. VIDEO MENSURATION

The coordinate transformation from a scene to an image in video is a projective transformation, which distorts geometrical properties of the scene, such as parallelism, ratio of lengths, etc. Observed images, as a result of projective transformations, preserve the following properties from a real scene (invariants): concurrency, collinearity, order of contact, and the ratio of ratio-of-lengths (a.k.a., the cross-ratio) [13]. Hence, difficulties arise in a mensuration (or metrology) problem, where we would like to make measurements of an object within a scene using only images taken by a camera. The problem becomes even more challenging when the internal calibration of the camera is also unknown.

In the vehicle fingerprinting problem, as complementary the features to the acoustic vehicle profile vector, we determine the following vehicle dimensions using a video sequence, collected by a stationary camera oriented perpendicular to the motion of the vehicles: vehicle wheelbase L and aspect ratio (AR). By using multiple video frames, we obtain a distribution of each vehicle dimension to reflect our estimation confidence in a Bayesian framework. We use a camera calibration scheme specifically designed for vehicle mensuration problem [14]. As a result, we assume that the vanishing line of the reference plane and the vertical vanishing point is available in our derivations. This calibration scheme is know as the minimal camera calibration [15].

2.1. Line Segment Metrology

In this section, we describe a robust and computationally efficient procedure of estimating the length of a line segment given a known length on the reference frame, i.e., in the scene. Unfortunately, this problem cannot be solved using a single reference length and it can be proved that at least three reference lengths are necessary unless the line segments are perfectly parallel. Hence, we demonstrate how to measure distances from any desired point using three reference lengths. Solution using more reference lengths has an efficient subspace solution and is described in detail in [14].

Figure 1 shows the basic idea of how to obtain the length of a line OC' using three known equal lengths $|A_iB_i| = r$, $i = 1, 2, 3$. Even in this simple 2D case, we still need three reference lengths. By the geometrical construction based on the vanishing line, $A_iB_iB_i'O$'s are all parallelograms. Hence, the points B_i' are all equidistant from the target point O and define a circle around O in the real scene, which then becomes an ellipse in the image plane due to the projective transformation. Hence, the length of the line OC is simply given by the ratio $|OC| : |OC'|$, where $|OC'| = r$.

Prepared through collaborative participation in the Advanced Sensors Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-02-0008.

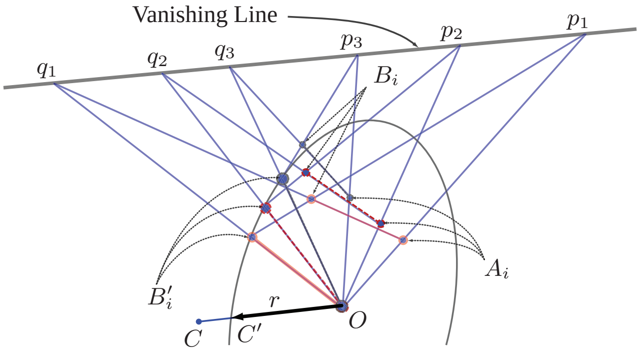


Fig. 1. Lines that start from the same vanishing point are parallel by the projective geometry. Then, $q_iO \| q_iA_i$ and $p_iO \| p_iB'_i$ by construction.

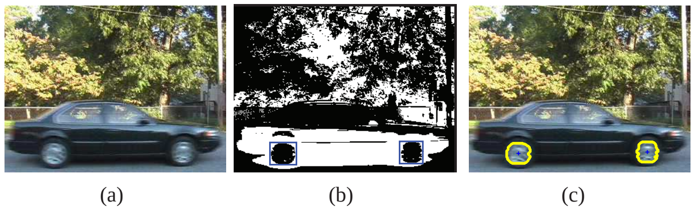


Fig. 2. (a) Original image (Nissan Maxima in Fig. 3(d,e,f) in Part I). (b) Thresholded image. (c) Estimated wheel covers and their centers (+).

2.2. Wheel Detection and Tracking

For wheel detection, we assume that the tire of a vehicle is always black and the wheel covers are silver or gray. Therefore, a simple intensity filter can separate the wheel cover from the tires as shown in Fig. 2. After the extraction of the blobs that represent the wheel cover, we determine their center position and mark as the wheel centers. When this estimation is done jointly with the tracking of the vehicle [14, 16, 17], it results in multiple wheelbase length estimates that defines a pdf $p(L)$ for L.

