Margin-Constrained Multiple Kernel Learning Based Multi-Modal Fusion for Affect Recognition

Shizhi Chen and Yingli Tian
Electrical Engineering Department
The City College of New York
New York, NY USA
{schen21, ytian}@ccny.cuny.edu

***Abstract*— Recent advances in multiple-kernel learning (MKL) show the effectiveness to fuse multiple base features in object detection and recognition. However, MKL tends to select only the most discriminative base features but ignore other less discriminative base features which may provide complementary information. Moreover, MKL usually employ Gaussian RBF kernels to transform each base feature to its high dimensional space. Generally, base features from different modalities require different kernel parameters for obtaining the optimal performance. Therefore, MKL may fail to utilize the maximum discriminative power of all base features from multiple modalities at the same time. In order to address these issues, we propose a margin-constrained multiple-kernel learning (MCMKL) method by extending MKL with margin constraints and applying dimensionally normalized RBF (DNRBF) kernels for application of multi-modal feature fusion. The proposed MCMKL method learns weights of different base features according to their discriminative power. Unlike the conventional MKL, MCMKL incorporates less discriminative base features by assigning smaller weights when constructing the optimal combined kernel, so that we can fully take the advantages of the complementary features from different modalities. We validate the proposed MCMKL method for affect recognition from face and body gesture modalities on the FABO dataset. Our extensive experiments demonstrate favorable results as compared to the existing work, and MKL-based approach.**

Keywords-multimodal fusion; affect recognition; multiple kernel learning;

I. Introduction

Recent research demonstrate that affect recognition from multiple modalities achieves better performance [6, 10, 12, 15, 21]. However, most literatures in affect recognition are either only based on features extracted from single modality [12, 16, 21] or just using simply concatenated features from different modalities [6, 10, 15, 22]. How to effectively fuse features from different modalities is still an open question.

The most popular methodology at feature level fusion is the simple concatenation of feature vectors from different modalities to form a large feature vector [6, 10, 15, 22]. However, this fusion method requires a careful design for selections of features and parameters, such as feature dimension etc. This is essentially the manual feature selection.

Shan *et al.* [15] apply the Canonical Correlation Analysis (CCA) to project facial features and body gesture features into a low dimensional space which maximizes their correlation. Then the authors simply concatenate the projected feature vectors together to train a Support Vector Machine (SVM) classifier for affect recognition. However, it is difficult to extend this method to more than two types of features, or base features hereafter, since it needs to find the correlated space between a pair of base features. In this paper, we define base features as a set of base descriptors, which can be combined to form an optimal kernel. In practice, it is very likely to have more than two base features for affect recognition due to the problem complexity.

Gunes and Piccardi [10] select frames, which are the common apex frames from both face and body gesture modalities for affect recognition, and then perform a direct concatenation to combine base features from both modalities. However, the apex frame selection is based on the knowledge of temporal dynamics, which is usually very difficult to predict in advance.

The direct concatenation fusion method is vulnerable to the contamination of less discriminative base features, especially those with large feature dimensions. The multiple-kernel learning (MKL) [1, 13, 18] is able to partially eliminate some drawbacks of the direct concatenation fusion method. MKL provides shielding from the contamination of the less discriminative base features by assigning very large weights to the most discriminative base features. It has recently shown the effectiveness to fuse multiple base features in object detection and recognition [18, 19]. However, MKL tends to select only the most discriminative base features and ignore other less discriminative base features. Therefore, MKL method cannot fully take the advantages of all types of base features from multiple modalities, which provide complementary information.

Moreover, MKL usually employs Gaussian RBF kernels for mapping each base feature to its high dimensional space $\mathcal{H}$. Generally, base features from different modalities require different kernel parameters to achieve their optimal performance. One of the reasons is due to the significant different feature dimensions from multiple modalities. Therefore, MKL may not utilize the maximum discriminative power of all types of features from multiple modalities at the same time.

In order to address these issues, we propose a margin-constrained multiple-kernel learning (MCMKL) by applying (1) additional margin constraints, and (2) dimensionally normalized RBF kernels (DNRBF).

The margin of a separating hyper-plane in SVM literature [3] is defined as the perpendicular distance between the support vectors of two classes. The large margin, obtained in the training set, usually indicates that the underlying feature is discriminative. If the underlying feature is not discriminative, the margin is usually small. So we use the margin to measure the discriminative power of each base feature. These margins then serve as rough guide to MKL when learning each base feature's weight.

