Audio-Visual Feature Fusion for Vehicles Classification in a Surveillance System

Tao Wang[1], Zhigang Zhu[2], Riad Hammoud[1]
[1] BAE Systems. 6 New England Executive Park, Burlington MA, 01803
tao.wang@baesystems.com; riad.hammoud@baesystems.com
[2] Department of Computer Science, The City College of New York, New York, NY 10031
zhu@cs.ccny.cuny.edu

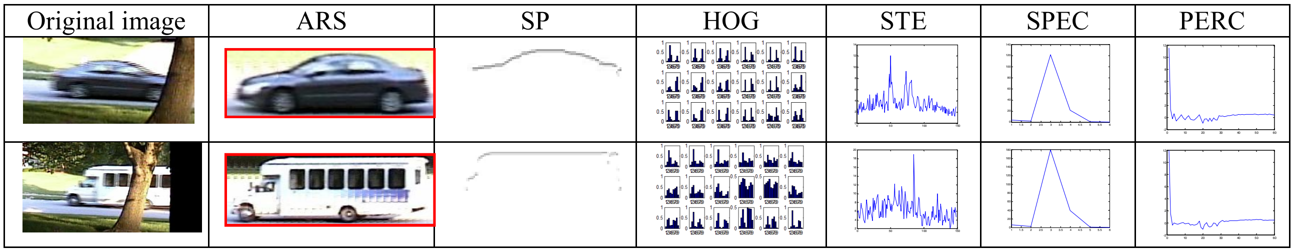

Figure 1. Multimodal features for vehicle classification (row 1: a sedan; row 2: a bus). From left two right: Original video frame, aspect-ratio and size (ARS) on reconstructed image, shape profile (SP), histogram of oriented gradients (HOG), short-time energy (STE) of audio wave, spectrum feature (SPEC) and perceptual feature (PERC) in MFCC.

Abstract

In this paper we tackle the challenging problem of multimodal feature selection and fusion for vehicle categorization. Our proposed framework utilizes a boosting-based feature learning technique to learn the optimal combinations of feature modalities. New multimodal features are learned from the existing uni-modal features which are initially extracted from the data acquired by a novel audio-visual sensing system under different sensing conditions (long range, moving vehicles, and various environments). Experiments on a challenging dataset collected with our long-range sensing system demonstrated that the proposed technique is robust to noise and can find the best among multiple good feature modalities from training in terms of classification performance than the feature modality selection using a sequential based technique which tends to stay on a local maxima.

1. Introduction

Recently, research and development efforts in moving object classification are gradually shifting their emphases from only analyzing visual information to using multiple sensing modalities. Multimodal data or features (Fig. 1), such as visual appearance, motion, range and acoustic signatures, are collected to make a better decision in either detection or classification. Many applications found in biometrics [1], activity recognition [2], traffic monitoring [3], and large area surveillance [4, 5] show that better results can be achieved using multimodalities. These sensor data could be in the forms of not only human signatures (biometrics), but other information such as vehicle signatures, scene description and other context information [6, 7]. Multimodal systems support users multiple ways of responding, according to their preference and needs.

However, some information is redundant, unimportant o even unrelated for a specific task. For example, in moving vehicle classification, if we only want to distinguish vehicles of different shapes, visual features may dominate the decision, and audio is irrelevant to this task. However, if we want to measure the volumes of the engine sound of particular vehicles, such as a truck vs. a minivan, audio information may influence our decision more. Moreover, if there is a large obstacle obscures most part of the vehicle, the audio analysis could be more effective. Therefore, understanding and selecting sensing data in such a system is critical. From the available feature set, which modalities should be selected to accomplish a specified task? The utility of those modalities could be different given different tasks. As the optimal feature subset changes over time, the confidence of the feature modality selected within the task is different. And it is an open problem for multimodal feature fusion and classification. It would be ideal to be able to learn the most representative data or features given a specified task.

