Multimodal and Multi-task Audio-Visual Vehicle Detection and Classification

Tao Wang
Department of Computer Science
The CUNY Graduate Center, New York, NY 10016
The City College of New York, New York, NY 10031
twang@cs.ccny.cuny.edu

Zhigang Zhu
Department of Computer Science
The City College of New York, New York, NY 10031
The CUNY Graduate Center, New York, NY 10016
zhu@cs.ccny.cuny.edu

Abstract

Moving vehicle detection and classification using multimodal data is a challenging task in data collection, audio-visual alignment, and feature selection, and effective vehicle classification in uncontrolled environments. In this work, we first present a systematic way to align the multimodal data based the multimodal temporal panorama generation. Then various types of features are extracted to represent diverse and multimodal information. Those include global geometric features (aspect ratios, profiles), local structure features (HOGs), various audio features in both spectral and perceptual representations. A flexible sequential forward selection algorithm with multi-branch searching is used to select a set of important features at different levels of feature combinations. Finally, using the same datasets for two different classification tasks, we show that the roles of audio and visual features are task-specific. Furthermore, in both cases, the combination of some of the features with multimodal and complementary information can improve the accuracy than using the individual features only. Therefore finer and more accurate classification can be achieved by two different levels of integration: feature level and the decision level.

1. Introduction

Moving vehicle detection and classification, in applications such as traffic monitoring [1-3], surveillance [4, 5], and checkpoint inspection [6], can be very challenging where the data are noisy, especially from a single sensor system at a large standoff distance in an uncontrolled environment. Using visual-only sensors may not be sufficient; audio information can provide complementary acoustic signatures, such as loudness and sharpness, for distinguishing different types of vehicles. Vehicle detection using multimodalities reduces false target detection and classification can be more efficient. Typical acoustic sensors are microphones or microphone arrays; however, laser Doppler vibrometers (LDVs) [7] start to show promises in long-range acoustic detection. Regular microphones or microphone arrays need to be placed at fixed locations and near to the targets of interest, whereas all sounds in between are captured if a long range directional microphone is used. The LDV can be used to listen to the target at very long distance (~ 300 meters) when the laser beam is reflected from a good vibration surface near the target, and only sounds close to the vibration surface are captured.

For vehicle classification, we use a multimodal sensor system that we have developed [7], which has a pair of pan-tilt-zoom (PTZ) cameras and an LDV for acquiring video, range and audio information of moving vehicles at a large distance. We analyze various types of visual features and audio features. The visual features include aspect ratio and size (ARS), histograms of oriented gradients (HOGs), shape profiles (SP), representing simple global scale features, statistical features, and global structure features, respectively. The audio features include short time energy (STE), spectral energy, entropy, flux and centroid feature, and Mel-frequency cepstral coefficients (MFCCs), which are grouped into three types: temporal features (STEs), spectral features (SPECs) and perceptual features (PERCs). The selected features are representative for various information, and we expect they can provide complementarities to each other. Next, we provide a flexible sequential forward selection algorithm that allows multi-branch searching in order to avoid local maxima. Then, a number of good features and their branches at different levels of combinations are automatically selected. We notice that the feature extraction and selection are task-dependent. Given different tasks, the same features may play different roles. In this work, we design two different types of classification tasks using the same set of features on the same dataset and provide a thorough study on the feature selection and combinations for vehicle classification using SVMs.

The rest of paper is organized as follows. Section 2 introduces the context of the work and some related work. Section 3 presents an improved algorithm for multimodal data alignment based on a multimodal temporal panorama representation. Section 4 describes multimodal feature extraction from audio signals and reconstructed visual images of vehicles. Section 5 presents a multi-branching sequential forward selection algorithm for robust feature selection. Then, Section 6 shows the experimental results and analysis with two classification tasks. Finally, conclusions are provided at Section 7.