By dimensionally normalizing RBF kernels, MCMKL eliminates the influence of the feature dimension difference in multiple modalities. Then, the optimal high dimensional space $\mathcal{H}$ of each base feature from different modalities corresponds to similar kernel parameter. Therefore, MCMKL is able to utilize the maximum discriminative power of all feature types from multiple modalities.

Unlike conventional MKL method, our proposed MCMKL is able to learn the most discriminative base features while still considering other base features, which are less discriminative, but can potentially provide complementary information. We apply MCMKL method on affect recognition from multiple modalities (e.g. both face and body gesture). The extensive experimental results on the FABO (Face and Body Gesture) facial expression database [9] demonstrate the effectiveness of the proposed method for multi-modal feature fusion.

II. Margin-Constrained Multiple Kernel Learning

A. *Multiple Kernel Learning*

Given a set of base features and their associated base kernels $\boldsymbol{K_k}$, we want to find the optimal kernel combination $\boldsymbol{K}_{opt} = \sum_k d_k {*} \boldsymbol{K}_k$, where d_k is the weight for the k^{th} base feature. The kernel combination $\boldsymbol{K}_{opt}$ approximates the best trade-off between the discriminative power and the invariance for a specific application.

Equations (1) to (3) show the objective cost function f and its constraints for the multiple-kernel learning (MKL) proposed in [18].

$$\underset{\mathbf{w},\xi_i,d_k}{Min} \quad f = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i + \sum_k \sigma_k d_k \tag{1}$$

subject to

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b) - 1 + \xi_i \geq 0 \tag{2}$$

$$\xi_i \geq 0 \ \ \forall i\,;\ d_k \geq 0 \ \ \forall k\,;\ \ \mathbf{Ad} \geq \mathbf{p} \tag{3}$$

where $\mathbf{\Phi}(x_i)$ is the combined features corresponding to $\boldsymbol{K}_{opt}$ in a high dimensional space for sample x_i, which is shown in Eq. (4). Equivalently, $\boldsymbol{K}_{opt}$ can be expressed in Eq. (5), where $\mathbf{\Phi}_k(x_i)\cdot\mathbf{\Phi}_k(x_j)$ forms the k^{th} base kernel $\boldsymbol{K}_k$.

$$\mathbf{K}_{opt}(x_i,x_j) = \mathbf{\Phi}(x_i)\cdot\mathbf{\Phi}(x_j) \tag{4}$$

$$\mathbf{K}_{opt}(x_{i,}x_j) = \sum_k d_k * \mathbf{\Phi}_k(x_i)\cdot\mathbf{\Phi}_k(x_j) \tag{5}$$

The optimization can be carried out in a SVM framework subject to additional regularization term of weight d_k in the objective function.

In order to handle large scale problems involving many base kernels, the minimax optimization strategy [5, 13, 18] is used in two iteration steps. In the first step, feature weight d_k is fixed, i.e., $\boldsymbol{K}_{opt} = \sum_k d_k{*}\boldsymbol{K}_k$ is fixed. Then, the optimization problem of Eq. (1) can be solved by any standard SVM solver using its dual form as in Eq. (6) since the term $\sum_k \sigma_k d_k$ is simply a constant.

$$\underset{\partial_i}{Max} \ f_D = \sum_i \partial_i - \frac{1}{2}\sum_{i,j}\partial_i\partial_j y_i y_j \mathbf{K}_{opt}(x_i,x_j) + \sum_k \sigma_k d_k \tag{6}$$

subject to

$$0 \leq \partial_i \leq C\,;\ \ \sum_i \partial_i y_i = 0 \tag{7}$$

In the second iteration step with the fixed α_i, projected gradient descent is employed to find updated feature weights d_k as shown in Eqs. (8) and (9).

$$\frac{\partial f}{\partial d_k} = \frac{\partial f_D}{\partial d_k} = \sigma_k - \frac{1}{2}\sum_{i,j}\partial_i\partial_j y_i y_j \mathbf{K}_k(x_i,x_j) \tag{8}$$

$$d_k^{new} = d_k^{old} - \frac{\partial f}{\partial d_k} \tag{9}$$

These two iteration steps are repeated until converge or the maximum number of iterations is reached. The final weights of base features can be determined.

Then we train SVM classifiers using the optimal combined kernel according the final weights of base features. The class label of a new sample x may be determined by the sign of Eq. (10).

$$g(x) = \sum_i \partial_i y_i \sum_k d_k K_k(s_i,x) + b \tag{10}$$

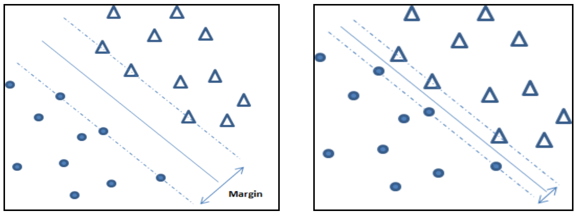
where s_i is the support vector. The multi-class problem can be solved by "one vs. one" or "one vs. all" strategy similar to SVM.

