Subject Adaptive Affection Recognition via Sparse Reconstruction

Chenyang Zhang
Media Lab, The City College of New York
New York, NY 10031
czhang2@ccny.cuny.edu

Yingli Tian
Media Lab, The City College of New York
New York, NY 10031
ytian@ccny.cuny.edu

Abstract

Multimedia affection recognition from facial expressions and body gestures in RGB-D video sequences is a new research area. However, the large variance among different subjects, especially in facial expression, has made the problem more difficult. To address this issue, we propose a novel multimedia subject adaptive affection recognition framework via a 2-layer sparse representation. There are two main contributions in our framework. In the subjective adaption stage, an iterative subject selection algorithm is proposed to select most subject-related instances instead of using the whole training set. In the inference stage, a joint decision is made with confident reconstruction prior to composite information from facial expressions and body gestures. We also collect a new RGB-D dataset for affection recognition with large subjective variance. Experimental results demonstrate that the proposed affection recognition framework can increase the discriminative power especially for facial expressions. Joint recognition strategy is also demonstrated that it can utilize complementary information in both models so that to reach better recognition rate.

1. Introduction

Human activity analysis is a significant component in image and video understanding and the large visual variance as well as semantic ambiguity underlying this topic makes it a difficult task. Applying advanced feature engineering and machine learning models, researchers in computer vision can build automatic software systems to recognize activity categories in controlled environments, such as smart-home surveillance and video gaming interactions. However, there are still numerous difficult problems in this domain, especially for subtle actions or emotions, such as affection recognition.

Affection is a disposition of mind or body, which is often expressed by facial expressions and body gestures. Some affection categories can be conveyed solely from facial expressions or solely from body gestures. But it is more natural and common that facial expressions and body gestures jointly express an affection.

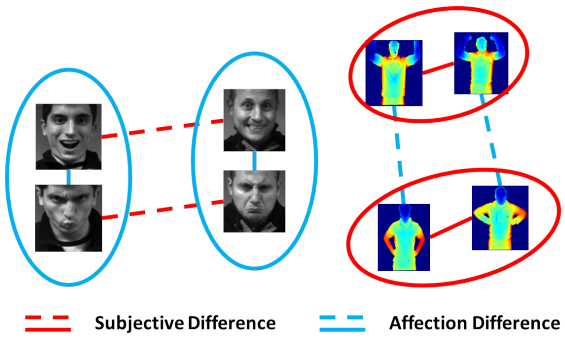

Figure 1. For facial expression, subjective (intra-class) variance is much larger than inter-class variances (expression). It brings benefits for expressional invariant face recognition but difficulties to subjective invariant expression recognition. This phenomenon is less severe for body gestures in Depth channel.

The success in facial expression recognition provides a plentiful of approaches to solve the problem in one perspective. Action Units (AUs) for Facial Action Coding System (FACS) [3] is a good modeling for facial expressions by decomposing the facial expressions into smaller organ-based movements, such as drawing brows and opening mouth. Facial expression recognition based on AUs is successful and has attracted a lot of attentions [14] [15] [20]. In addition, recognizing human emotions from body gestures is also a growing research area in recent years [12]. Especially after the debut of Kinect depth camera [7], the new type of sensor together with its technologies provides powerful tools for human activity analysis. The depth channel makes it easier to segment human from clutter background and therefore research based on this novel information channel has been conducted on an unprecedented scale [16] [19] [13].

However, the difficulty in facial expression recognition is always proportional to the degree of subjective variance. As illustrated in Figure 1, subjective variance in image space is