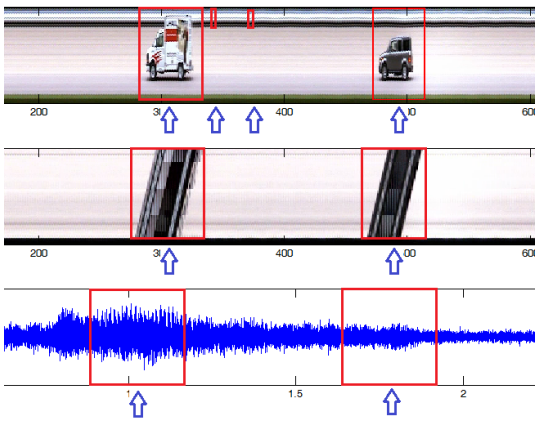

Figure 1: Multimodal data representations in the multimodal temporal panorama (cropped). The top image shows the appearance, the middle image shows the motion, and the bottom image shows the audio wave. The red rectangle boxes show the detection, and the blue arrows show the centers of detected regions determined by the multimodal data alignment.

2. Background and Related Work

For moving vehicle detection, multimodal data (including appearance, motion and acoustic) need to be aligned in order to extract synchronized multimodal features for the targets (vehicles). In our previous work [8], we showed that the video clip and audio data can be represented into a 2D *multimodal temporal panorama* to facilitate both alignment and detection. The key point is to extract from the x- y- t video volume to a y-t appearance panorama and x-t motion panorama, so that both of them can be represented with audio signals in the same temporal domain. We used the same single 1D vertical line cross all frames in a video clip for detecting the moving vehicle passing through a detection location. The detection results are much reliable and consistent than using the image differencing affected by background variations [1, 2]. 1D x-t images along epipolar lines are extracted from all frames and concatenated together to form the motion estimation panorama. In practice, the selected epipolar line should lie on the relative horizontal direction as the motion path of the vehicles in the scene. This concept was first introduced in [9], and it has been used to display features to trace the horizontal motion [10]. Fig. 1 shows a short segment of multimodal temporal panorama representing visual appearance, motion, and audio signals.

We collected audio data using a special "acoustic" device LDV in [7] at a large distance in addition to the visual information. The audio data contain clear vehicle sounds after filtering, and most of them can be separable cleanly. We will deal with the mixture of moving vehicle sounds in the future work. In this paper we investigate the strengths and weaknesses of both visual and audio modalities and examine the effectiveness of multimodal integration; a few related works [11, 12] showed that the combination of both video and audio can provide more reliable results in surveillance. Vehicle detection using both video and audio can be achieved at multiple integrating levels: data level, feature level, and decision level. In [13], the authors discussed a fusion model and a decision model used for the vehicle identification. However, only pre-selected audio and video features are used. They do not study the impact of multimodal feature selection given different tasks. Atrey et al [22] provided a thorough survey on multimodal fusion for multimedia analysis. Of particular interest to us is the work on multimedia synchronization between text and audio–video modalities by using alignment in [23] and [24]. The alignment was performed by maximizing the number of matches of the text and audio-video events - they are considered matched when both are within a temporal range, occur in the same sequential order, and the audio–video event adapts to the modeling of the text event.

3. Multimodal Data Alignment

Because our system allows us to collect multimodal data at different locations and selecting various detection zones, the visual detection and audio detection of vehicles may not be aligned. In other words, depending on the viewing angles of the camera and the directions of the moving vehicles, the system may hear the sound before or after it actually sees them. Also, noise from the background subtraction and ambient sounds may also cause invalid alignment. Here, we present a systematic way to align the multimodal data using the multimodal temporal panorama, similar to the approaches in [23, 24]. Let I_i^D denote intensity map of a vehicle whose center body is detected at the time i in the appearance panorama D. Let I_i^M denote the intensity map displayed at the time i in the motion panorama M. We want to select a correct range $(j\text{-}m, j\text{+}m)$ in an audio clip that corresponds to the detected vehicle, as:

$$\text{argmax}_j \frac{1}{N}\left(\sum I_i^D + \sum I_i^M + \sum_{j-m}^{j+m} A\right) \quad (1)$$

where j is the center of the audio clip and m is the half-duration of the audio. N is the normalization factor, and A is the energy of the audio signals. Unlike human speech signals, the sound of a vehicle is much consistent during a period of time, so usually an audio clip of 3-5 seconds (for m) can describe the signature of a vehicle sufficiently. A constraint Ψ in subject to the Eq. (1) is set as:

$$\Psi = \left(\sum I_i^D > \tau\right)\left(\sum I_i^M > \tau\right)\left(\sum_{j-m}^{j+m} A > \varphi\right) = 1 \quad (2)$$

where τ, φ are thresholds to penalize the visual background noise and the ambient sound. Note that the detection results do not rely on restricted thresholds. Indeed, a clean background subtraction with a filtered audio signal can guarantee a good detection results using very small values for the thresholds τ and φ.