MKL method provides an elegant framework to fuse many base features by assigning larger feature weight to the most discriminative base feature. Compared to the direct concatenation method, MKL can avoid the contamination from less discriminative base features, especially when those features have large dimensions.

However, MKL method tends to select very few base features from the feature pool. It often only selects one or two base features, which are discriminative at a particular high dimensional space $\mathcal{H}$. Moreover, the kernel parameter associated with the optimal high dimensional space $\mathcal{H}$ may be significantly different for the base features from different modalities. Therefore, traditional MKL cannot utilize the maximum discriminative power of each base feature at the same time.

B. Margin Constraints

To address these issues, we propose a Margin-Constrained Multiple Kernel Learning (MCMKL) method. This is motivated by the observations that base feature which is more discriminative usually finds a hyper-plane with larger margin to separate support vectors of opposite classes during training of SVM machines. A hyper-plane of base feature "*a*" in Figure 1(a) has a larger margin than that of base feature "*b*" in Figure 1(b) to separate the class of solid dot from the class of triangle. This suggests that the base feature "*a*" is more discriminative than the base feature "*b*" for the classification of the solid dot and the triangle class.



(a) (b)

Figure 1: (a) The hyper-plane of base feature "*a*" has a large separation margin to separate solid dot class and triangle class; (b) The hyper-plane of base feature "*b*" has a small margin to separate solid dot class and triangle class.

Therefore, the separation margin for each base feature in its high dimensional space provides a rough measurement on the base feature's discriminative power. Nevertheless, these rough measurements can effectively guide MKL when searching for the optimal feature combination. The separation margin for each base feature can be calculated using Eq. (11) as the inversed square root of its own objective cost function.

$$m_k = \frac{2}{\|\mathbf{w}_k\|} \approx \frac{\sqrt{2}}{\sqrt{f_k}} = \frac{\sqrt{2}}{\sqrt{\frac{1}{2}\|\mathbf{w}_k\|^2 + C\sum_i \xi_i + \sigma_k d_k}} \qquad (11)$$

After obtaining the separation margin m_k for each base feature, we select one of the base feature as the reference base feature, which has the feature weight of d_s and the margin m_s. The weight d_k of k^{th} base feature is constrained in the range, which has the lower bound of LB_k and the upper bound of UB_k according to the margin ratio between m_s and m_k during training. LB_k and UB_k can be calculated as in Eq. (13). The additional weight constraints in Eq. (12) are enforced during the multiple-kernel learning.

$$LB_k \le d_k \le UB_k \qquad \forall k \qquad (12)$$

$$LB_k = \left(\frac{m_k}{m_s}\right)^n * d_s\,; \quad UB_k = \left(\frac{m_k}{m_s}\right)^n * d_s * (1+\delta) \qquad (13)$$

where n is a parameter that controls the margin sensitivity on the feature weight ratio between d_k and d_s. As n increases, the values of LB_k and UB_k become more sensitive to the ratio of m_k and m_s. δ is a constant to control the range width of the feature weight d_k. In our experiments, we set n to 1.5 and δ to 1.

C. Dimensionally Normalized Kernel

Gaussian RBF kernel is one of the most popular non-linear kernels due to its excellent performance in numerous applications. It is defined in Eq. (14).

$$\mathbf{K}(x_i, x_j) = \exp(-\gamma \sum_{q=1}^{D} (x_{i,q} - x_{j,q})^2) \qquad (14)$$

where x_i and x_j are the i^{th} sample and the j^{th} sample along with $x_{i,q}$ and $x_{j,q}$ as the q^{th} element in a feature vector. D is the sample's feature dimension.

γ is the RBF kernel parameter, which determines the mapping from a low dimensional feature space $\mathcal{L}$ to a high dimensional space $\mathcal{H}$.

Assuming that feature vectors x_i and x_j have been properly normalized between 0 and 1 along each feature dimension [4], the kernel value decreases when the feature dimension increases at a fixed γ as shown in Eq. (14). Hence, Eq. (14) suggests the inverse relationship between the optimal γ and the feature dimension. This intuition is confirmed in our experiments, which will be analyzed in Section 4.