We now describe the video wheelbase pdf estimation procedure by an example. Figure 3 shows the wheelbase detection and pdf estimation results for the control vehicle Nissan Maxima on a two-way street (Fig. 2). The video resolution is 320×240. In the figure, the lower left corner is taken as the origin. Figure 3(a) shows the wheelbase estimation results, corresponding to the different runs of the same vehicle. We acquire 24 frames from a total of four runs. We use the line segment metrology to project the estimates on the same wheel center and calculate the mean and the variance of the wheelbase distribution. This variance is then used to construct the pdf in Fig. 3(b) using Parzen methods [18].

3. SHAPE AND APPEARANCE

In this paper, we use extracted vehicle silhouettes to determine the vehicle shapes, because the video camera used in our experiments is oriented perpendicular to the motion of the vehicles and is parallel to the ground plane. At other camera configurations, it is possible to build 3D models using planar motion constraints via tracking [11] or use pre-built models for discrimination under varying illumination conditions [3]. To encode the vehicle appearance, we propose to use illumination invariant (for both matte and shiny surfaces) $l_1l_2l_3$ color measure [19, 20] under white illumination, which is satisfied for vehicles during the day. Equations (27)-(29) in [19] define l_1, l_2, l_3 in terms of RGB values (Table 1).

To determine the vehicle silhouettes, we first determine the statistical parameters of the background [4, 5]. Then, the objects are detected using background subtraction. The resulting image is median filtered and averaged over multiple frames while compensating for the target motion. We choose the largest blob in the resulting image and then remove the vehicle shadows [6] to obtain the silhouette (Fig. 4).

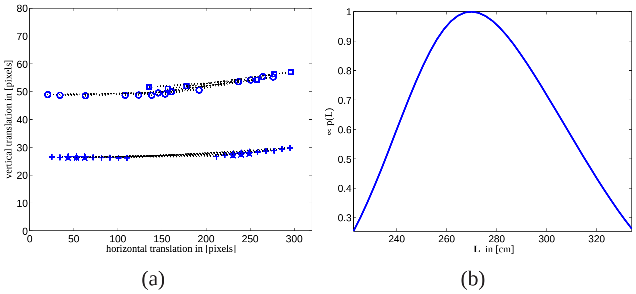


Fig. 3. (a) Wheel detection results of the calibration vehicle Nissan Maxima (+ corresponds to the run in Fig. 3(d,e,f) in Part I). (b) The estimated pdf for the wheelbase length. The mode of the distribution is set to 270cm, corresponding to the manufacturer's specification. The mean and the variance of the distribution is 276.3cm and $(27.3\text{cm})^2$.

Table 1. Vehicle Shape and Appearance

Vehicle	ω_1	ω_2	ω_3	ω_4	ω_5	l_1	l_2	l_3
$\mathcal{V}_1$	8.6177	7.4597	7.7005	5.2211	7.1051	0	0.5000	0.5000
$\mathcal{V}_2$	8.7402	5.6839	6.5009	5.9914	6.5673	0.0055	0.5495	0.4451
$\mathcal{V}_3$	6.9970	2.7016	2.4357	6.9836	3.5238	0.6429	0.2857	0.0714
$\mathcal{V}_4$	7.8884	3.7738	4.0195	6.1920	4.6284	0.3810	0.5952	0.0238
$\mathcal{V}_5$	6.7251	2.5737	2.9351	5.5913	2.9551	0.2368	0.6579	0.1053
$\mathcal{V}_6$	4.2321	5.9754	2.3604	7.8313	1.1111	0.1667	0.6667	0.1667
$\mathcal{V}_7$	7.3992	2.4707	3.2560	5.2688	3.6770	0.2302	0.6594	0.1104
$\mathcal{V}_8$	7.9148	2.5141	3.5488	7.1726	4.3853	0.3228	0.6271	0.0501
$\mathcal{V}_9$	11.6709	9.4643	10.6410	6.5768	9.8081	0.1162	0.6609	0.2228
$\mathcal{V}_{10}$	8.7663	3.5790	4.6002	6.1067	4.9960	0.2162	0.1216	0.6622