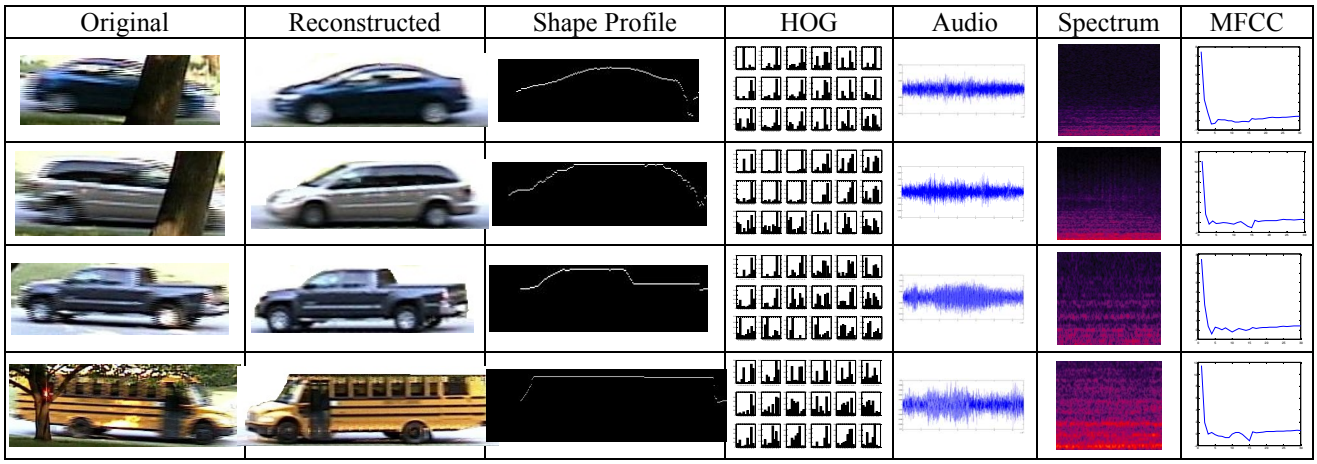

Figure 2: Samples of four types of vehicles: sedan, van, pickup truck and buses: original image shots from video, reconstructed visual images, shape profile, HOG features, audio wave, spectrum and MFCC features are shown in columns from left to right.

Fig. 1 shows the detected objects in red rectangle boxes and alignments in blue arrows. For the first object (vehicle), the detection time of the visual appearance and the audio signals are different; however, they can be aligned using Eq. (1) by finding an audio region that yields the highest total energy with respect to the visual detecting region. In addition, false targets (the 2nd and 3rd objects) in the appearance panorama can be removed, since there is no motion presented to indicate a moving vehicle. In other words, if there is a vehicle detected in all the three panoramas (appearance, motion and audio), the result should be 1 in Eq. (2); otherwise, the result should be 0. Note that the constraint is task dependent; and we assume a moving vehicle could be detected at both video and audio. It is also possible to only be able to hear the sound of a moving vehicle without actually seeing it, then the constraint needs to be redesigned to fit in this situation.

4. Feature Extraction

Various types of features are extracted from both the visual and audio detection results of vehicles. Some sample results are shown in Fig. 2 and the details are described in the following sub-sections.

4.1. Visual Feature Extraction

The visual features are extracted using the reconstructed image results. The detail of the algorithm for moving vehicle image reconstruction from video is described in [14]. The objective of reconstruction is to make vehicles' visual images invariant to perspective views and distances. Also, the results have occlusions and motion blur removed. Therefore, both metric features as well as statistical features can be used more effectively. Four samples representing four different types of vehicles are shown in Fig. 2. Column 1 shows the original images retrieved from the video collected at different locations and interested zones. Column 2 shows their reconstruction results so that all vehicles' visual images are normalized in the same perspective view. For visual feature extraction based on the reconstructed images, the first one can be used is simple aspect ratio and size (ARS) feature, as $f_{ARS} = [w, h, w/h]$, where w is the width and h is the height. It can classify vehicles into various sizes. Next we use the shape profile (SP), which is a strong indicator of the vehicle's type, to represent the top boundary of a vehicle. Only top half of vehicle images are used since only top boundary can be make significant distinguish among different types of vehicles. Then, the top boundary curves of all reconstructed vehicle images are sampled into the same number of bins. Each bin B_i presents the average of height of a part of a vehicle in the image, and to form a feature vector f_{SP} of the same dimension after normalization as:

$$f_{SP} = \frac{1}{max\ B}[B_1, B_2, \ldots, B_N] \quad (3)$$

where N is the number of bins for the SP. Note this normalization loses the size information, but it has been captured by the aspect-ratio and size (ARS) feature.

The other type of visual feature is the histograms of oriented gradients (HOG) feature [15]. It is a statistical feature that preserves some texture and local structure. It counts occurrences of gradient orientation in localized dense grid cells uniformly, thus, forming a feature vector of histogram H as $f_{HOG} = [H_1, H_2, \ldots, H_M]$, where M is the number of bins for the HOGs. Since it uses local contrast normalization, it is invariant to illumination changes, thus, it is good at people detection as well as vehicle detection [16].

4.2. Audio Feature Extraction

In general, audio features can be categorized into three groups: time-series features, spectral features and perceptual features. The time-series features represent audio samples in their raw waveforms. Short time energy (STE) is used to calculate the energy over a time [17]. It is usually good at distinguishing a vehicle sound with a silent background. Since the audio signals of a moving vehicle are much consistent over a short time period, we form the STE feature vector using only its mean and standard deviation as $f_{STE} = [\mu_{STE}, \sigma_{STE}]$.

In the second group, the spectral features (SPEC) represent spectral moments and flatness [18]. Here we select spectral energy, entropy, flux and centroid composed into a spectral feature vector f_{SPEC}=*[Eng, Ent, Flux, Cent]*. The spectral energy *Eng* calculates the energy of the power spectrum defined as:

$$Eng = \sum |F\{x(t)\}|^2 \quad (4)$$

where x(t) is the audio signal and F{} is its Fourier transform. The spectral entropy *Ent* measures the energy changes and defined as

$$Ent = -\sum |F\{x(t)\}| \frac{\log |F\{x(t)\}|}{Eng} \quad (5)$$

The spectral flux *Flux* measures how quickly the power spectrum of a signal is changing and defined as:

$$Flux = \sum (|F\{x(t)\}| - |F\{x(t-1)\}|)^2 \quad (6)$$

The spectral centroid *Cent* indicates the center of the spectrum defined as:

$$Cent = \frac{\sum w|F\{x(t)\}|}{\sum |F\{x(t)\}|} \quad (7)$$

where *w* is the weighted mean vector of the same dimension as the *F*.

In the third group, the perceptual features (PERC) represent the spectral variation and sharpness. Mel-frequency cepstral coefficients (MFCCs) [19] are commonly used to perceptually represent the frequency band responses of the human auditory system. The mel-frequency cepstrum (MFC) equally spaces the frequency band on the mel scale of *F{x(t)}*, and then transformed using the DCT after log of powers at each mel frequency. Then the coefficients of the results forms the perceptual feature $f_{PERC} = [\mu_{MFCC}, \sigma_{MFCC}]$, where μ_{MFCC}, σ_{MFCC} are the mean and the standard deviation vectors of all coefficients, respectively.

5. Feature Selection and Classification

Feature selection may be a task dependent problem. Given two different tasks, the classification results may be different using the same features or feature combinations. To evaluate the individual features or feature combinations, the support vector machines (SVMs) using RBF kernels [21] are used. For the multi-class problem, one-against-one technique is used by fitting all binary sub-classifiers and finding the correct class using a voting mechanism. To evaluate the classifier for a given feature or feature combination *f*, confusion matrix *C* is generated and a classification accuracy value is calculated as $\omega(f)=trace(Diag(C)/sum(C))$ is calculated to indicate what percentage the true labels and expected labels are on the diagonal.