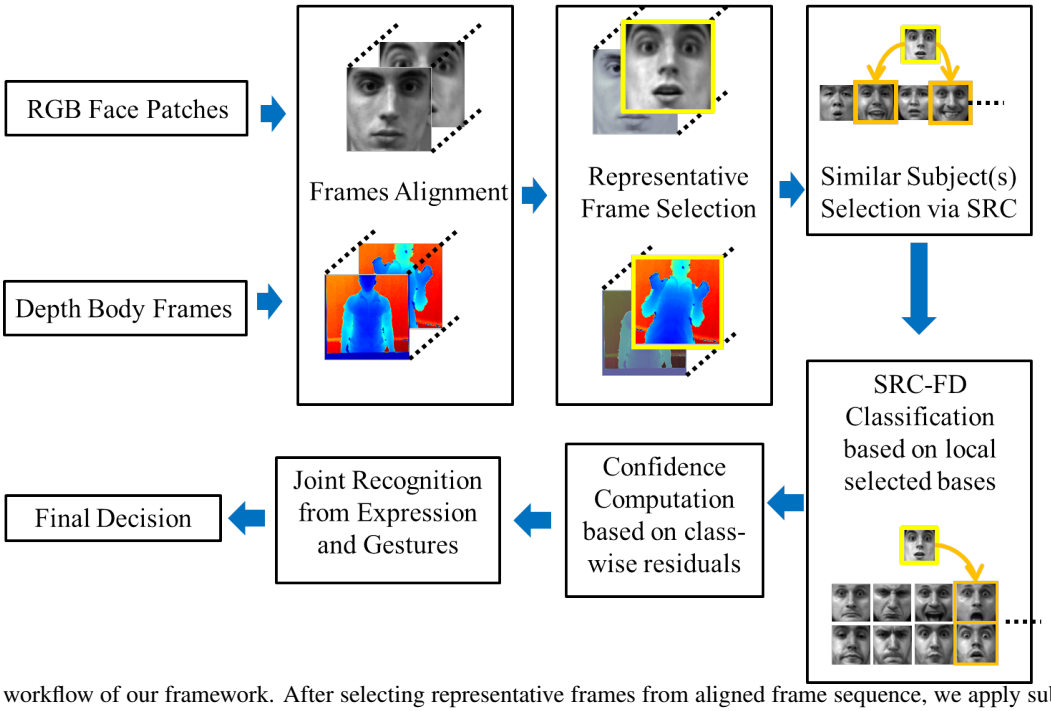

Figure 2. The workflow of our framework. After selecting representative frames from aligned frame sequence, we apply subject selection for a given testing subject. Affection class is recognized for testing queries with selected training data using sparse representation based classification. Then a joint decision from expression model and body gesture model is made based on confidences calculated from class-wise reconstruction residuals.

much larger than expressional variances. In the domain of Face Recognition (FR), the subjective variance is inter-class variance and expressional variance is intra-class class variance. However in the domain of Expression Recognition (ER), the roles of the two kinds of variances are reversed and therefore this phenomenon brings benefits to FR but harms to ER. Approaches are proposed to reduce the within class variance and increase the between-class variance, such as Linear Discriminant Analysis (LDA), or Fisher's Linear Discriminant [4]. Sparse Representation based Classification (SRC) [17] provides an informative way to image classification. In SRC, a query image is coded using a sparse dictionary whose bases (columns) are training samples with or without sparsity constraint; then the query image is reconstructed by the bases with sparse coefficients as well as sparse residuals. In [18], the authors combined SRC and Fisher Discriminant criteria to propose an algorithm to learn a structured dictionary and providing informative reconstruction residual for class recognition. A low-rank regularization constraint is added to FDDL is also demonstrated to be useful in FR [6].

In this paper, we utilized the classification scheme proposed in [18], which uses the residuals from class-wise reconstructions as classification criteria. We argue that instead of using all training samples for sparse reconstruction with the huge subjective variance, it makes more senses to select a subset of subjects using FR first and recognize affection then. Then we propose a joint affection recognition combining facial expressions (from RGB channels) and body gestures (from the Depth channel) with subjective adaption and joint decision making based on reconstruction confidence in sparse representation. The contributions of our work have two aspects:

1. First, we propose a subject adaptive sparse representation approach by combining the idea from [17] and [18] and reconstruct the query image from subject related subgroups.

2. Second, we address the joint recognition problem using the confidence computed from the residuals of sparse representation and experiment results demonstrate that the combination can be effective without additional computational cost.

Additionally, we also provide a combinatory dataset for joint affection recognition with both facial expressions and body gestures. Both color images and depth images are collected for multi-modal recognition.

An overview of the proposed framework is illustrated in Figure 2. Face patches and body gesture patches are extracted from RGB channels and depth channel of the input video respectively. We firstly apply Robust Alignment (RASL) [10] to align the frames in each video sequence. Then representative frames (queries) are selected based on

"apex" position, where expressional intensity is highest as discussed in [1]. Subjective adaption is to select a group of most similar subjects based on SRC based face recognition. The query image is then reconstructed from the selected subjective dictionary. Fisher Discriminative Sparse Representation Classification (SRC-FD) is used for class inference. The final decision of affection class is made according to the confidence score based on class wise reconstruction residuals.

The organization of the rest of this paper is as following. Section 2 introduce how we align frames and select the most representative ones from them. Proposed subject adaptive affection recognition and joint decision making framework is introduced in Section 3. We also describe our new collected affection recognition data set in Section 4 and proposed framework on this data set is evaluated and discussed in Section 5. We conclude in Section 6.