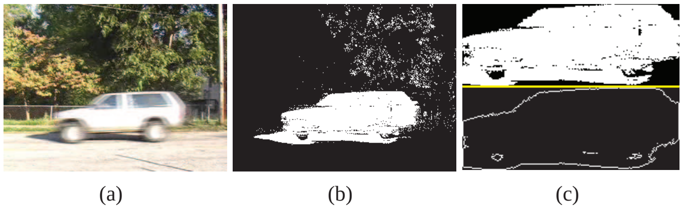


Fig. 4. (a) Nissan Rodeo. (b) Background subtraction. (c) Denoising and shape extraction: (i) select the largest blob in (b), (ii) apply a median filter, (iii) average over multiple frames, and (iv) extract the silhouette.

We use five different classes ω_i to represent vehicle silhouette shapes for bus, sedan, mini van, truck, and SUV, as shown in Fig. 5. To determine the similarity of each of the calculated silhouettes to each of the classes ω_i, we use the Hausdorff distance [21], which is relatively insensitive to the perturbations of the image and is computationally efficient. Table 1 summarizes the Hausdorff distance for the ten test vehicles in Fig. 5. In the calculations, we discard the bottom %30 of the calculated silhouette to decrease the detrimental effect of the shadows. Using the Hausdorff distance, only $\mathcal{V}_3$ is misclassified as in ω_3, whereas it judiciously belongs to ω_2. However, as can be seen in the Table 1 and Fig. 5, the vehicle silhouette is close to both classes. Hence, the misclassification is mainly due to our choice of the generic class shapes.

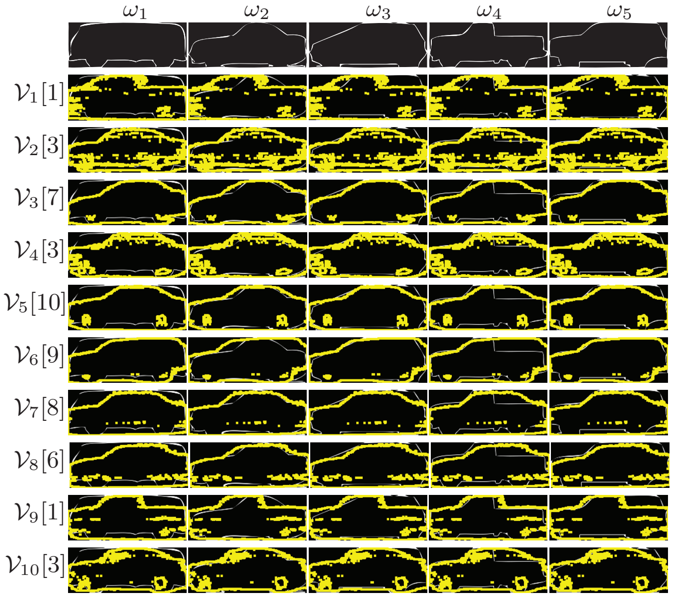


Fig. 5. Vehicle silhouette results. The notation $\mathcal{V}_i[j]$ refers silhouette results for the ith vehicle in Table 2 obtained by using j frames. Note that as j increases, the estimated silhouettes improve.

4. JOINT ACOUSTIC-VIDEO VEHICLE PROFILE VECTOR

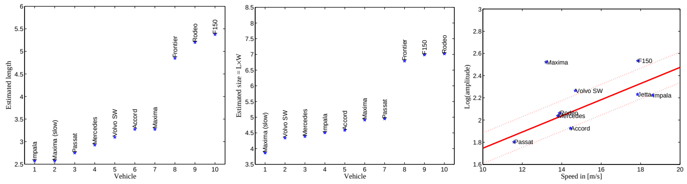
In Part I, we defined an acoustic vehicle profile vector using the physical parameters and the envelope shape (ES) component parameters. We now extend the the acoustic vehicle profile vector to also include (i) the video mensuration results, (ii) the shape information in terms of the Hausdorff distances to each classes ω_i, and (iii) the appearance information l_j to form a fingerprint of the vehicle. As opposed to choosing one class for each vehicle, we deliberately choose to keep the Hausdorff distance to each vehicle class since it allows the propagation of the identity in a probabilistic framework. If other shape, appearance, and mensuration results are available, they should also be included in the profile vector.

