MULTIMODAL INFORMATION FUSION FOR VIDEO CONCEPT DETECTION

Yi Wu, Ching-Yung Lin**, Edward Y. Chang*, John R. Smith***

*Electrical & Computer Engineering, University of California Santa Barbara, CA 93106.
*IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532.

ABSTRACT

Video media carries multimodal information including visual, audio, textual data. Considerable research has been focused on utilizing multimodal features for better understanding of video content. However, many problems remain such as how to combine multimodal features and what are the effects of different combinations. In this paper, we propose to find the optimal combination of multimodal information in order to improve the performance of video concept detection using two methods, one is *gradient-descent-optimization linear fusion* and the other is *super-kernel nonlinear fusion*. *Gradient-descent-optimization linear fusion* learns an optimal weighted linear combination of single modalities based on fusing individual kernel matrices with gradient descent techniques. *Super-kernel nonlinear fusion* trains separate classifiers for single modalities as the first step. Once individual models have been designed, *super-kernel nonlinear fusion* learns an optimal non-linear combination of individual models by fusing single-modality classifers. Our experiments show that both methods improve performance significantly on TREC-Video 2003 benchmarks.

1. INTRODUCTION

The growing amount of digital video data is driving the need for more effective methods of video content indexing, searching, and understanding. Recent advances in shot boundary detection, visual feature extraction, and speech recognition are improving the capabilities for effective video analysis. However, video concept detection based on classification techniques is still a challenging research issue.

The major task of concept detection is to use the labeled training video content to classify unknown video content. Video carries multimodal information including visual data, audio, closed caption, speech, etc. At this time, no single descriptor is sufficient to encompass all aspects of video content. There has been considerable research on utilizing multiple features for video concept detection. For example, Dimitrova et al. [6] presented a method for video classification using face and text trajectories based on Hidden Markov Models (HMMs). Reaaijmakers et al. [8] proposed a multimodal classification of news video, which classifies the video content based on visual, audio and textual information. However, very few studies have been made on how to combine multimodal features and what are the effects of different combinations.

In this paper, we study the problem of optimally combining multimodality information in order to improve the performance of video concept detection. We proposed two methods. The first method is *gradient-descent-optimization linear fusion*. The second method is *super-kernel nonlinear fusion*. *Gradient-descent-optimization linear fusion* learns an optimal weighted linear combination of single modalities based on fusing individual kernel matrices with gradient descent techniques. *Super-kernel nonlinear fusion* trains separate classifiers for single modalities as the first step. Then, based on soft decisions that emanate from models constructed independently for individual modalities, *super-kernel nonlinear fusion* learns an optimal non-linear combination of individual models by fusing single-modality classifers. Our extensive empirical studies show that both methods achieve markedly improved performance on TREC-Video 2003 benchmark.

2. GRADIENT-DESCENT-OPTIMIZATION LINEAR FUSION

For each single modality (e.g., visual, audio, text, etc.), we have its feature descriptor and similarity measurement, resulting in an individual kernel matrix where each element of this kernel matrix defines the similarity between two training samples in terms of that specific single modality.

Assume we have a total of D modalities for depicting the targe video concept, and $\mathbf{K}_d$ is the individual kernel matrix defined for the d^{th} modality. Each $\mathbf{K}_d$ extracts a specific type of information from given data, thereby providing a partial view of the data. We formulate the multimodal fusion problem as a convex optimization problem, and propose a *gradient-descent-optimization linear fusion* method to learn the optimal combination fashion. Our linear fusion can provide the flexibility to learn the relative importance of these sources according to a target learning task.

Wu and Chang are supported by NSF grants IIS-0133802 and IIS-0219885.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

Definition 1 (Linear fusion) *We define linear fusion as a convex combination of individual kernel matrices:*

$$\underline{\mathbf{K}} = \sum_{d=1}^{D} \mu_d \mathbf{K}_d$$

$$tr\underline{\mathbf{K}} = c, \;\; \mu \geq 0, \;\; \mathbf{K}_d \geq 0, \;\; d = 1, \cdots, D \qquad (1)$$

where $\underline{\mathbf{K}}$ is the fused kernel matrix, and c is some constant to bound $tr\underline{\mathbf{K}}$ (the trace of $\underline{\mathbf{K}}$). Parameter μ is the weight vector for individual kernels; $\mu \geq 0 \Leftrightarrow \mu_d \geq 0, d = 1, \cdots, D$. $\mathbf{K}_d$ is the individual kernel matrix for the d^{th} modality, and $\mathbf{K}_d$ is positive semi-definite for all $d = 1, \cdots, D$.