We'd like to evaluate a large number of features and select only a few of representative features or feature combinations for both efficiency and effectiveness. Such problem can be formulated as: given a feature set $F = \{f_i | i=1,..., N\}$, where f_i represents a single modal feature, find a subset of these features and their combinations (bimodal, tri-modal, etc) that maximizes an objective function *J(S)*. The commonly used selection strategy is sequential forward selection (SFS) [20], which starts from the empty set and sequentially adds the feature that maximizes *J(S)*, then the process is repeated testing each remaining feature combinations with those previously preserved until all features have been evaluated. The problem of the SFS algorithm is that only a single best feature is selected at each round so that it has a tendency to become trapped in local maxima.

To alleviate this problem, we provide a *multi-branching sequential forward selection* (MB-SFS) algorithm which selects a number of good features at each round (level) that satisfy some criteria. First, let us define the following symbols:

- N: the number of uni-modal features.
- K: the number of levels of feature combinations, k=1: uni-modal, k=2: bi-modal, and so forth.
- M_k: the number of selected features and/or feature subsets at the level k.
- S_M: the M_th selected feature subset, where M = $M_1+M_2+\ldots+M_k$.
- F': an available feature set, a subset of F.
- F*: features in S_M.
- $\omega(f)$: classification accuracy of a feature or feature combination, *f*
- ε: a small tolerance value.

In the first level, a classifier is trained for each of the N uni-modal features and its classification accuracy is calculated. Then a subset S_{M1} with the top M_1 uni-modal features are selected whose classification accuracy drops from the best one is within a small percentage ε. Then in the second level, each of these M_1 features will be paired with the other un-selected features in the first round to generate multiple bi-modal features to train their classifiers. The same selection rule is used to select the top M_2 bi-modal features. This process continues to level K and therefore the selected feature subset S_M include M features or feature combinations, and $S_M = S_{M1} \cup S_{M2} \ldots \cup S_{MK}$. Last, the feature with the best accuracy among all levels of feature combinations in S_M is selected. This usually will be a multimodal feature, but it could be a unimodal or bimodal feature.

The algorithm is formulated as the following:

1. Start with the empty set $S_0 = \{\varnothing\}$, k=1;
2. *Let* $F' = \{f'_i | i = 1, \ldots, N\} = F$;
3. Select the next subset of *k*-modal features S_{Mk} in the level *k*, by combining a feature f'_i in F' with every feature F^*_j in the subset $S_{M(k-1)}$, $j = 1, \ldots, M_{k-1}$, s.t. $\omega(F^*_j + f'_i) \geq \omega_{max} - \varepsilon$ and $F^*_j + f'_i \notin S_M$ where ω_{max} is the accuracy of the best classifier in *kth* level;
4. Update $S_M = S_{M(k-1)} \cup S_{Mk}$;
5. $F' = F' - \{f'_i\}$, if $F' = \{\varnothing\}$, k=k+1, go to step 2, else go to step 3.

6. Experimental Results

6.1. Experiment Setup

We collected both video and audio data at different locations at a local road. Only vehicles' visual images are reconstructed so that all image results are normalized in the same perspective views and with the occlusions and motion blurs removed. Each image is aligned with its corresponding sampled 5-10 seconds audio clip which presents the audio signature of the vehicle. In total, there are 485 samples with 280 for training and 205 for testing. To demonstrate the feature selection process of individual features and feature combinations for different tasks, we labeled them into two types of categories. In the first type, the vehicles are labeled into four categories: *sedan, vans, pickup trucks*, and *buses*. This type is classified mainly based on the shape information, so we expect the visual features can dominate the classification performance. In the second type, the vehicles are labeled into *light, medium*, and *heavy* vehicles. This type is based on the sounding levels of the moving vehicles. So the light vehicles include economic cars and minivans; the median vehicles include sports cars, large vans; and the heavy vehicles include buses and trucks. We expect the audio features can play important roles.