5. EXPERIMENTS

5.1. Acoustic Vehicle Fingerprinting

Tables 1 and 2 summarize the results, using the vehicle profiling methods outlined both in Part I and Part II. In Table 2, it is apparent from the variance and bias estimates that acoustic estimation of the vehicle speeds improves, compared to the classical methods such is by Couvreur and Bresler ([7] in Part I), when they are jointly estimated with the classification features. Another possible reason for the improvement is the multiplicative noise model used in our acoustic observations.

The vehicle profile vector also provides a natural basis for classifying vehicles. Figures 6(a) and (b) show that the vehicles can be separated into two classes based on their length and size. Note that the even though estimated vehicle lengths are not exact vehicle lengths, they can separate small vehicles from large vehicles . Figure 6(c) also illustrates that it is possible to identify loud vehicles such as vehicles with mechanical problems or heavily loaded SUV's or pick-up trucks, which are expected to be louder than usual. Hence, given two similar vehicles, it may be possible to identify if one of them is heavily loaded or has mechanical problems even if they move at different speeds.



(a) (b) (c)

Fig. 6. (a) Estimated vehicle lengths are compared. There is a clear separation between small and large vehicles (Rodeo is misclassifed). (b) Estimated vehicle sizes are compared. (c) Logarithm of the vehicle signal amplitudes are plotted with respect to their speed. There is a linear trend in the plot as also indicated by [22]. The solid line represents a least squares fit to the data without Nissan Maxima. The dotted lines are one standard deviation away from the mean. Nissan Maxima is louder than the other cars because the vehicle has mechanical problems.

5.2. Video Vehicle Fingerprinting

The shape, appearance, and mensuration features in video render this modality more capable in distinguishing vehicles. The video can separate the vehicles into finer classes such as sedans, trucks, and SUV's (as opposed to rougher classes such as small or large); and can much accurately calculate vehicle wheelbase length, when compared to acoustics alone. Although the estimated vehicle aspect ratios (AR) are biased, it is easy to see that when de-meaned, AR is an effective discriminative feature (Table 2). Unfortunately, the authors could not automatically determine the wheelbase lengths for some of the test vehicles using the video data, as marked by - in Table 2. A possible reason for this is the faster vehicle movements, which result in fewer frames that are also significantly motion blurred, for parameter estimation due to a constant (and narrow) camera field-of-view.

5.3. Joint Acoustic Video Fingerprinting

In our experiment, the acoustic and video modalities provided complementary information to each other and overcame each other's weaknesses. As can be seen in Table 2, acoustics provided mensuration results when the video failed. Moreover, Fig. 7 illustrates a case where the video helps acoustics to resolve a bi-modal mensuration result. The acoustics-only mode estimate for (L, W) is $(5.20, 1.35)$ m, over the second mode in the likelihood function around $(L, W) = (2.93, 1.5)$m, which is only fractionally lower in the dynamic scale of the log likelihood function. The video mensuration result for the wheelbase length is $L = 2.71$m. When combined with acoustics, the final wheelbase and width estimates become $(L, W)_{\text{joint}} = (2.76, 1.50)$m ($\sigma_{L,\text{joint}} = 7.27$cm), which consequently corrects the acoustic classification result in Fig. 6(a) and (b); and further improves the width estimate. This combination is achieved by multiplying two Laplacian approximations of the modality pdf's. Hence, the resulting fusion result is a Gaussian approximation. Other combined length estimates are $(L, \sigma_L)_{\text{joint}} = (2.96, .20)$m ($\mathcal{V}_3$), $(2.88, .22)$m ($\mathcal{V}_4$), $(2.62, .16)$m ($\mathcal{V}_5$), and $(2.86, .31)$m ($\mathcal{V}_{10}$). In addition, notably, as the video estimates deteriorate when the vehicles increase their speed, the acoustic results tend to improve because the vehicles become louder, thereby improving the acoustic signal-to-noise ratio.