Boosting is a rather general approach for improving the performance of any weak classifiers. Usually a weaker classifier is defined as any classifier that can achieve classification accuracy above 50%. Classification performance is boosted by combining many weak classifiers to produce a strong classifier. Based on this learning capability, we propose a boosting-based feature learning technique to automatically select a number of same or different feature modalities for each weak learner, and then the ensemble classifiers can further improve the classification accuracy. For facilitating research and development of algorithms in multimodal feature selection and vehicle classification, we have made our dataset publicly available for research purpose [8]. The contributions of this work include: (1) a unique audio-visual vehicle dataset collected and processed by a long-range multimodal sensing and processing system; (2) a boosting-based feature learning (BBFL) algorithm applied to multiple sensing modalities; and (3) a comprehensive experimental study for both feature/modality selection and vehicle classification.

The rest of paper is organized as follows. Section 2 introduces some related work on feature modality selection. Section 3 describes our boosting based feature learning technique. Section 4 describes the dataset we used to evaluate our algorithm. Then experiment results are provided in Section 5. Finally, conclusions are given in Section 6.

2. Related Work

In term of optimal feature modality selection, the method in [9] first finds statistically independent modalities from raw features, then determines the optimal combination of individual modalities using a super-kernel. When all feature components were combined and treated as a one-vector representation, it suffers from the curse of dimensionality. On the other hand, the large number of separate modalities reduces the curse of dimensionality, but the inter-modality correlation increased. An optimal selection of modalities could balance between the curse of dimensionality and the inter-modality correlation. A summary of the approaches proposed in a few papers can be found in [10].

The fusion of features that are obtained from different modalities usually result into a large feature vector; therefore many feature reduction techniques are applied. Commonly used are principle component analysis (PCA), and linear discriminant analysis (LDA). PCA is used to project higher dimensional data into lower dimensional space while preserving as much information as possible. LDV is used for determining the linear combination of features, which is not only a reduced set of features but it is also used for classification [11]. Many researchers have used these methods for feature vector dimension reduction for the multimodal fusion, for example: PCA in [12] for video classification, singular vector decomposition (SVD) in [13] for biometric person authentication, and adopted LDA [14] for speech recognition.

In the boosting literature, feature fusion is achieved by using the available features to create a new combination of these features. One example is the method in [15], which learned the classifiers using products of decision trees. Alternatively, some [16] suggested the addition, of logical (and, or) combinations of previously selected weak learners; others [17] derived more sophisticated combinations of weak learners for boosting feature selection and extraction. The resulting boosting algorithms grow a predictor by selecting among a pair of pre-defined operations, which could be sums and products or "ands" and "ors", among others. However, their work cannot be directly applied to sensing *modality* selection. In our previous work [3], we described a technique using a multi-branch feature searching (MBFS) to select only the best combination of feature modalities. Here we propose a more robust and general approach using a boosting based feature learning (BBFL) technique to select the optimal combinations of features and modalities for higher performance. Although our framework is a variant of some existing related work in boosting, it is different in aspects of 1) multiple feature/modality selection using boosting base method, 2) flexible feature space and dimensionality, and 3) thorough data collection and analysis in challenging sensing conditions (long-range, moving vehicles, various environments).

3. Boosting based feature learning (BBFL)

The basic idea of the BBFL technique is to not only learn the weak classifiers given input training samples, but also learn the best feature modalities for the weak classifiers. Then the "winner-takes-all" approach selects the best classifier of the corresponding feature modalities in each round. Our method uses an exhaustive search that learns weak classifiers for all feature modalities and their combinations. In our experiments, we use decision trees as the weak classifiers.

The original boost works for binary classification problems. For multiclass problems, a meta-classifier is designed for a general n-class problem. Two straightforward combination schemes are the one-again-all classifier and the one-against-one (or pairwise) classifier [18]. With the one-against-all classifier, n classifiers are trained, each of which is able to distinguish one class from all of the others. At the end, the testing vector is assigned the class corresponding to that of the machine producing the largest positive score. The one-against-one classifier uses $\frac{(n)(n-1)}{2}$ binary classifiers to separate each class from each other class. A voting scheme is then used at the end to determine the correct classification.