Once we get the fused kernel matrix, we use Support Vector Machines (SVMs) [10] as a supervised classification tool for concept modeling because of their good generalization ability. SVMs achieve discrimination by mapping the feature vectors into a higher dimensional space through a linear or nonlinear function and constructing the separating hyperplane with maximum margin.

Proposition 1 *Fused kernel matrix $\underline{\mathbf{K}}$ is positive semi-definite.*

Proof. Since the weights $\{\mu_d\}$ are constrained to be nonnegative and the $\mathbf{K}_d$ are positive semi-definite, thus $\underline{\mathbf{K}} \geq 0$ is satisfied, which means $\underline{\mathbf{K}}$ is positive semi-definite.

The simplest way of doing linear fusion is to exhaustively search the optimal values of $\{\mu_d\}$ by using cross-validation. However, this approach is clearly limited to a small number of parameters and requires expensive computation cost.

We propose a *gradient-descent-optimization linear fusion* method, using a bound on the generalization error and computing the gradient of this bound with respect to kernel parameters. By performing *gradient descent optimization*, we can effectively handle a large number of parameters.

Proposition 2 *If the trace of $\underline{\mathbf{K}}$ is bounded, we need only to take the gradient of squared norm of hyperplane weight vector $\mathbf{w}$ to minimize the generalization error of support vector machines.*

Proof. For a discriminative function $f(x)$ and a kernel matrix $\mathbf{K}$, the proportion of errors on the test data is, with probability $1 - \delta$, bounded by [4]

$$\frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i)) + \frac{1}{\sqrt{n}}\left(4 + \sqrt{2log(1/\delta)} + \sqrt{\frac{tr\mathbf{K}}{n\gamma^2}}\right) \qquad (2)$$

Therefore, to minimize the generalization error of a fixed kernel matrix $\mathbf{K}$, we need to keep the trace of $\mathbf{K}$ fixed and maximize the margin γ [3]. In our case, if we fix the trace of fused kernel matrix $\underline{\mathbf{K}}$ as some constant c, we need only to maximize the margin γ, which is equivalent to minimizing the squared norm of hyperplane weight vector $\mathbf{w}$ [1].

Given a fused kernel matrix $\underline{\mathbf{K}}$, the gradient of $\|\mathbf{w}^2\|$ according to kernel parameters θ can be computed as [5]:

$$\frac{\partial\|\mathbf{w}^2\|}{\partial\theta} = -\alpha^T diag(\mathbf{y}) \frac{\partial\underline{\mathbf{K}}}{\partial\theta} diag(\mathbf{y})\alpha, \qquad (3)$$

where $\mathbf{y} = [y_1, \cdots, y_n]^T$ is the label vector, $diag(\mathbf{y})$ is a diagonal matrix with y_i as its diagonal elements, and $\alpha = [\alpha_1, \cdots, \alpha_n]^T$ is the vector of support values. In our case, θ refers to parameters μ. Obviously, the gradient of $\|\mathbf{w}^2\|$ according to each μ_d can be computed as:

$$\frac{\partial\|\mathbf{w}^2\|}{\partial\mu_d} = -\alpha^T diag(\mathbf{y})\mathbf{K}_d diag(\mathbf{y})\alpha. \qquad (4)$$

By using gradient-descent-optimization linear fusion, we can find not only optimal linear discriminant boundaries but also the optimal values of weights for problems involving multiple kernels. Figure 1 summarizes the algorithm of finding the best linear fusion using a gradient descent optimization algorithm. First, we need to normalize all the individual kernel matrices to the same scale and set the initial values of their weights $\{\mu_d, d = 1 \cdots D\}$ (steps $1 - 3$). Second, we train an SVM to find the optimal value of α with the current kernel matrix $\underline{\mathbf{K}}$ (steps $5 - 6$). Then, we make gradient steps according to Equation 4. The values of $\{\mu_d, d = 1 \cdots D\}$ are updated using gradient descent optimization methods [2], and we make sure that they satisfy nonnegative constraints (steps $7 - 9$). Using the updated weight values, we can form a new matrix $\underline{\mathbf{K}}$ and train another SVM. This iteration stops after at least one of the *gradient termination criteria* has been satisfied (step 4). Here, we set three termination criteria. 1) The maximum number of iterations is reached. 2) The performance has been minimized to the predefined *goal*. 3) The performance gradient falls below predefined *minimum gradient values*.