In MKL fusion, base features from different modalities may have significantly different feature dimensions, which will result very different optimal γ values for each base feature. Therefore, MKL cannot utilize the maximum discriminative power of all base features from different modalities at the same time.

We can treat γ as a feature selection parameter in MKL, which select only few base features at a time. This intuition also explains the observations reported in [18] that MKL tends to select only very few most discriminative base features. Therefore, MKL cannot take the full advantages of all types of features from multiple modalities.

Based on these observations, we propose a dimensionally normalized RBF kernel (DNRBF) which is defined in Eq. (15).

$$\mathbf{K}(x_i, x_j) = \exp(-\frac{\gamma}{D} \sum_{q=1}^{D} (x_{i,q} - x_{j,q})^2) \qquad (15)$$

This normalization step is essential to eliminate the effect of feature dimension on γ selection, so that all base features have a similar optimal γ. Therefore, MCMKL can utilize the maximum discriminative power of all base features from multiple modalities.

III. MULTI-MODAL FUSION FOR AFFECT RECOGNITION

Affect recognition from multiple modalities is a challenging problem. Our study focuses on fusion of features from visual modalities, i.e., face and body gesture modality.

Different from conventional approaches to fuse features from multiple modalities, which simply concatenate all feature vectors from different sources together and feed the concatenated feature vector into a classifier, such as SVM, we apply a margin-constrained multiple-kernel learning (MCMKL) method to fuse features from both face and body gesture modalities. MCMKL can effectively combine all types of features for affect recognition by assigning an appropriate feature weight to each type of features and calculate the optimal kernel for affect recognition.

A. *Overview of MCMKL-based Affect Recognition*

Figure 2 shows an overview of our affect recognition system, which consists of five major parts, i.e., facial feature extraction, body gesture feature extraction, expression temporal segmentation, temporal normalization, and MCMKL-based classification.

Two types of facial features, i.e., Image-HOG and MHI-HOG [6] are extracted in our experiments. Here, HOG stands for Histogram of Gradients [8], and MHI stands for Motion History Image [2, 17]. Image-HOG features capture facial appearance changes, while MHI-HOG features represent facial motion information.

Four types of gesture features are extracted, which include location features, motion area features, Image-HOG features, and MHI-HOG features around both hands.

Each expression in video sequences can be first temporally segmented into onset, apex, offset and neutral phases [6]. Then, we perform a temporal normalization procedure to handle different temporal resolutions of expressions. Finally MCMKL method is employed to find the optimal feature combination and recognize affects.

B. *Facial Features*

Active Shape Model [7, 20] is first applied to track 53 facial landmark points including brows, eyes, nose, mouth, and face contour, as shown in Figure 3(a). Then we locate the corresponding positions of the facial points in the Motion History Image (MHI), as shown in Figure 3(b).

The next step is to extract Image-HOG and MHI-HOG features on original video frames and the corresponding MHI images respectively.

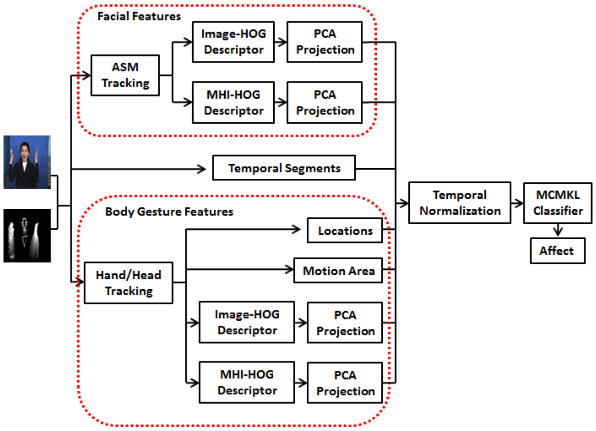
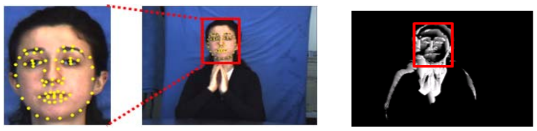


Figure 2: The overview of our proposed MCMKL-based multi-modal fusion for affect recognition through both face and body gestures.

We use 48 by 48 pixels patches with the number of orientation bin equals to 6 and 8 for the Image-HOG and the MHI-HOG features respectively. The MHI image captures motion information of each selected facial point, while the original video frame conveys the appearance information. Finally, we concatenate the Image-HOG descriptor of all the 53 facial points and apply Principal Component Analysis (PCA) to reduce the feature dimension of the concatenated Image-HOG feature from 2862 to 40. Similarly, we can obtain the MHI-HOG descriptor for the corresponding frame and reduce the feature dimension of the concatenated MHI-HOG from 3816 down to 40 for each frame.