In the following, we will present our algorithm on a binary classification problem, which is extended for multiclass problems using the one-against-one technique in our experiments (due to its efficiency). Let $S = (x_i, y_i)_{i=1}^{M}$ be the set of M training data, s.t. $x_i \in R^D$ and $y_i = \{-1, +1\}$ is the corresponding class label. Let *h(x)* be a weak classifier which projects an input vector *x* into {-1, +1} considering only binary classifiers, so that $h_t^j(x_i)$ is the label predicted by the *t-th* weak classifier $h_t(.)$ for the datum x_i using *j-th* feature subset. This can be applied to any real-valued weak classifiers. In fact, one-against-one multiclass problem is considered as a combination of all binary classifiers. Given *J'* uni-modal features, F^J is the set of all possible uni-modal features and their linear combinations into single vectors, where J is the total number of feature subsets. The same weak classifiers *h(.)* are trained on all possible linear combinations of multimodal features at each learning step. A valid weak classifier should have an overall training error rate r larger than 0 (meaning more than 50% correctness), therefore the total number (J) of useful feature subset may be different at each iteration. In each step, only the best classifier who has the largest important factor is selected to boost the ensemble classifier. As a result, the best feature set is obtained.

Before we formally describe the algorithm, let us first summarize all the notations:

$x_i$is a vector the *i-th* training sample
$y_i$is class label of the ith training sample
M is the total number of training samples
$h_t^j(x_i)$is the label predicted by the *t-th* weak classifier $h_t(.)$ for the datum x_i using *j-th* feature subset
w_t is the weight distribution of samples at the *t-th* weak learner.
J' is total number of uni-modal features
$F^J$is the set of all possible uni-modal features and their linear combinations, with
J is the total number of feature subsets.
T is the number of weak learners
$r_t^j$is the overall error rate for the *t-th* weak classifier using *j-th* feature subset
$\alpha_t^j$is the important factor the *t-th* weak classifier using *j-th* feature subset
H(x)is the final ensemble classifier.

The algorithm of exhaustive BBFL can be described as the follows**:**

```
Input: S, T, F^J
Initialize:  t=0; $w_t^i = 1/M$, the weight for the i-th sample
Feature Learning: For t=1 to T:
    (1) For each possible feature $f^j \epsilon F^J, j = 1, \dots, J$
        a. Train a weak classifier $h_t^j(.)$
        b. Compute error: $r_t^j = \sum_i w_t^i y_i h_t^j(x_i) \ / \sum_i w_t^i$
        c. Compute important factor: $\alpha_t^j = \log \frac{1-r_t^j}{r_t^j}$
    (2) Select the best weak classifier $h_t^{j*}$ who has the largest important factor $\alpha_t^{j*}$
    (3) Re-weight samples:
            $w_{t+1}^i = w_t^i \exp\left(-\alpha_t^{j*} y_i h_t^{j*}(x_i)\right) / Z_t$,
        where $Z_t$ is the normalization factor so that $\sum_i w_{t+1}^i = 1$
Output: An ensemble classifier using the best feature subset
        $H^{j*}(x) = argmax_{j*} \sum_{t=1}^{T} \alpha_t^{j*} h_t^{j*}(x_i)$
```

Note that this algorithm is very similar to a classic boosting algorithm, but with an additional feature modality selection at each learning step. The weak classifiers of individual feature modalities and their combinations can be learned independently, and then the one with best classification accuracy is selected at each step. Then the samples are reweighted based on the classifier of the best feature modality/combination at each time. Note that every new learner can use different or similar feature modalities as previous learners, solely based on the learning results. Therefore, the BBFL can provide automatic and optimal feature modality selection.

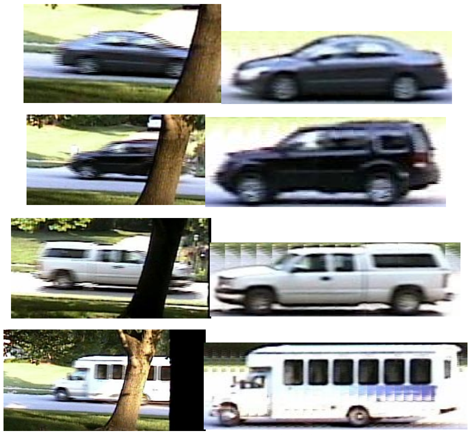

Figure 2. Sample four types of vehicles: sedan, van, pickup truck and bus are shown from top to bottom. Original image shots and their reconstructed results are shown on the left and right columns, respectively.