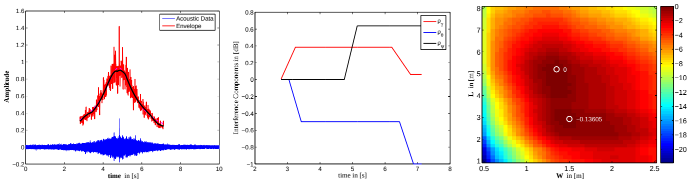
6. CONCLUSIONS

In this paper, we discussed video fingerprinting features that are complimentary to the acoustic profile vector, introduced in Part I. We showed

Table 2. Field Test Results

Ground Truth						Estimation using $\boldsymbol{\lambda}$						Estimation using video-only			Estimation using $\boldsymbol{\lambda}_v^{\dagger}$	
Vehicle	y_m	v_{camera}	L	W	AR	v	C	$L(\mu \pm \sigma)$	$W^{\ddagger}(\mu \pm \sigma)$	χ	$p^{\natural}$	$L(\mu \pm \sigma)$	AR$^{\odot}$	Silhouette	v	C
$\mathcal{V}_1$: Ford F150	6.3	17.54m/s	3.20m	1.70m	2.87	17.86m/s	12.60	5.38 ±.55m	1.30±.09m	3038	8	-	2.41		21.39m/s	21.27
$\mathcal{V}_2$: Chevy Impala	5.8	18.68m/s	2.80m	1.58m	3.49	18.60m/s	9.23	2.58±.31m	1.75±.17m	3300	6	-	3.06		15.05m/s	10.90
$\mathcal{V}_3$: Honda Accord	4.3	16.74m/s	2.71m	1.55m	3.29	14.44m/s	6.86	3.28±.29m	1.40±.17m	3074	6	2.68±.27m	2.95	[§]	14.49m/s	9.67
$\mathcal{V}_4$: Nissan Maxima[*]	4.6	13.32m/s	2.70m	1.53m	3.4	13.20m/s	12.45	3.28±.38m	1.50±.18m	3825	6′	2.70±.27m	2.84		14.27m/s	14.49
$\mathcal{V}_5$: Nissan Maxima[*]	4.1	4.14m/s	2.70m	1.53m	3.4	4.49m/s	6.34	2.58±.19m	1.50±.09m	3150	4	2.70±.27m	2.71		3.46m/s	9.20
$\mathcal{V}_6$: Isuzu Rodeo	8.1	13.44m/s	2.70m	1.51m	2.64	13.89m/s	7.87	5.20±.53m	1.35±.25m	3450	6	2.71±.29m	2.45		11.79m/s	7.95
$\mathcal{V}_7$: Mercedes E	8.1	13.94m/s	2.83m	1.54m	3.34	13.80m/s	7.68	2.93±.96m	1.50±.41m	3075	6	-	3.20		11.78m/s	9.93
$\mathcal{V}_8$: Volvo 850	8.1	14.11m/s	2.66m	1.51m	3.29	14.69m/s	9.60	3.10±.27m	1.40±.27m	2250	10′	-	3.00		11.22m/s	8.63
$\mathcal{V}_9$: Nissan Frontier	4.3	17.56m/s	3.20m	1.56m	2.94	17.84m/s	9.31	4.85±.63m	1.40±.25m	2625	6	-	2.67		17.56m/s	9.74
$\mathcal{V}_{10}$: VW Passat	5.1	11.66m/s	2.70m	1.50m	3.19	11.58m/s	6.06	2.75±.66m	1.80±.26m	1950	6	2.90±.35m	2.72		8.66m/s	6.11
Error STD						0.8246m/s		0.9821m	0.1917m			0.0929m	0.1702		2.2627m/s	
Error STD[¶]						0.2777m/s		0.3292m	0.1613m						1.5126m/s	
Bias						-0.0735m/s		-0.7730m	0.0610m			-0.0360m	0.3840		-1.1458m/s	
Bias[¶]						0.1737m/s		-0.2000m	0.0063m						-1.7013m/s	

[†] Using the method of reference [7] in Part I. [*] Same vehicle. [¶] Calculated by removing the outliers in each method. [‡] A fixed bandwidth of $\mathcal{W} = 600$Hz is used to determine the car widths. Hence, the width estimates of the F150 and Nissan Frontier are biased because they have a significantly different tire profile than the sedan vehicles. When $\mathcal{W} = 800$Hz is used, the width estimates of F150 and Nissan Frontier become 1.60m and 1.80m, respectively. In turn, their wheelbase length estimates also change to 4.20m and 3.98m. [♮] Estimated by finding the frequency F_0 with the maximum power spectral density between frequencies 85-210Hz and then dividing F_0 by the CFR f_0 estimate. [′] Incorrectly estimated. The actual values are 4 (Maxima) and 5 (Volvo). [§] The silhouette is misclassified by a slight margin. The second best silhouette is . [⊙] Aspect ratio estimates are negatively biased because of the imperfect shadow removal.