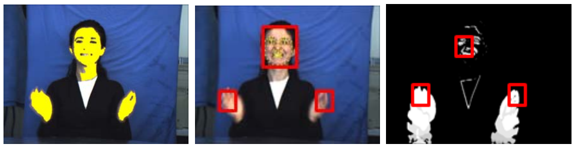


(a) (b)

Figure 3: (a) Facial landmark points tracking; (b) Motion History Image.

C. *Body Gesture Features*

To extract body gesture features, we first track both hands and head in an expression video. The head position is simply the center point of the facial points from the ASM model (see Figure 4(b)). To track hands, we apply a skin color detection [11] followed by the removal of the face regions, which has already been tracked by the ASM model as shown in Figure 4(a) and 4(b).



(a) (b) (c)

Figure 4: (a) skin color detection; (b) head and hand position in the original video frame; (c) head and hand positions in the MHI image.

In addition to the positions of head and hands, we also calculate the motion areas (e.g. the numbers of motion pixels in MHI image) within the detected regions of head and hands. Figure 4(c) shows the head and hand regions in a MHI image.

We further extract Image-HOG and MHI-HOG features in hand regions by uniformly sampling interest points. Then a bag of words representation with the codebook size of 80 is used to describe the distribution of Image-HOG and MHI-HOG features of hand regions. Finally, we perform PCA to reduce their feature dimensions.

D. *Temporal Segmentation*

An expression is a sequence of facial movements which can be roughly described by neutral, onset, apex and offset temporal segments.

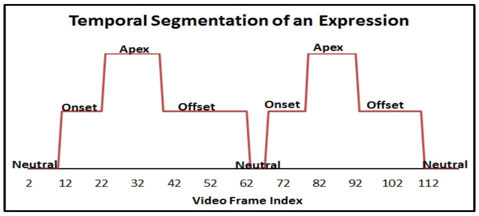


Figure 5: temporal segmentation of an expression video.

Figure 5 shows a sample of the ground truth temporal segmentation of an expression video. The temporal segmentation procedure is necessary to accurately model the expression dynamics, which has been proven crucial for facial behavior interpretation [14]. In our experiments, we simply use the ground truth temporal segmentation and the affect recognition is performed on the complete expression cycle, i.e., onset, apex and offset.

E. Temporal Normalization

In general, the temporal resolution of an expression is generally different when performed by different people. Even same expression performed by same person at a different time, the temporal resolution may not be the same. In order to resolve this time resolution issue in expression videos, we adopt the temporal normalization approach by normalizing all types of features over a complete expression cycle.

The temporal normalization over an expression cycle can be easily implemented by linear interpolation over frame's feature vector along the temporal direction.

F. MCMKL Based Multi-Modal Feature Fusion

Features from multiple modalities may have different forms. Therefore dimensions of different types of features may vary significantly. Our proposed margin-constrained multiple kernel learning (MCMKL) method can effectively fuse all base features from different modalities, i.e., face and body gesture modality, by assigning a feature weight to each base feature.

We concatenate the Image-HOG and the MHI-HOG of facial points as one base feature, i.e., the face feature. The other four base features are from the gesture channel, i.e., location, motion area, and both hands' Image-HOG and MHI-HOG features. Using the margin of each individual base feature as a guide, along with the DNRBF to synchronize the optimal kernel parameter, the MCMKL learns the optimal combined kernel by selecting a proper weight for each base feature during the fusion.

For our multi-classes application, we choose one vs. one classification, and then using the maximum voting scheme to label testing samples.

IV. EXPERIMENTS

A. Experimental Setups

We use a bi-modal face and body gesture database, i.e., FABO database in our experiments [9]. The database is collected using two cameras, i.e., one for face and one for body gesture in a laboratory environment. A sample video is shown in Figure 6. However, we only employ the videos captured by the body camera to extract features for both modalities, since the videos from the body camera already contain both face and body gesture information.

After removing the categories in the database with very small number of samples, there are 8 expression categories, i.e., "Anger", "Anxiety", "Boredom", "Disgust", "Fear", "Happiness", "Puzzlement", and "Uncertainty". The total number of videos used in our experiment is 255 and each video has 2 to 4 complete expression cycles.

We randomly divide the videos into three subsets. Two subsets are used in training and the remaining subset is used in testing. No same video appears in both training and testing. But same subject may appear in both training and testing due to the random selection process.