4. DATASETS

Since there is few public available dataset for multimodal moving vehicles, especially at long range, to compare with, we made our own dataset [8] collected using a long-range multimodal sensing platform we have designed [6], with two pan-tilt-zoom (PTZ) cameras for video and range and a laser Doppler vibrometer (LDV) for acoustic acquisition. The data is collected at very challenging scenarios where there are obstructions from trees and passing-by moving vehicles, motion blur and various perspective views. The acoustic signals are also noisy the inherent nature of the optical system and weak signal returns from a large distance. Vehicles may be out of field of view, therefore we have designed a real-time algorithm to recover the whole of vehicle in motion via automatic detection and reconstruction; otherwise the classification task would be more challenging. Thus, the dataset consists of reconstructed visual images of moving vehicles and audio clips aligned with those detected targets. By the reconstruction, all vehicles visual images have the same side view, heading to the left, with occlusions and motion blurs removed (Fig. 2). The detail of the reconstruction algorithm and analysis can be found in [19]. There are 667 samples vehicles, 400 samples for training and 267 samples for testing. There are four general types of vehicles: sedan, pickup truck, van, and bus. For each general type, there are many variations. For example, sedan type includes 4-door and 2-door cars, or cars with hatchback. Pickup truck type also includes those with wagons or trailers behind. Van type includes those mini-vans, SUVs and regular vans. Bus type consists of school buses and regular trucks. Three visual features are used: aspect ratio and size (ARS), histograms of oriented gradients (HOGs), shape profile (SP), representing simple global scale features, statistical features, and global structure features, respectively. The audio features include short time energy (STE), spectral features (SPECs) which consists of spectral energy, entropy, flux and centroid, and perceptual features (PERCs) are Mel-frequency cepstral coefficients (MFFCs) for the perceptual features. The six features of two vehicle samples are shown in Fig. 1. We mainly test on those 4 general classes, but we can easily extend to more classes. Given enough data for various sub-classes, we can simply categorize them into four general types and then from there we apply classifiers again to distinct them into more detailed classes using a tree-like classification structure.

5. EXPERIMENTAL RESULTS

5.1. Selection results using BBFL

In the boosting framework, we use decision trees as the weak classifiers. We applied up to 20 rounds of weak learners. Fig. 3 shows the training and testing error rates over the numbers (rounds) of weak learners. Table 1 shows the feature modalities selected at each learning step (round) corresponding to Fig. 3. The 2nd column shows the accumulated overall training error rates from previous classifiers. The 3rd column shows the testing errors. The advantages of boosting is that it can select multiple feature modality sets and continuously learn a new feature modality set without stopping at a local maximal one. The last three columns show the top three selected feature modality sets. Note that the training accuracy keeps improving as the number of weak learners increases. However, the best testing error in this experiment (in Fig. 3) stays at 0.2247 (round 9). This is because the decision trees selected are determined only on the re-weight samples in the training dataset. The accuracy of those testing samples, on the other hand, seems to largely depend on the number of possible feature modalities selected. For example, in this experiment in Table 1, the first three weak learners select the same feature combination: ARS+HOG. Since they only use visual signatures, the testing errors are the worst. No. 4 to No. 6 weak learners select the same feature combination: ARS+HOG+PERC. They include some acoustic signatures, so the testing errors reduce. Then No. 8 and No. 9 weak learners select the combinations of the most representative uni-modal feature modalities: ARS, HOG, PERC and SPEC, and those could provide the most complementary information to each other. And using No. 9 weak learners gives the best testing accuracy among all 20 weak learners. Note that when different feature combinations are selected later, such as using No. 12 to No. 18 weak learners, the testing errors actually increase.