2. Pre-processing

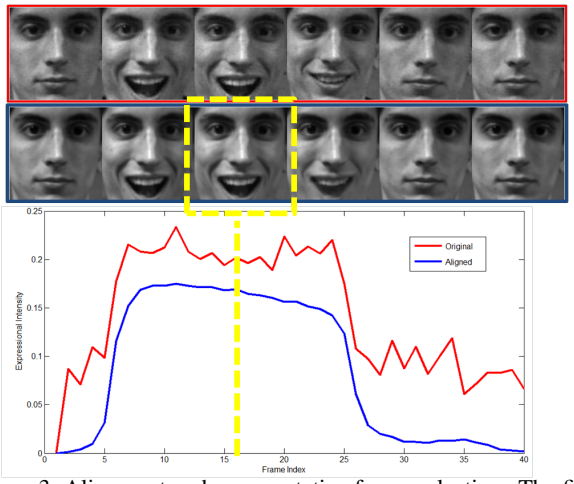

Figure 3. Alignment and representative frame selection. The first row with red boundary shows the original un-aligned face patches, one can observe variances due to poses and minor changes. The second row with blue boundary shows the aligned face patches via RASL [10], one can observe that differences due to factors other than expressions are eliminated. Expressional intensity curves are shown in the bottom. Selected representative frame is indicated by yellow dashed box.

Given a sequence of face patches, we firstly need to align them and select the most representative frame out of them. The misalignment in the sequence is introduced by both human movement and noise in face detection. To align them, we apply the RASL [10] algorithm which uses sparse and low-rank matrix decomposition. Sparse learning based frame alignment takes advantages of the inner structure of the given sequence of similar frames (*e.g.* face patches of the same subject) and reduces the noises with rare occurrence. For representative frame selection, we select the middle frame of the apex area [1] according to expressional intensity.

As illustrated in Figure 3, RASL [10] smooths the expressional intensity curve by representing the "intermediate" frame with "apex" or "neutral" frame. The red curve indicates the intensities of un-aligned sequence and blue for aligned sequence. Yellow dashed box shows the final selected representative frame. For body gestures, the pre processing step is the same as facial expression.

3. Subject Adaptive Joint Affection Recognition via Fisher Discriminant Sparse Representation

In this section, we firstly review sparse representation based classification (SRC) [17] and Fisher Discriminant. Then our proposed two layered subject adaption framework for affection recognition is described. Finally, a joint recognition framework is proposed based on the class-wise reconstruction residuals.

3.1. Sparse Representation Classification with Fisher Discriminant

Sparse representation based classification (SRC) was proposed in [17] by Wright *et al.*. Given C as the set of class labels, we have $A = [A_{C_1}, A_{C_2}, ..., A_{C_c}]$ as the dictionary of training samples. In our approach, A is the matrix of vectored frames, *i.e.*, $A_{C_i \in C} = [vec(x_1^{C_i}), vec(x_2^{C_i}), ..., vec(x_n^{C_i})]$, where $x_j^{C_i}$ is the j^{th} image (face path or gesture patch) of class i. Given a query image q and its vectored instance $y = vec(q)$, the SRC via l_1-minimization is given as:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \quad s.t. \quad \|y - A\alpha\|_2 \leq \lambda \tag{1}$$

Therefore, Classification rule is given as:

$$identity(y) = \underset{C_i}{\operatorname{argmin}} r_{C_i}(y) \tag{2}$$

where class-wise reconstruction residual $r_{C_i}(y)$ is given as:

$$r_{C_i}(y) = \|y - A\delta_{C_i}(\hat{\alpha})\|_2 \tag{3}$$

where δ_{C_i} is the characteristic function that selects the coefficients associated with that class.

According to FDDL [18], the SRC classification rule can be re-written as an "SRC-FD" form:

$$\hat{\alpha_{C_i}} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \quad s.t. \quad \|y - A_{C_i}\alpha\|_2 \leq \lambda \tag{4}$$

where A_{C_i} is the sub-dictionary associated with class C_i. Thus the class-wise residuals in Eq. 3 is re-written as:

$$r_{C_i}(y) = \|y - A_{C_i}(\hat{\alpha_{C_i}})\|_2 \tag{5}$$

Noted that Eq. 5 corresponds to FDDL with global cost weight as 0 (the global reconstruction residual weight is used in FDDL [18] to force that the reconstruction should not come from all the data points, for more details, please refer to [18]). We called this classification method as Fisher Discriminant SRC (SRC-FD). Comparing Eq. 3 and Eq. 5,