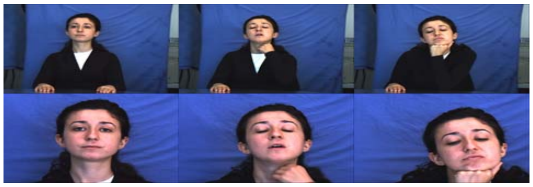


Figure 6: sample video in FABO database recorded by body (top) and face (bottom) camera;

B. Comparison to Existing Work and MKL

In order to evaluate the effectiveness of the proposed MCMKL fusion method, we compare it to the most recent work on the FABO database [6] by using same features, and same training and testing dataset. We further compare MCMKL with MKL method. The performance of the comparison is displayed in Figure 7.

The five base features are used in our experiments, which include face feature, location feature, motion area feature, Image-HOG and MHI-HOG feature of both hands. The face feature is the concatenation of the Image-HOG and the MHI-HOG from the face modality. The Table 1 shows the corresponding feature dimension for each base feature. These base features are fused through the concatenation, MKL, and MCMKL methods.

Table 1: The feature dimension for each base feature, i.e., Face, Loc (location), MA (motion area), Img-HOG (Image-HOG from gesture), and MHI-HOG (from gesture).

Base Feature	Face	Loc	MA	Img-HOG	MHI-HOG
Dimension	2400	180	90	120	30

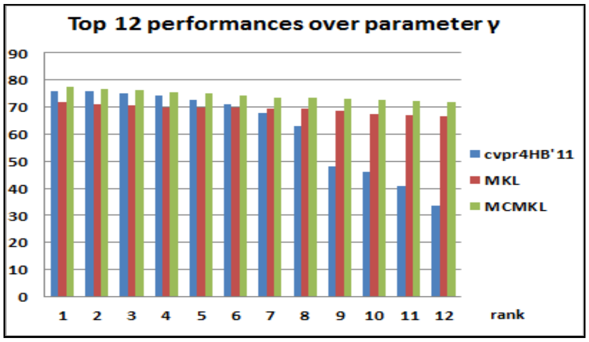
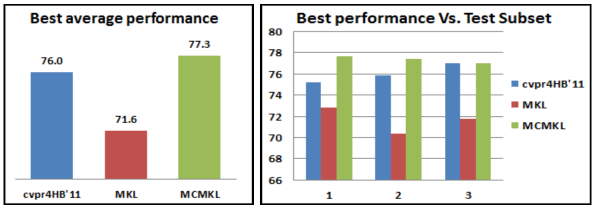


Figure 7: The average performance of the top 12 ranks by sweeping kernel parameter log2(γ) from -15 to 8 for each of the three methods, i.e., concatenation (cvpr4HB'11), MKL, and MCMKL.

To make a fair comparison, we sweep kernel parameter log2(γ) from -15 to 8, and select the top 12 performances for each fusion method. Then we rank these 12 performances by a descending order of their accuracy. We repeat same experiment for three different subsets and the average performances are reported in Figure 7. Figure 8 shows more details of the rank 1 comparison, i.e., the comparison of the best performances for the three fusion methods: MCMKL, MKL, and direct feature concatenation (cvpr4HB'11).

MCMKL outperforms the other two methods over all the top 12 ranks, as shown in Figure 7. If we look at the rank 1 comparison in Figure 8, we can see that the proposed MCMKL achieves better performance than the concatenation method. Note that the five base features have been carefully selected, and the parameters, e.g., PCA projection dimensions etc. are also carefully chosen for the concatenation fusion method in [6]. On the other hand, MCMKL effectively select those feature vectors, and it can still outperform the concatenation fusion method by the average of 1.3%.



(a) (b)

Figure 8: (a) The best average performance by sweeping kernel parameter log2(γ) from -15 to 8 for each fusion method, i.e., cvpr4HB'11, MKL, and MCMKL; (b) The best performance of the three fusion method in three different testing subsets.

Our proposed MCMKL outperforms traditional MKL method by an average recognition rate of 5.7% on these base features, as shown in Figure 8(a). Figure 8(b) shows the performance comparison over three different testing subsets.

C. *Evaluate Feature Weight Distribution*

In this section we verify that our proposed MCMKL is more effective than MKL to incorporate less discriminative features, which provide complementary information to the base features with the maximum discriminative power. We select the kernel parameter γ of 2-15 for MKL and 2-1 for MCMKL method, which yield the best performance for MKL and MCMKL method respectively.