For both types of categories, the same feature extraction and selection methods are used. In our experiments, the aspect ratio and size (ARS) feature includes aspect ratio (height: length), height and length of vehicles. The HOG divides every vehicle's image into the same 6x3 cells with 9 bins thus has 162 dimensions. The shape profile (SP) feature uses 30 bins across the top profile of the vehicle. The short-time energy (STE) consists of a mean and a standard deviation of a vehicle temporal energy. The spectral feature (SPEC) contains the 4 means of spectral features and their standard deviations. The perceptual feature (PERC) uses the first 30 coefficients of MFCCs and the means and standard deviations are calculated and store into a feature vector of 60 dimensions. For the feature selection process, a cut-off threshold of ε=2% is set below the accuracy of the best classifier using a feature or feature combination among all features in the same level is set.

6.2. Result Analysis

Table 1 shows the feature selection and classification results. For the individual bimodal features, we can see that the SP feature and the HOG feature have better performance than others for the type 1 task. This is mainly because the data are labeled based on the visual appearance. However, they are no longer better than the SPEC and PERC features for the type 2 task since the data are labeled based on audio signatures. As a result, the SP and HOG are selected as the starting nodes for the level-2 bi-modal feature evaluation for the type 1, and so are the SPEC and PERC for the type 2. The detailed accuracy results of the 6 individual features are also shown in confusion matrix in Table 2.In the bi-modal feature

Table 1. Feature selection and classification results. Except that the accuracies of all individual features are presented at the first level, only the selected features passing the thresholds at level 2 and 3 are presented. After level 3, only the best features are shown. Note: ARS-aspect ratio and size, SP-shape profile, STE-short time energy, SPEC-spectral features, PERC-perceptual features.

Type 1: Sedan, Van, Pickup Truck, and Buses				Type 2: Light, Medium, Heavy		
lvl	Features	Acc		lvl	Features	Acc
1	ARS	66.3%		1	ARS	76.6%
	SP	**82.0%**			SP	81.0%
	HOG	**83.9%**			HOG	83.9%
	STE	46.3%			STE	76.6%
	SPEC	43.9%			**SPEC**	**84.4%**
	PERC	75.6%			**PERC**	**86.8%**
2.1	**SP**, HOG	84.9%		2.1	**SPEC**, HOG	84.9%
	SP, PERC	83.4%		2.2	**PERC**, HOG	**93.2%**
2.2	**HOG**, PERC	86.3%		3	**PERC, HOG,** SPEC	92.7%
3.1	**SP, HOG**, ARS	92.7%		4	**PERC, HOG, SPEC**, SP	88.8%
	SP, HOG, PERC	**93.7%**		5	**RPEC, HOG, SPEC, SP,** STE	84.4%
3.2	**SP,PERC**, ARS	93.2%		6	**PERC, HOG, SPEC, SP, STE**, ARS	85.4%
4	**SP, HOG, ARS,** PERC	94.1%				
5	**SP, HOG, ARS, PERC**, SPEC	91.2.%				
6	**SP, HOG, ARS, PERC, SPEC**, STE	90.2%				

Table 2. Confusion matrices of the 6 individual features and their accuracies, 1-ε. Actual label on rows, expect label on columns.