3. SUPER-KERNEL NONLINEAR FUSION

Gradient-descent-optimization linear fusion learns an optimal linear combination of multimodal information. However, the optimal combination of multimodal information might be non-linear instead of linear. In this section, we explain how *super-kernel nonlinear fusion* learns the optimal non-linear combination by fusing single-modality classifers.

For each single modality (e.g., visual, audio, text, etc.), we first learn its classification model using SVMs, which results in multiple single modality classifiers. After formulating single modality classification models, the next step

[1] SVMs try to maximize the margin $\gamma = 1/\|\mathbf{w}^2\|$ between positive and negative classes. $\mathbf{w} = \sum_{i=1}^{SN} \alpha_i y_i \Phi(x_i)$, in which SN is the number of support vectors [4].

Input:
$\{\mathbf{K}_1, \cdots, \mathbf{K}_D\}$; /* *A set of kernel matrices* */
$\mathbf{y} = \{y_1, \cdots, y_n\}$; /* *Label of training shots* */
Output:
$\underline{\mathbf{K}}$; /* *Fused kernel matrix* */
Variables:
α /* *Support values* */
μ; /* *Optimal weighting vector of* $\{\mathbf{K}_1, \cdots, \mathbf{K}_D\}$ */
Function calls:
SVMTrain($\mathbf{K}, \mathbf{y}$); /* *SVMTrain* */
Initialize(μ); /* *Assign initial weighting values* */
Update(μ); /* *Update vector* μ *along optimal direction* */
Terminate(); /* *Check whether gradient termination criteria are satisfied, yes=1, no=0* */

Begin
1) **for** each kernel matrix $\mathbf{K}_d(x_d, y_d)$
2) $\quad \mathbf{K}'_d(x_d, y_d) \leftarrow \frac{\mathbf{K}_d(x_d, y_d)}{\sqrt{\mathbf{K}_d(x_d, x_d)\mathbf{K}_d(y_d, y_d)}}$;
3) Initialize(μ);
4) **while**(!Terminate()){ /* *Iteration to find optimal linear fusion* */
5) $\quad \underline{\mathbf{K}} \leftarrow \sum_{d=1}^{D} \mu_d \mathbf{K}'_d$;
6) $\quad \alpha \leftarrow$ SVMTrain($\underline{\mathbf{K}}, \mathbf{y}$);
7) $\quad$ **for** each μ_d
8) $\quad\quad \frac{\partial \|\mathbf{w}^2\|}{\partial \mu_d} \leftarrow -\alpha^T diag(\mathbf{y}) \mathbf{K}_d diag(\mathbf{y}) \alpha$;
9) $\quad$ Update(μ);} /* *End of iteration* */
10) **return** $\underline{\mathbf{K}}$;
End

Fig. 1. Gradient-descent-optimization Linear Fusion

is multimodal fusion. Multimodal fusion forms a complete picture of the relationship between different modalities of the original video data. Each shot in the training set has associated confidence scores using the predeployed basis classification models, giving it a vector of model classification confidence scores. Then, we take confidence scores as new features and apply a nonlinear kernel function to form a new kernel matrix. Based on this new kernel matrix, we can learn the non-linear boundary for different modalities. This learning process can be viewed as operating in a new feature space (of classifier scores) and finding a decision boundary [7].

Figure 2 summarizes the algorithm of super-kernel nonlinear fusion. The inputs to the algorithm are shots $\{\mathbf{s}_1, \cdots, \mathbf{s}_n\}$, their labels $\{y_1, \cdots, y_n\}$, For the d^{th} modality, we use the label information and the similarity information encoded in $\mathbf{K}_d$ to train a *discriminative function* f_d (steps $1-2$). This is essentially a supervised learning procedure, which trains a function for each modality to predict semantics for shots.