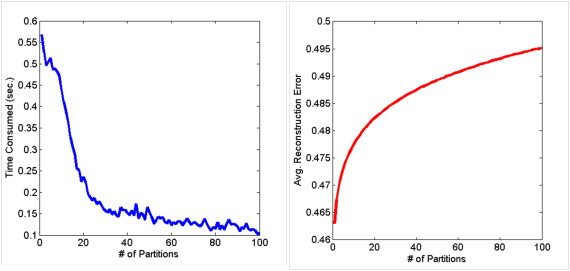

Figure 4. Timing and avg. reconstruction errors on Simulated data with 10,000 training samples and 1 query instance. We can observe that time consumed decreased as the number of partitions increases.

the latter one is more intuitive. The classification is actually to find the optimal space spanned by bases of a certain class which minimize the reconstruction error. In addition, the residual formula in SRC-FD makes the inference much efficient, which is critical in applying SRC to real applications. To illustrate this argument, we test the l_1-minimization using Least Angle Regression [2] on a simulated dataset which has 10,000 instances and each instance is of dimension 1000; a query instance is given to reconstruct. The simulated data is randomly generated and each instance is normalize to unit l_2 norm. We partition the training set into n non-overlapping subsets and the query is reconstructed on each subset. The overall computation time and averaged reconstruction error for different n are as illustrated in Figure 4. With the number of partition increases, the computation time decreases. This is because l_1 minimization is not linearly proportional to the number of training instances. However, using a larger dictionary can achieve better reconstruction error. In our work, since the absolute reconstruction error is not a concern *per se*, so we use the classification form as in Eq. 5.

3.2. Subject Adaptive Affection Recognition

Both SRC [17] and FDDL [18] achieve impressive recognition rates in face recognition (FR) and have been demonstrated to be robust to varying illuminations, occlusion and expressions. However, subjective robust expression recognition (ER) is harder than expressional robust face recognition. The reasons are two-folded: 1) in term of pixel-wise variance, the distance between subjects of the same expression is much larger than the distance between expressions of the same subjects, as shown in Figure 1; 2) behavior habits of different subjects make the subjective variance much larger and the expressions harder to model. The two reasons jointly make ER a harder problem than FR.

The similar phenomenon also exists in gesture recognition. Instead of facial appearance, subjective appearance variances in body gestures are due to subjective body sizes, types and clothes. However, because 1) the body gestures are always much more drastic than facial expressions and 2) we extract body gesture information from Depth channel which ignore much of the appearance variances (*e.g.*, different clothes), this phenomenon is actually not crucial at all. Therefore, the effect of applying the subject adaptive framework on gesture recognition is limited to overcome subjective behavioral variance.

In this paper, we propose a two layer recognition structure to pursue subjective robust affection recognition. The first layer is actually a face recognition problem. The motivation of the first layer recognition is straightforward: given a query image of an affection, it is more naturally to identify who is the subject and check if previous records exist in our training data, if so, using the data of the same subject is more efficient and accurate. However, in reality most of the time the assumption of already seeing the same subject is invalid. Therefore, we define the first layer recognition step as to find a fixed number of most similar subjects such that the identity information (appearance, behavioral habits, *etc.*) can be approximated using the instances from the selected subjects.

The first layer recognition is as illustrated in Algorithm 1. To reach better consensus subjects selection, we use a batch of queries (from the same subject) each time, *i.e.*, $|Y| > 1$. An upper bound limit on maximal available subjects (N) is also given. The propose algorithm selects a subgroup (A^*) of training instances (A) for expression recognition using SRC-FD with l_1-minimization. One may doubt that it is not valid to know that the given batch of frames are from the same subject. We concede that sometimes it is the case but in reality, an affection recognition system can always capture more than one frames from the same subject and select similar subjects in training set as an off-line initialization. In the other hand, algorithm 1 is still functional when $|Y| = 1$.

Based on the selection of A^* (which can be represented as $A^* = [A_1^*, A_2^*, ..., A_c^*]$, where A_i^* is the class-wise subset of A^*), we can determine the affection class of each query y according to Eq. 4 and 5.

Figure 5 illustrates the proposed 2-layer recognition framework. When a query image is given, the first layer recognition process seeks at most N ($N = 2$ in this example) subjects whose samples can best approximate the query image. After we have selected the subset of subjects (colored as green rows in the training data matrix), the selected subset is further partitioned into C classes, $i.e.$, subsets with

Algorithm 1: Subject Adaptive sub-dictionary selection.