Type 1: Sedan-S, Van-V, Pickup Truck-P, and Buses-B

	S	V	P	B
Train	134	103	30	13
Test	93	75	24	13
Total	227	178	54	26

Type 2: Light-L, Medium-M, Heavy-H

	L	M	H
Train	200	67	13
Test	157	35	13
Total	357	102	26

ARS: 66.3%

	S	V	P	B
S	65	28	0	0
V	11	63	0	1
P	12	12	0	0
B	0	2	3	8

ARS: 76.6%

	L	M	H
L	157	0	0
M	35	0	0
H	13	0	0

SP: 81.0%,

	S	V	P	B
S	80	11	2	0
V	13	60	2	0
P	2	6	16	0
B	0	2	1	10

SP: 81.0%

	L	M	H
L	157	0	0
M	29	6	0
H	5	5	3

HOG: 83.9%

	S	V	P	B
S	83	8	2	0
V	11	63	1	0
P	2	6	16	0
B	0	3	0	10

HOG: 83.9%

	L	M	H
L	155	2	0
M	24	11	0
H	4	3	6

STE: 46.3%, ε=

	S	V	P	B
S	93	0	0	0
V	75	0	0	0
P	22	1	1	0
B	11	0	1	1

STE: 76.6%

	L	M	H
L	157	0	0
M	35	0	0
H	13	0	0

SPEC: 43.9%

	S	V	P	B
S	88	5	0	0
V	73	2	0	0
P	22	2	0	0
B	13	0	0	0

SPEC: 84.4%

	L	M	H
L	156	1	0
M	25	10	0
H	4	3	7

PERC: 75.6%

	S	V	P	B
S	86	7	0	1
V	20	53	0	2
P	1	16	7	0
B	0	3	1	9

PERC: 86.8%

	L	M	H
L	157	0	0
M	22	13	0
H	2	3	8

selection and classification, we will see both the visual-only and the visual+audio cases. In visual feature combinations, HOG and SP are applied on size-normalized images, but their combination includes both interior and exterior information of vehicles, thus providing some classification improvement. The PERC feature adds acoustic signatures of vehicles in addition to their visual information, thus providing significant improvement over the audio–only result, as well as visual-only results. The testing accuracy using PERC with HOG is also better than using HOG itself, indicating features from two different sources (audio and visual) are better than the single source, even though individually, visuals do better than audio.

In the multimodal level, adding one or more features may improve the classification depending on how well the complementary information that the new features can provide. For example, for the type 1 task, the combination of SP, HOG and ARS (all visual features) increases the accuracy since each of them captures distinct signatures of vehicles. When combining visual features with audio features, the results are also improved, even though just slightly. Based on the results, the accuracies with three modalities, between two different visual-audio combinations (SP, HOG, PERC and SP, PERC, ARS) are almost the same. However, ARS feature has only 3 dimensions whereas HOG uses 162 dimensions. Therefore, if the reconstructed images are accurate, the ARS can be used to replace HOG while combining with other features to reduce computational costs for the vehicle classification task. Between the visual-audio combinations (SP, HOG, PERC) and the visual-only combinations (SP, HOG, ARS), heterogeneous multimodal combinations seem to win, by 1% in this experiment. However, this is not the case for the type 2 task, where adding another type of feature (SPEC) will not improve the performance as original bimodal combination (PERC, HOG). In fact, adding additional "bad" features (such as SP, ARS) will even drop the performance. This is mainly because of the small distinguishability among classes using the shape or size information for the type 2 task.

Finally, an interesting result may be derived. Using two different ways in classifying the vehicles, the features and their combinations from the same feature set will be used differently in order to have the best performance in each case. Then in the decision level, finer vehicle classification can be achieved using a simple decision-level integration strategy. In the examples we used, we can obtain classification results of 12 classes of vehicle, such as light sedan, medium sedan, and etc.

Based on the experimental results, we can summarize our observations as follow:

1. Feature selection is task dependent problem. Same feature may perform differently given different tasks.
2. Multimodal features can improve the performance than using single modalities, especially data from various sources (such as audio and video).
3. Adding a new feature or more features may affect the performance depending on how well the complementary information it or they can provide.
4. Finer classification can be achieved by integrating results in the decision level.

7. Conclusions

In this paper, we first describe the multimodal data alignment using the multimodal temporal panorama for reliable feature extraction. Then various types of visual and audio features are presented. A robust multi-branching sequential forward selection method is provided to select one or more good features at the same levels of feature

combinations (uni-modal, bi-modal, tri-modal, and etc.). We also provide a thorough study on the experimental results given two different tasks using the same set of samples and features, to see the impact of various features for specific tasks, and the possibility to integrate the results to achieve finer classification.

Acknowledgments

This work is supported by AFOSR under Award #FA9550-08-1-0199 and the 2011 Air Force Summer Faculty Fellow Program (SFFP), by NCIIA under E-TEAM grant No. 6629–09, by ARO under DURIP Award # W911NF-08-1-0531, and by a PSC-CUNY Research Award. The work is also partially supported by NSF under award #EFRI-1137172.

References

[1] S. Gupte, O. Masoud, R. F. K. Matrin and N. P. Papanikolopoulos, Detection and classification of vehicles, *IEEE Transactions on Intelligent Transportation System*, vol.3, no.1. p37-47, March 2002

[2] W. L. Hsu, S. H. Yu, Y. S. Chen and W. F. Hu, An automatic traffic surveillance system for vehicle tracking and classification, *IEEE Trans. on Intelligent Transportation Systems*, vol. 7, no. 2, 175-187, 2006

[3] Z. Zhu, G. Xu, B. Yang, D. Shi, X. Lin, VISATRAM: A Real-time vision system for automatic traffic monitoring, *J. of Image and Vision Computing*, pp. 781-794, July 2000

[4] J. Xiao, H. Cheng, H. Sawhney and F. Han, Vehicle detection and tracking in wide field-of-view aerial video, *CVPR 2010*, pp. 679-684, 13-18 June, 2010.