Since we choose one vs. one strategy for our multi-class expression classification, the total number of models we need to train is C_2^n, where n is the total number of expression classes in the dataset. Therefore, we have trained 28 models for 8 categories of expressions, in which each model contains one set of feature weights for the base features, i.e., face, location (loc), motion area (MA), and both hands' Image-HOG (imgHOG) and MHI-HOG (mhiHOG) features. Then we take the mean feature weight of the 28 models over each base feature, followed by the proper normalization.

Figure 9 shows the distribution of average feature weights over the 5 base feature types for MKL and MCMKL method. As expected, MKL selects only the most discriminative base feature, i.e., face feature. More specifically, it assigns more than 98% of the total feature weights to the face feature. The MKL method ignores all the gesture features, i.e., location, and motion area etc., even though these gesture features have been proven to provide complementary information to the face feature [6].

As shown in Figure 9, the proposed MCMKL obtains a more reasonable feature weight distribution. Similar to MKL methods, it recognizes the face feature as the most discriminative base feature by assigning the largest feature weight of 48%. At the same time, it also incorporates other less discriminative gesture features according to their discriminative power.

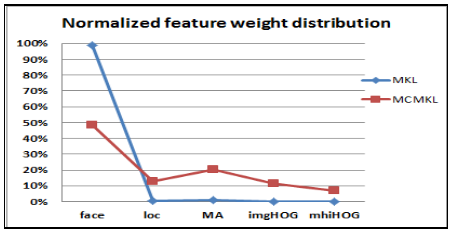


Figure 9: Comparing the average feature weight distribution of the 5 base features, i.e., face, location (loc), motion area (MA), and both hands' Image-HOG (imgHOG) and MHI-HOG (mhiHOG) features, for MKL and MCMKL methods.

Figure 9 has verified the effectiveness of the proposed margin constraints and the dimensionally normalized RBF kernel (DNRBF). It is obvious that the additional constraints on the feature weights according to the separation margin of each base feature can enforce the model to assign small weights to the less discriminative base features. However, it may not be intuitive how the DNRBF contribute to a more reasonable feature weight assignment.

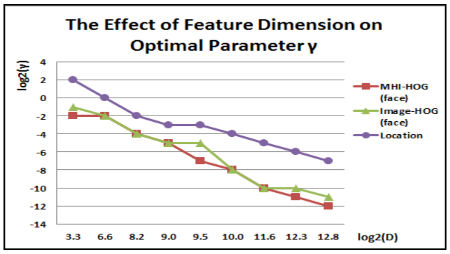


Figure 10: The effect of feature dimension of three base features over the selection of the optimal RBF kernel parameter γ.

Before we provide such intuition, we examine the relationship between the optimal γ value and feature dimension experimentally. We select three base features, i.e., facial point's MHI-HOG, the facial point's Image-HOG and the location feature. Then we manually vary the PCA dimension of the Image-HOG and the MHI-HOG, or the number of normalization frames of the location feature, so that their feature dimensions can be gradually increased. Then we use SVM's 5-fold cross validation to find out the optimal kernel parameter γ for each of the three base features at the selected feature dimension. Figure 10 has suggested the inverse relationship between the optimal γ value and the feature dimension, which has verified our analysis in section 2.

In the experiments of the last section, the most discriminative feature, i.e., the face feature, has the optimal γ of 2-15. Since other less discriminative gesture feature has much smaller feature dimension as we can see from Table 1. Their optimal γ value is much larger. Therefore, at the γ of 2-15, the other gesture features has almost no discriminative power since their optimal γ values are very far away from 2-15. Therefore, MKL method assigns almost zero feature weights to other gesture features.

After we perform the dimensionally normalization as in Eq. (15). The optimal γ values become very close for different base features regardless the differences of their feature dimensions. Therefore, MCMKL can utilize the maximum discriminative power of all base features at the same time. That is another reason why MCMKL method can incorporate other less discriminative base features, which provide complementary information.

D. Contamination from Less Discriminative Features

In this section, we examine the contamination from the less discriminative base features, particularly those with large feature dimensions. From the feature weight distribution in Figure 9, we know that the Image-HOG and MHI-HOG of hands are the least discriminative features. So we intentionally increase their feature dimension to 1200 by including more PCA dimensions. At the same time, we also decrease the dimension of the most discriminative feature, i.e., the face feature, down to 90. Now, we also sweep kernel parameter log2(γ) and select the top 10 performances for each fusion method, i.e., concatenation, MKL, and MCMKL. Then we rank these 10 performances by the descending order of their accuracy. The experimental results are shown in Figure 11.