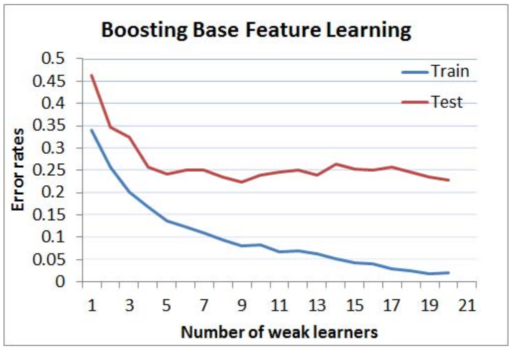

Figure 3.Training and testing errors up to 20 weak learners

Table 1. The first 20 rounds of BBFL using all possible features & combinations (see Fig. 1 for definitions of individual features)

T	Train	Test	Top Best	2nd Best	3rd Best
1	0.3409	0.4627	ARS+HOG	ARS+HOG+STE	ARS+HOG+PERC
2	0.2581	0.3470	ARS+HOG	ARS+HOG+STE	ARS+HOG+SPEC
3	0.2005	0.3246	ARS+HOG	ARS+HOG+STE	ARS+HOG+PERC
4	0.1679	0.2575	ARS+HOG+PERC	ARS+HOG+STE+PERC	ARS+HOG+PERC+SPEC
5	0.1378	0.2425	ARS+HOG+PERC	ARS+HOG+STE+PERC	ARS+HOG+PERC+SPEC
6	0.1228	0.2512	ARS+HOG+PERC	ARS+HOG+STE+PERC	ARS+HOG+PERC+SPEC
7	0.1103	0.2497	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	ARS+HOG+PERC+SPEC
8	0.0952	0.2359	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC	ARS+HOG+PERC
9	0.0800	0.2247	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC	> Err
10	0.0827	0.2388	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC	> Err
11	0.0677	0.2463	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC	> Err
12	0.0702	0.2500	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	ARS+HOG+PERC
13	0.0627	0.2388	ARS+HOG+PERC	ARS+HOG+STE+PERC	> Err
14	0.0526	0.2649	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	> Err
15	0.0426	0.2537	ARS+HOG+PERC	ARS+HOG+SPEC	ARS+HOG+STE+PERC
16	0.0401	0.2500	ARS+HOG+STE+SEPC	ARS+HOG+SPEC	> Err
17	0.0301	0.2575	HOG+PERC+SPEC	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC
18	0.0251	0.2463	ARS+HOG+SPEC	ARS+HOG+STE+SEPC	> Err
19	0.0175	0.2351	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC	> Err
20	0.0201	0.2276	ARS+HOG+PERC+SPEC	ARS+HOG+STE+PERC+SPEC	> Err

Table 2. The best testing results of the boosting based feature learning using the first 9 weak learners. S: sedan, T: pickup truck, V: van, B: bus & truck

Training: 92.00%						Testing: 77.53%				
	S	T	V	B			S	T	V	B
S	165	1	11	0		S	86	1	21	0
T	3	45	3	0		T	2	20	6	0
V	11	3	136	0		V	16	5	91	1
B	0	0	0	22		B	3	0	5	10

Until No. 19 and No, 20 weak learners are used, the testing errors decrease again closing to the error using No. 9 weak learner. The training and testing errors as well as the confusion matrices using No. 9 weak learners are show in Table 2. We have noticed that the number of feature modalities selected at the 2nd best (as well as the 3rd best) is much larger than the top best in training at most weak learning steps. Therefore, this is important since we can safely select minimal number of feature combinations (i.e. only the 1st best) to reduce the amount time in testing.

The classification performance of feature modality selection is also affected by the number of training samples (Fig. 4). With the training size decreases while keeping the same number of testing samples, both the training accuracy and test accuracy decrease. The horizontal axis in Fig 4 from left to right shows from the 100% of original training size used to 60% percentage of that used; the latter is almost equivalent to the size of testing samples. The training accuracy drops 4.3%, and the testing accuracy drops 9.1%.

5.2. Uni-modal feature learning using BBFL

The above results showed several good feature modality combinations that were learned using BBFL. Note that most of the time, multimodal features are selected instead of individual uni-modal features. Here we use the same technique to learn individual uni-modal features without any combinations. If the performance is comparable to that using combinations, then uni-modal could be more efficiency. Table 3 shows the training and testing errors up to 20 rounds of weak learners. We can see that unfortunately the training error doubles, and the testing error increases to more than 1.5 times (from 22.47% to 37.84%). The most top best uni-modal features selected are HOGs which count the interior structure of vehicles. Because the data samples are mainly categorized based on the visual appearance, HOG type of feature performs well inevitably comparing to other uni-modals features. So does for the SP features for the 2nd best selection. However, since both HOG and SP features preserve similar visual information, the combination of both may not produce better classification results. As the result in Table 1 shown, the best combinations include at least one visual feature and one audio feature. It is reasonable since the audio signature can provide complementary information in addition to the visual appearance. At this point we can see that heterogeneous multimodal feature combinations have significant advantages over uni-modal features for this classification task.