Input: training instance matrix A, testing instance matrix Y, subject number limit N, subject set S

Output: subject adaptive sub-dictionary A^*

A^* = *empty*;
for $y \in Y$ **do**
  for $S_i \in S$ **do**
    $\hat{\alpha} = \text{argmin}_\alpha \|\alpha\|_1 \quad s.t. \|y - A_{S_i}\alpha\|_2 \leq \lambda$;
    $r_{S_i} = \|y - A_{S_i}\alpha\|_2$;
  end
  $\hat{S} = \text{argmin}_{S_i} r_{S_i}$;
  vote($\hat{S}$) += 1;
end
$n = 1$;
while $n \leq N$ *and* $!allzeros(vote)$ **do**
  $s = \text{argmax}_{S_i}$ vote(S_i);
  $A^* = [A^*, A_s]$;
  vote(s) = 0;
end
return A^*

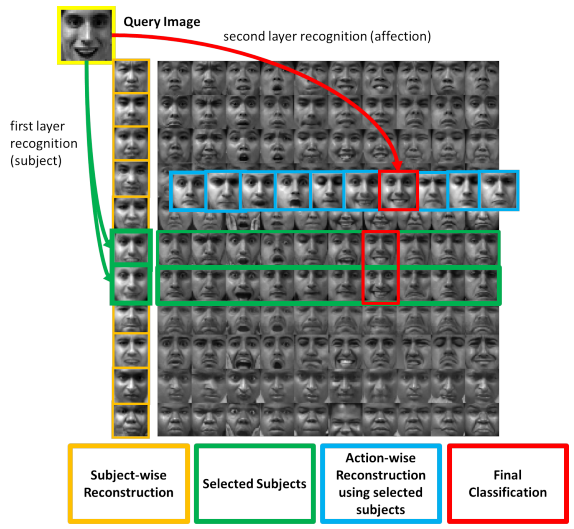

Figure 5. Illustration of our two-layer affection recognition framework. The first layer is to select a subset of subjects which can best approximate the query image(s). The selected rows (colored as green) are used for second-layer affection recognition using SRC. The action-wise approximations are colored as blue. The final decision is to find which class can best approximate the query image. The final decision are shown in red.

affection labels. The second layer recognition seeks the "row" which can best approximate the query image (final decision and selected subset are colored as red boxes).

Algorithm 2: Joint Decision Making from Face and Gesture.

Input: Facial expression dictionary $A^{*,f}$ and body gesture dictionary $A^{*,g}$, a query $y = [y_f, y_g]$

Output: Affection label $\hat{C}$ of y

for $C_i \in C$ **do**
  $\hat{\alpha^f} = \text{argmin}_\alpha \|\alpha\|_1 \quad s.t. \|y - A^{*,f}_{C_i}\alpha\|_2 \leq \lambda$;
  $\hat{\alpha^g} = \text{argmin}_\alpha \|\alpha\|_1 \quad s.t. \|y - A^{*,g}_{C_i}\alpha\|_2 \leq \lambda$;
  $r^f_{C_i} = \|y - A^{*,f}_{C_i}\hat{\alpha^f}\|_2$;
  $r^g_{C_i} = \|y - A^{*,g}_{C_i}\hat{\alpha^g}\|_2$;
end
$conf^f = F(r^f)/(F(r^f) + F(r^g))$;
$conf^g = F(r^g)/(F(r^f) + F(r^g))$;
$\hat{C} = \text{argmin}_{C_i} r^f_{C_i} * conf^f + r^g_{C_i} * conf^g$;
return $\hat{C}$

3.3. Joint Decision Making via Confident Reconstruction Prior

When there are multiple models have the same set of class labels, as in this paper, facial expression model and body gesture model, how to effectively combine them to make a joint decision is an issue. One can combine the models in an early phase by feature concatenation or make a joint decision only based on the decision scores given by different models. In this paper, we apply the latter one since the early fusion in feature representation level can be overwhelmed by the dominant feature channel, if there exists one.

In our work, we have two models: facial expression model and body gesture model. Each model uses the same classification rule based on Sparse Representation (SRC). Since the decision in each model is made according to the smallest residual in term of l_2 norm. It is straightforward to derive the confidence score of an decision, denoted as $F(.)$ by the margin between the smallest residual and the second smallest residual. The assumption under the confidence score formula is that a "confident" decision should be made more easier with a more comparative significant smallest reconstruction error. Then the confidence scores of both models are used for weighing the reconstruction error and we make the joint decision by selecting the smallest weighted sum of the class-wise reconstruction residuals. The procedure is as shown in Algorithm 2.