To determine the best combination of the individual functions, the algorithm records the prediction scores generated by individual functions on each training shot $\mathbf{s}_i$. As a result, $\mathbf{s}_i$ is represented by a vector X_i of D dimensions, generated by the D discriminative functions (steps $3-5$). Next, the algorithm employs a *super-kernel* function SKF to compute similarity between vectors and generate an $n \times n$ super-kernel matrix $\underline{\mathbf{K}}$ (steps $6-8$). The super kernel function SKF can be *Gaussian radial basis kernel function*, *Laplacian kernel function*, or *Polynomial kernel function*.

Algorithm Super-kernel Fusion
Input:
$\{\mathbf{K}_1, \cdots, \mathbf{K}_D\}$; /* *A set of kernel matrices* */
$L = \{\mathbf{s}_1, \cdots, \mathbf{s}_n\}$; /* *A set of training shots* */
$\mathbf{y} = \{y_1, \cdots, y_n\}$; /* *Label of training shots* */
Output:
$\underline{\mathbf{K}}$; /* *Super-kernel matrix* */
Variable:
$\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$; /* *A set of* D *dimensional vectors* */
$\{f_1, \cdots, f_D\}$; /* *A set of discriminative functions* */
Function calls:
$f_d(s_i)$; /* *Prediction score of* s_i *from* f_d */
SVMTrain($\mathbf{K}, \mathbf{y}$); /* *SVMTrain* */
SKF($\mathbf{X}_i, \mathbf{X}_j$); /**Kernel function between two vectors* */

Begin
1) **for** each kernel matrix $\mathbf{K}_d$
2) $\quad f_d \leftarrow$ SVMTrain($\mathbf{K}_d, \mathbf{y}$);
3) **for** each shot $\mathbf{s}_i \in L$
4) $\quad$ **for** each discriminative function f_d
5) $\quad\quad \mathbf{X}_i[d] \leftarrow f_d(s_i)$;
6) **for** each vector $\mathbf{X}_i$,$\mathbf{X}_j$
7) $\quad \underline{\mathbf{K}}_{ij} \leftarrow$ SKF($\mathbf{X}_i, \mathbf{X}_j$);
8) **return** $\underline{\mathbf{K}}$;
End

Fig. 2. Super-kernel Nonlinear Fusion

4. EXPERIMENTS

Our experiment was designed to evaluate the effectiveness of using gradient-descent-optimization linear fusion and super-kernel nonlinear fusion to learn the optimal multimodal information fusion for video concept detection. Specifically, we wanted to answer the following questions:

- Can gradient-descent-optimization linear fusion learn the optimal linear combination of multimodal information effectively and efficiently?
- Can super-kernel nonlinear fusion which learns the optimal nonlinear combination of multimodal information work better than those which use a linear combination?

We conducted our experiments on TREC-2003 Video Track benchmark. TREC-2003 Video Track used 133 hours digital video (MPEG-1) from ABC and CNN news. The concept detection benchmark is summarized as follows: 60% of the video shots were randomly chosen from the corpus to be used solely for the development of classifiers. The remaining 40% is used for concept validation[2].

We chose the NIST Average Precision[3] as our evaluation criteria. A maximum of $2,000$ entries were returned ranked according to the highest probability of detecting the

[2]This sampling ratio worked best in our experiments.

[3]Average Precision is a system-wide number used by NIST to evaluate retrieval systems.

presence of the concept. The ground-truth of the presence of each concept was assumed to be binary (either present or absent in a video shot). Sixteen concepts were defined in this benchmark for concept detection (See Table 1).

For each video shot, we extracted a number of features [1]: *Color histogram*, *Edge orientation histogram*, *Color correlogram*, *Co-occurrence texture*, *Motion vector histogram*, *Visual perception texture*, and *Speech*.