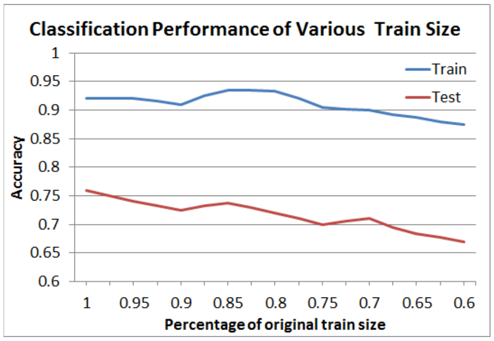

Figure 4.Classification performance of various train sizes.

5.3. Comparison between MBFS and BBFL

The main difference between the MBFS [3] and BBFL algorithms is that the MBFS only selects one best combination of feature modalities, whereas BBFL proposed in this paper selects many feature modality sets. The MBFS starts with selecting the best feature modality from all uni-modal features then combine it with those not selected in the next step. These procedures are repeated until combinations of all feature modalities are evaluated. The worst computation is when all feature modalities or their combinations have similar classification accuracies and fall into the decision boundary, so that all are selected at each level (or each round). So the time complexity for MBFS is $C_{SVM}\sum_{k=1}^{n}\binom{n}{k}$, where $\sum_{k=1}^{n}\binom{n}{k}$is the total number of all possible combinations of n uni-modal feature modalities, and C_{SVM} is the time to evaluate a feature modality using the *support vector machine (SVM)* (for the best performance). This assumes that all feature modalities have same number of vector dimensions. This is always not true. For example, ARS has only 3 dimensions, whereas HOG has 162 dimensions, so the

Table 3. The first 20 iterations of BBFL using six individual feature modalities

T	Train	Test	Top Best	2nd Best	3rd Best
1	0.3500	0.4007	ARS	SP	SPEC
2	0.2750	0.3670	SP	HOG	PERC
3	0.2550	0.3820	HOG	PERC	SP
4	0.2400	0.3670	HOG	PERC	SP
5	0.2225	0.3558	HOG	SP	PERC
6	0.1900	0.3483	HOG	SP	PERC
7	0.1925	0.3708	HOG	SP	PERC
8	0.1800	0.3371	HOG	SP	PERC
9	0.1700	0.3184	HOG	SP	PERC
10	0.1800	0.3296	HOG	SP	PERC
11	01700	0.3446	HOG	SP	PERC
12	0.1525	0.3521	HOG	SP	ARS
…	…	…	Same as above	Same as above	Same as above
20	0.1225	0.3446	HOG	SP	ARS

time to train using ARS feature is much faster than that using HOG feature for the SVM. The BBFL evaluates the same number of re-weighted samples using a number of weak learners. So if the decision tree weak classifier is used, the time complexity for the BBFL will be $\mathrm{TC_{DT}} \sum_{k=1}^{n} \binom{n}{k}$,where T is the number of weak learners and $\mathrm{C_{DT}}$ is the time to evaluate a feature modality using the *decision tree* weak classifier. Even though the classification time using decision trees is much faster than using SVMs on individual feature modalities, the large number (T) of the weak learners will make the computation expensive using the BBFL in feature modality selection. In our experiments as shown above, the time to select the best combination of feature modalities in training using the MBFS is about 0.61 seconds, whereas the time to learn all feature modality sets of 20 weak learners using the BBFL is about 1.43 seconds, which is 2 to 3 times slower than MBFS. The computer that we used has Intel CPU 3.06GHz with installed 4GB memory. So, if the training criteria are not met, larger number of weak learners could be used in order to obtain robust results, and the computational time increases.