4. Face and Gesture RGBD Dataset

In this section, we introduce a new Face and Gesture RGBD dataset (FG-RGBD) we collected for affection recognition with 1920 affection samples.

In [5], the authors presented a widely used bi-modal

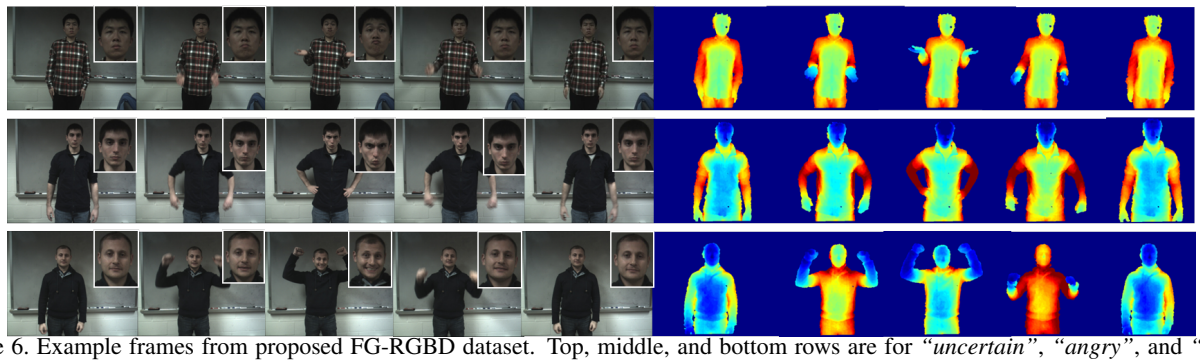

Figure 6. Example frames from proposed FG-RGBD dataset. Top, middle, and bottom rows are for *"uncertain"*, *"angry"*, and *"happy (cheering)"* respectively. RGB frames and enlarged face patches are shown on the left and depth frames are shown on the right.

public dataset for combinatory recognize affective behavior categories from both facial and gesture model. However, as recent success in Kinect and related research in the depth channel, there is a trend that researchers are mining more complementary information from this novel information source instead. There are a lot of datasets have been presented for using depth channel as a counterpart for research topics and have proved their effectiveness, such as MSR Gesture 3D, MSR Action 3D, and MSR Daily Activity 3D [8]. However, to the best of our knowledge, there is not such a dataset for affection recognition jointly from face and gesture combining both RGB channel and Depth channel. To fulfill this vacant slot, we thus present a Face and Gesture RGBD dataset for affection recognition (FG-RGBD) dataset which contains videos from both RGB channels and depth channel from a Kinect camera. Basic statistics will be introduced briefly in this section.

There are ten affection categories in our FG-RGBD dataset, they are: *"uncertain"*, *"angry"*, *"surprise"*, *"fear"*, *"anxiety"*, *"happy (cheering)"*, *"happy (clapping)"*, *"disgust"*, *"boredom"* and *"sad"*. There are twelve subjects are recruited to perform the ten categories of affections according to a simple instruction. The subjects were asked to perform each affection in 4 different records (video clips), each record (video clip) the subjects were asked to repeat 4 times. The dataset contains a significant subject-variance because of two reasons: 1) the instruction used to direct the subjects has no more than two sentences for each action, so the subjects have a big freedom to perform the actions spontaneously, which is more close to reality. 2) The subjects are from different races and genders: there are 1 American-African, 2 Latinos, 4 Caucasians and 5 Asians; there are 2 women and 10 men.

In our FG-RGBD dataset, both RGB frames and depth frames are provided, skeleton joint estimations computed from off-the-shelf software [11] are also provided yet not used in this work. There are in total of 480 videos as well as 1920 affection samples collected.

In this work, the 1920 samples are divided into training set with 960 samples and testing set with 960 samples while none of the subjects appears in both training and testing sets. Resolutions for RGB frames and depth frames are 1280×1024 and 640×480, respectively. Some sample frames from the FG-RGBD dataset are shown in Figure 6.

5. Experimental Results

In this section, we use FG-RGBD dataset to evaluate tasks for facial expression recognition and body gesture recognition and joint affection recognition. Quantitative results in term of recognition rates are reported and compared with several baselines and state-of-the-art methods. Qualitative results in terms of cross-subject facial expression and gesture reconstruction are also illustrated for future discussion.

5.1. Selection on subject limit N

In this part, we discuss the effect of subject limit N selection in subject adaptive phase. As can be inferred from Figure 4, if we select a small N, the reconstruction error should be large but the time consumed in testing phase is reduced. Although reconstruction error is not a concern *per se* in this work, a over-relaxed reconstruction error can bring bad recognition accuracy. Therefore, we need to find a good tradeoff when selecting N.