We observe that the rank 1 result of the MCMKL method outperforms the concatenation fusion method by almost 10%, which indicate that MCMKL method is more effective to shield the contamination from the less discriminative base features, as compared with the concatenation fusion method.

V. CONCLUSION

In this paper, we have proposed a margin-constrained multiple-kernel learning (MCMKL) method which extends the multiple-kernel learning (MKL) method by constraining feature weight range according to the separation margin of each base feature. The dimensionally normalized RBF kernel (DNRBF) is also proposed and employed in MCMKL in order to fuse the features from multiple modalities, which is possible to have very different feature dimensions. Our experimental results demonstrate favorable results as compared to the state-of-the-art results on the FABO database. We also demonstrate the significant improvement as compared to the conventional MKL method.

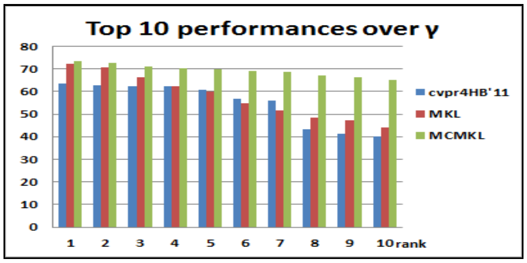


Figure 11: The average performance of the top 10 ranks by sweeping kernel parameter γ for each of the three methods, i.e., concatenation (cvpr4HB'11), MKL, and MCMKL.

ACKNOLEDGEMENT

This work was supported in part by NSF IIS-0957016.

REFERENCES

[1] F. Bach, G. Lanckriet, M. Jordan, “Multiple kernel learning, conic duality and the SMO algorithm”, NIPS, 2004.

[2] A. Bobick and J. Davis, “The recognition of human movement using temporal templates”. PAMI, 2001.

[3] C. Burges, “A tutorial on support vector machines for pattern recognition”, Data Mining and Knowledge Discovery, 2(2), 121–167, 1998.

[4] C. Chang and C. Lin, “LIBSVM : a library for support vector machines”, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, “Choosing Multiple Parameters for Support Vector Machines”, Machine Learning, 2002.

[6] S. Chen, Y. Tian, Q. Liu, D. Metaxas, “Recognizing Expressions from Face and Body Gesture by Temporal Normalized Motion and Appearance Features”, IEEE Int'l Conf. Computer Vision and Pattern Recognition workshop for Human Communicative Behavior Analysis (CVPR4HB). 2011.

[7] T. Cootes, C. Taylor, D. Cooper and J. Graham, “Active Shape Models – Their Training and Application”, Computer Vision and Image Understanding, 1995.

[8] N. Dalal, B. Triggs, “Histogram of Oriented Gradients for Human Detection”, CVPR 2005

[9] H. Gunes and M. Piccardi, “A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior”, International Conferenece Pattern Recognition, 2006.

[10] H. Gunes and M. Piccardi, “Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display”, IEEE Transaction on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 39, NO. 1 2009.

[11] J. Kovac, P. Peer, F. Solina, “Human Skin Colour Clustering for Face Detection”, EUROCON – Computer as a Tool, 2003.

[12] M. Pantic, L. Rothkrantz, “Automatic Analysis of Facial Expressions: The State of the Art”, PAMI, 2000.

[13] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, “More Efficiency in Multiple Kernel Learning”, ICML, 2007.

[14] K. Schmidt and J. Cohn, “Human Facial Expressions as Adaptations: Evolutionary Questions in Facial Expression Research”, Yearbook of Physical Anthropology, 2001.

[15] C. Shan, S. Gong and P. McOwan, “Beyond facial expressions: learning human emotion from body gestures”, British Machine Vision Conference, 2007.

[16] Y. Tian, T. Kanade, J. Cohn, “Facial Expression Analysis”, Handbook of Face Recognition, pp. 247-276, Springer, 2005.

[17] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, “Hierarchical Filtered Motion for Action Recognition in Crowded Videos”, IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews, 2011.

[18] M. Varma, D. Ray, “Learning The Discriminative Power-Invariance Trade-Off”, ICCV, 2007.

[19] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, “Multiple Kernels for Object Detection”, ICCV, 2009.

[20] Y. Wei, “Research on Facial Expression Recognition and Synthesis”, Master Thesis, 2009, software available at: http://code.google.com/p/asmlibrary.

[21] Z. Zeng, M. Pantic, G. Roisman, T. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”, PAMI, 2009.

[22] X. Zhou, B. Bhanu, “Feature fusion of side face and gait for video-based human identification”, Pattern Recognition 41, 2008.