For the classification performance, the selected best classifier with the MBFS technique achieves a training accuracy of 88.50% and a testing accuracy of 74.53%. The selected feature combination is ARS+HOG+PERC. The best performance with the BBFL technique achieves a training accuracy of 92.00% and a testing accuracy of 77.53%. The testing accuracy is 3% higher than that of the MBFS, which is achieved with the ensemble of 9 weak learners. Notably, among the 9 weak learners, the most important modalities are ARS+HOG+PERC, which is consistent with the results using the MBFS technique.

6. Conclusions

Multimodal features can have significant improvement in classification over that using single modality. Experimental results show that the combinations of heterogeneous features produce better classification performance in both training and testing. The boosting based feature selection technique not only can learn many best uni-modal features from weak learners; but also learn their combinations to further improve the classification performance. We welcome other researchers to design more effective features, develop novel algorithms, and explore different tasks in classifications and identification, using our multimodal vehicle dataset.

Acknowledgments

This research has been partially funded by NSF through Award #EFRI-1137172, by NCIIA under Award #6629-09 and by AFOSR through Grant #FA9550-08-1-0199, as well as a PSC-CUNY Research Award (Round 41).

References

[1] W. Chen, S.B. Oetomo, L. Feijis, P. Andriessen, F. Kimman, M. Gerates and M. Thielen, "Rhythm of life aid (ROLA): an integrated sensor system for supporting medical staff during cardiopulmonary resuscitation (CPR) of newborn infant," *IEEE Trans. Information Technology in Biomedicine*, iss. 99, 2010.

[2] T. Petsatodis, A. Pnevmatikavakis and C. Boukis,"Voice activity detection using audio-visual information,"*16th Int. Conf. Digital Signal Processing*, August, 2009.

[3] T. Wang and Z. Zhu, "Multimodal and multi-task audio-visual vehicle detection and classification,"*9th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, Sept. 2012.

[4] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, 9(2): 257-267, Feb., 2007.

[5] 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

[6] T. Wang, R. Li, Z. Zhu and Y. Qu "Active stereo vision for improving long range hearing using a laser Doppler vibrometer," *IEEE Workshop on Applications of Computer Vision (WACV)*, pp 564 – 569, Jan 5-6, 2011

[7] Z. Zhu and T.S. Huang, (Eds.). *Multimodal Surveillance: Sensors, Algorithms, and Systems*, Artech House, Norwood, MA, 2007

[8] T. Wang, Z. Zhu, Audio visual vehicle dataset. Website: http://www.cse.ohio-state.edu/otcbvs-bench/

[9] Y. Wu, E.Y. Chang, K.C.C. Chang and J.R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proceedings of ACM Multimedia*, pp. 572–579, 2004.

[10] P.K. Atrey, M.S. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems,"*Multimedia Systems*, vol. 12, no. 3, pp. 239-253, September, 2006.

[11] M. E. Wall, A. Rechtsteiner and L. M. Rocha, "Singular value decomposition and principal component analysis,"Chap.5,pp.91-109,Kluwel,Norwell,MA,2003.

[12] M. Guironnet, D. Pellerin and M. Rombaut, "Video classification based on low-level feature fusion model," in *13th European Signal Processing Conf.* Antalya, Turkey, 2005.

[13] G. Chetty and M. Wagner,"Robust face-voice based speaker identity verification using multilevel fusion,"*Image and Vision Computing*, vo. 26, iss. 9, 2006.

[14] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview,"in*Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[15] B. Kegl and R. Busa-Fekete, "Boosting products of base classifiers," In *ICML*, pages 497-504, 2. 5, 2009

[16] D. Danielsson, B. Rasolzadeh and S. Carlsson, "Gated classifiers: Boosting under high intra-class variation," In *CVPR*, pages 2673-2680, 2. 5, 2011.

[17] M. Saberianand N. Vasconcelos, "Boosting algorithms for simultaneous feature extraction and selection," In *CVPR*, 2012.

[18] D.M.J. Tax and R.P.W. Duin, "Using two-class classifiers for multiclass classification,"*Proceedings of Int. Conf. on Pattern Recognition*, Quebec City, Canada, August, 2002.

[19] T. Wang, Z. Zhu, "Real time moving vehicle detection and reconstruction for improving classification," *IEEE Workshop on Applications of Computer Vision (WACV)*, Colorado, 2012.