As illustrated in Figure 7, with the increasing N, recognition rate increases. But after $N = 8$, it becomes stable. Two examples of "happy" are shown for illustrating the progress of reconstruction. With the increase of subjects available, the reconstruction error can be reduced from a "hybrid" reconstructed face.

5.2. Affection Recognition Performance Evaluation

In this part, we evaluation our affection recognition framework and compare our system with state-of-the-art methods [17] [18] in term of recognition rate. In our experiments, we conduct "leave-one-out" cross-subject tests for all methods and report the averaged recognition rates. We use the face detector in [9] to localize face patches in

Approach	Recog. Rate for expression	Recog. Rate for gestures	Joint Recog. Rate
Logistic Regression	38.49%	46.35%	54.89%
FDDL [18]	43.39%	61.25%	64.84%
SRC[17]-FD	46.72%	62.34%	69.3%
Proposed Method	**48.80%**	**62.66%**	**69.7%**

Table 1. Performance comparison of different methods.

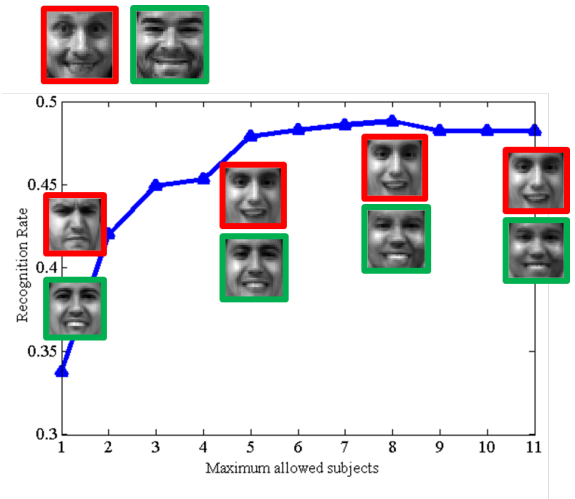

Figure 7. Recognition increases with N grows but after some certain value, it does not change. In this curve, the recognition rate reaches maximum as 48.8% when $N = 8$. An example of original query face y and reconstructed face using selected subjects ($A^*\hat{\alpha_y}$) are shown. We can observe that with the increase of N, the reconstructed face is de-personalized. Two samples of original query faces and reconstruction sequences with varying N are shown in red and green boxes ($N = 1, 5, 8, 11$, respectively). It can be observed that the reconstructed faces "evolve" to be more similar to the original queries.

RGB channels. The extracted face patches are normalized to 150×150 resolution and aligned by RASL [10]. We extract representative frame in the low-rank part of the aligned face sequence (for details, please refer to [10]). The selected face patch is down-sampled to 30×30 and vectored as feature vector. The subject limit we select for facial expression channel is 8, but in average there are only 5 to 6 subjects are selected. As for body gesture model, we apply the same procedure as for face patches using [10], then the body gesture patch is normalized to 38×38 and vectored to be the feature vector for body gesture. For single model evaluation, the subject limit is also set to 8. In all our l_1 minimization process, we force the reconstruct coefficients $\alpha \leq 0$.

Table 1 shows the comparative results among several state-of-the-art methods with proposed framework. We also compare with a baseline method, logistic regression, since logistic regression can explicitly output classification probabilities of each class label. In Joint recognition, we apply the classification probability of facial expression model and body gesture model to the joint decision. As for FDDL, we directly use the published code for evaluation. Joint recognition with FDDL is accomplished by early fusion of facial expression frame and body gesture frame.

From Table 1, we observe that proposed method outperforms other methods, especially in facial expression recognition part. We also observe that in joint recognition, if we relax the subject limit constraint in body gesture channel, the joint recognition result is better. We report our best performance in Table 1.

Figure 8 illustrates the confusion matrices for affection recognition from facial expression, body gesture and joint decision making. We observe that the gesture recognition model solely perform superior than facial expression recognition model, especially between classes *"happy (cheering)"* and *"happy (clapping)"* since the facial expressional attributes are very similar while gestures vary drastically. Body gesture model performs much better in classes *"surprised"*, *"happy (cheering)"*, and *"sad"* since their gestures are much more distinct with others. However, this model is a little ambiguous in distinguishing between classes *"anxiety* and *"happy (clapping)"* because both gestures have similar attributes; similarly, classes *"happy (disgust)"* and *"boredom"*, because both gestures contain action attribute like "raising hands in front of chest". Although facial expressional recognition rate is lower than body gesture recognition rate in almost every class, we can observe that the information contained in each model is quite complementary to each other: thus jointly recognizing affection classes reaches much higher recognition rates, such as in class "anxiety", "happy (clapping)" and "disgust".