(a) Isuzu Rodeo (b) Isuzu Rodeo (c) Isuzu Rodeo

Fig. 7. (a) Observed acoustic envelope and the estimated envelope using the ES components are shown. (b) Estimated ES components are displayed. (c) The log-likelihood surface for Isuzu Rodeo for the vehicle dimensions is bi-modal. Acoustics-alone chooses dimensions farther away from the actual vehicle dimensions.

examples where each modality helped the other overcome its shortcomings. The video-only wheelbase estimates are shown to be comparable to acoustic-only wheelbase estimates, when only a few video frames are used for estimation. The joint acoustic video profile vector provides natural vehicle statistics that can be used in sensor networks to propagate vehicle identity in a communication constrained manner.

7. REFERENCES

[1] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. Image Processing*, vol. 11, pp. 530–544, May 2002.

[2] M. Isard and A. Blake, *Active Contours*, Springer, 2000.

[3] O. C. Ozcanli, A. Tamrakar, B. B. Kimia, and J. L. Mundy, "Augmenting Shape with Appearance in Vehicle Category Recognition," in *CVPR 2006*, NYC, June 2006.

[4] A. N. Rajagopalan and R. Chellappa, "Vehicle detection and tracking in video," in *ICIP*, 2000, pp. 351–354.

[5] S. Joo and Q. Zheng, "A temporal variance-based moving target detector," in *IEEE VS-PETS*, 2005.

[6] M. Kilger, "A shadow handler in a video-based real-time traffic monitoringsystem," in *IEEE Workshop on Apps. of Comp. Vision*, 1992, pp. 11–18.

[7] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in *IEEE Workshop on Apps. of Comp. Vision*, Princeton, NJ, October 1998, pp. 8–14.

[8] J. M. Ferryman, A. D. Worrall, G. D. Sullivan, and K. D. Baker, "A generic deformable model for vehicle recognition," in *Proc. of British Machine Vision Conf.*, 1995, vol. 136.

[9] S. Z. Der and R. Chellappa, "Probe-based automatic target recognition in infrared imagery," *IEEE Trans. on Image Processing*, vol. 6, no. 1, pp. 92–102, 1997.

[10] Y. Asokawa, K. Ohashi, M. Kimachi, Y. Wu, and S. Ogata, "Automatic vehicle recognition by silhouette theory," in *Proc. of the 5th World Cong. on ITS*, Korea, October 1998.

[11] A. C. Sankaranarayanan, J. Li, and R. Chellappa, "Finger printing vehicles for tracking across non-overlapping views," in *Army Science Conference*, Orlando, FL, 2006.

[12] V. Cevher, R. Chellappa, and J. H. McClellan, "Joint acoustic-video fingerprinting of vehicles, part I," submitted to ICASSP 2007, available at *http://www.umiacs.umd.edu/users/volkan/javf1.pdf*.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in computer vision*, Cambridge University Press, 2003.

[14] F. Guo and R. Chellappa, "Video Mensuration Using a Stationary Camera," *Lecture Notes in Computer Science, Springer-Verlag*, vol. 3953, pp. 164–176, 2006.

[15] A. Criminisia, A. Zisserman, L. Van Gool, S. Bramble, and D. Compton, "New approach to obtain height measurements from video," in *Proceedings of SPIE*, 1999, vol. 3576, pp. 227–238.

[16] V. Cevher, A. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," accepted with minor revisions to IEEE Transactions on Multimedia, available at *http://www.umiacs.umd.edu/users/volkan/joint%20tracker.pdf*.

[17] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Wiley New York, 2001.

[19] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, no. 3, pp. 453–464, 1999.

[20] A. Diplaros, T. Gevers, and I. Patras, "Combining color and shape information for illumination-viewpoint invariant object recognition," *IEEE Trans. on Image Processing*, vol. 15, no. 1, pp. 1–11, 2006.

[21] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[22] U. Sandberg and A. J. Ejsmont, *Tyre/road noise reference book*, Infomex, SE-59040 Kisa, Sweden, 2002.