CONCEPT	BSM	CLF	GLF	NLF
AIRPLANE	20.9	**23.8**	23.5	20.1
ANIMAL	6.1	5.5	**8.6**	8.2
BUILDING	4.1	6.4	4.7	**8.4**
FEMALE SPEECH	67.2	67.2	67.2	**67.2**
MADELEINE ALBRIGHT	30.1	**47.4**	33.9	43.3
NATURE VEGETATION	31.3	37.8	33.7	**39.4**
NEWS SUBJECT FACE	7.0	6.3	**7.9**	7.1
NEWS SUBJECT MONOLOGUE	11.8	13.3	8.9	**13.5**
NIST NON-STUDIO SETTING	64.1	68.0	66.4	**69.9**
OUTDOORS	55.4	59.6	53.9	**60.2**
PEOPLE	10.2	11.8	16.4	**18.9**
PHYSICAL VIOLENCE	**1.7**	0.6	1.4	0.2
ROAD	3.5	10.0	**12.4**	8.4
SPORT EVENT	28.9	47.8	40.5	**52.8**
VEHICLE	8.9	16.5	15.6	**16.5**
WEATHER NEWS	43.8	53.6	81.1	**86.7**
Average	24.6	29.8	29.8	**32.5**

Table 1. Average Precision (%) of Video Concept Detection

Table 1 compares the best single modality (BSM), weighted linear combination using cross validation (CLF) [9], gradient-descent-optimization linear fusion (GLF), and super-kernel nonlinear nusion (NLF) based on Average Precision of video concept detection. For the 16 concepts in TREC-Video benchmark, both Linear and Nonlinear Fusion models improved detection performance compared with single modality models. NLF performed 8.0% better than BSM, and GLF and CLF performed 5.0% better than BSM on average. In addition, NLF performed 3.0% better than the linear multimodal combination including GLF and CLF on average.

We also noticed that the average precisions of GLF and CLF were very close, but GLF is more efficient than CLF. Assuming two different modalities for the given concept, we tried ten different values for each weighting factor and performed 5-fold cross-validation. The number of training iterations needed by CLF was $10 \times 10 \times 5 = 500$. Obviously, exhausive search by cross-validation requires expensive computational cost. Table 2 compares the average number of training interations required by GLF and CLF when involving around 4 to 6 modalities for each concept.

MODEL	AVERAGE TRAINING ITERATIONS
CLF	668, 790
GLF	43.5

Table 2. Average Training Iterations

5. CONCLUSION

In this paper, we have proposed two methods to find an optimal combination of multimodal information in order to improve the performance of video concept detection. Both methods train separate classification models for individual modalities as the first step. *Gradient-descent-optimization linear fusion* learns an optimal weighted linear combination of individual modalities based on fusing individual kernel matrices with gradient descent techniques. *Super-kernel nonlinear fusion* learns an optimal non-linear combination of individual modalities based on fusing individual-modality classifiers. Our extensive empirical studies show that both methods achieve markedly improved performance on TREC-Video 2003 benchmarks.

Acknowledgement

We thank Arnon Amir, Shih-Fu Chang, Giridharan Iyengar, Apostol Natsev, Chalapathy Neti, Harriet Nock, Milind Naphade, Winston Hsu, Belle Tseng and Donqing Zhang for providing video features.

6. REFERENCES

[1] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C. Y. Lin, A. Natsev, M. Naphade, C. Neti, H. J. Nock, H. H. Permutery, R. Singhx, J. R. Smith, S. Srinivasany, B. L. Tsengz, T. V. Ashwin, and D. Q. Zhang. Ibm research trec-2002 video retrieval system.

[2] Y. Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, (8), 2000.

[3] O. Bousquet and D. Herrmann. On the complexity of learning the kernel matrix. *Proc. of Advances in Neural Information Processing Systems*, 2003.

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998.

[5] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3), 2002.

[6] N. Dimitrova, L. Agnihotri, and G. Wei. Video classification based on hmm using text and faces. *European Conference on Signal Processing*, 2000.

[7] G. Iyengar, H. J. Nock, and C. Neti. Discriminative model fusion for semantic concept detection and annotation in video. *ACM Multimedia*, 2003.

[8] S. Raaijmakers, J. Hartog, and J. Baan. Multimodal topic segmentation and classification of news video. *IEEE Multimedia and Expo*, 2, 2002.

[9] B. L. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, and J. R. Smith. Normalized classifier fusion for semantic visual concept detection. *IEEE Conf. Image Processing*, 2003.

[10] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.