6. Conclusion

In this paper, we have investigated on affection recognition from the perspective of facial expression and body gesture combination in RGB-D videos. To address the issue that subjective variance in affection recognition is always larger than inter-class variance, we have proposed a novel subject adaptive algorithm to mining category-related variance by using sparse representation with Fisher discriminant. Instead of using all training data for each test-

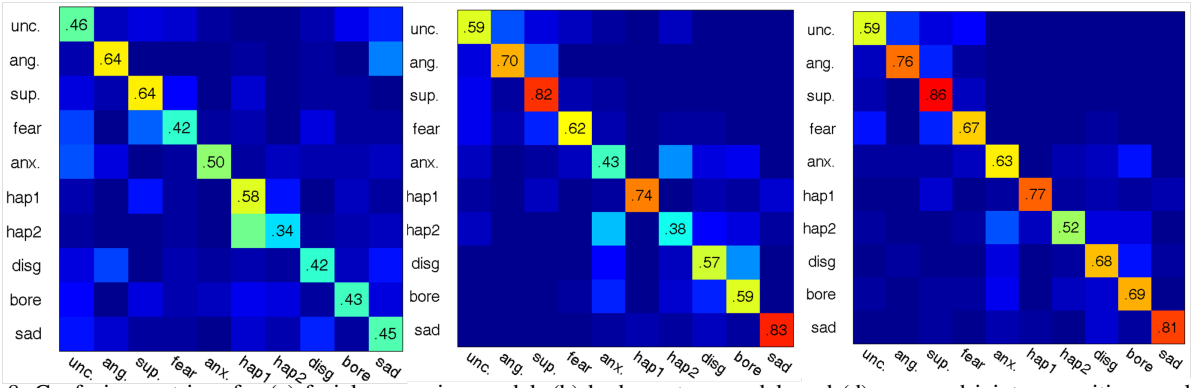

Figure 8. Confusion matrices for (a) facial expression model, (b) body gesture model, and (d) proposed joint recognition result. Obvious improvement is achieved in joint recognition infers that facial expressions and body gesture model contains very complementary discriminative information.

ing query, we firstly select a subject adaptive subset using sparse representation based classification. Then affection class is recognized in the selected subject adaptive subset of training data. To jointly recognize affection class from facial expressions and body gestures, we propose a confident reconstruction based joint decision making strategy. We also presented a novel dataset which contains 10 different affection categories and 12 subjects, which is challenging due to large subjective variance. Our proposed recognition framework and joint recognition approach is evaluated on the dataset. Experimental results demonstrate that joint recognition results can be improved by combing two complementary discriminative models.

Acknowledgment

This work was supported in part by NSF grant EF1137172, IIP-1343402, and FHWA grant DTFH61-12-H-00002.

References

[1] S. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. Conn. Improved facial expresion recognition via uni-hyperplane classification. In *CVPR*. IEEE, 2012. 3

[2] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), 2004. 4

[3] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978. 1

[4] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936. 2

[5] G. H and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *ICPR*, 2006. 5

[6] L. Li, S. Li, and Y. Fu. Discriminative dictionary learning with low-rank regularization for face recognition. In *FGR*, 2013. 2

[7] Microsoft Corporation. Kinect for Xbox 360. http://www.xbox.com/en-US/kinect, 2010. 1

[8] Microsoft Research. MSR Action Recognition Datasets. http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/default.htm. 6

[9] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, 2008. 6

[10] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. 2, 3, 7

[11] PrimseSense. OpenNI. http://www.openni.org/. 6

[12] C. Shan, S. Gong, and P. McOwan. Beyond facial expressions: learning human emotion from body gestures. In *BMVC*, 2007. 1

[13] J. Shotton, A. Fitzgibbon, , M. Cook, T. Sharp, M. Finocchio, M. moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011. 1

[14] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on PAMI*, 23, 2001. 1

[15] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPR workshop*, 2006. 1

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012. 1

[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, , and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 31(2):1–18, 2009. 2, 3, 4, 6, 7

[18] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*. IEEE, 2011. 2, 3, 4, 6, 7

[19] C. Zhang and Y. Tian. Edge enhanced depth motion maps for dynamic hand gesture recognition. In *CVPR Workshop*, 2013. 1

[20] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huamg, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012. 1