[5] G. Zhang, R. P. Avery, Y. Wang, Video-based vehicle detection and classification system for real-time traffic data collection using uncalibrated video cameras, *Transportation Record: Journal of the Transportation Research Board*, vol 1993/2007, pp. 138-147, Oct 25, 2007

[6] P. Dickson, J. Li, Z. Zhu, A. R. Hanson, E. M. Risemen, H. Sabrin, H. Schultz, G. Whitten, Mosaic generation for under vehicle inspection, *IEEE WACV*, Dec. 2002.

[7] T. Wang, R. Li, Z. Zhu, and Y. Qu, Active stereo vision for improving long range hearing using a laser Doppler vibrometer, *IEEE Computer Society's Workshop on Applications of Computer Vision (WACV)* , Jan 5-6, 2011

[8] T. Wang, Z. Zhu, and C. N. Taylor, Multimodal temporal panorama for moving vehicle detection and reconstruction, *IEEE ISM International Workshop on Video Panorama (IWVP)*, De. 5-7, 2011, California, ISM 2011: 571-576

[9] R. C. Bolles, H. H. Baker, and D. H. Marimont, Epipolar plane image analysis: An approach to determine surface from motion, Int. J. Computer Vision, vol.1, no.7, 1987

[10] J. Y. Zheng, and X. Wang, Pervasive views: area exploration and guidance using extended image media, ACM Multimedia Conference, 986-995, Singapore, 2005

[11] Y. Dedeoglu, B. U. Toreyin, U. Gudukbay and A. E. Cetin, Surveillance using both video and audio, *in Multimodal Processing and Interaction: Audio, Video, Text*, P. Maragos, A. Potamianos and P. Gros Eds., 143-156, 2008

[12] R. Chellappa, G. Qian and Q. Zheng, Vehicle detection and tracking using acoustic and video sensors, *IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP)*, May, 2004

[13] A. Klausner, A. Tengg, C. Leistner, S. Erb and B. Rinne, An audio-visual sensor fusion approach for feature based vehicle identification, AVSS, 2007.

[14] T. Wang and Z. Zhu, Z., Real time vehicle detection and reconstruction for improving classification, *IEEE Computer Society's Workshop on Applications of Computer Vision (WACV)*, January 9-11, 2012, Colorado.

[15] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection. *CVPR,* 2005.

[16] L. Mao, M. Xie Y. Huang and Y. Zhang, Preceding vehicle detection using histograms of oriented gradients. *Int. Conf. on Communications, Circuits and Systems (ICCCAS)*. pp. 354-358, July, 2010

[17] L. Lu, H-J. Zhang and H. Jiang, Content analysis for audio classification and segmentation*, IEEE Trans. On Speech and Audio Processing*, vol 10, no. 7, October 2002

[18] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1993

[19] F. Zheng, G. Zhang and Z. Song, Comparison of different implementations of MFCC, *J. Computer Science and Technology*, 16(6): 582-589, 2001

[20] I. A. Gheyas and L. S. Smith, Feature selection in large dimensionality domains, *Pattern Recognition*, vol. 43, issue 1, pp 5-13, Jan. 2010

[21] C. Cortes and V. Vapnik, Support-vector network, *Machine Learning,* 273-297, 1995

[22] P. K. Atrey, M. A. Hossain, A. El Saddik and M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, *Springer Multimedia Systems Journal*, vol. 16(6), pp:345-379, 2010

[23] N. Babaguchi, Y. Kawai, T. Ogura and T. Kitahashi, Personalized abstraction of broadcasted American football video by highlight selection, *IEEE Trans. Multimed*. 6(4), 575–586, 2004

[24] H. Xu, T. S. Chua, Fusion of AV features and external information sources for event detection in team sports video, ACM Trans. Multimed. Comput. Commun. Appl. 2(1), 44–67, 2006