Multiple Kernel Learning for Emotion Recognition in the Wild

Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana,
Gwen Littlewort, Marian Bartlett
Machine Perception Laboratory
UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093
{ksikka, kdykstra, ssathyanarayana}@ucsd.edu

ABSTRACT

We propose a method to automatically detect emotions in unconstrained settings as part of the 2013 Emotion Recognition in the Wild Challenge [16], organized in conjunction with the ACM International Conference on Multimodal Interaction (ICMI 2013). Our method combines multiple visual descriptors with paralinguistic audio features for multimodal classification of video clips. Extracted features are combined using Multiple Kernel Learning and the clips are classified using an SVM into one of the seven emotion categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. The proposed method achieves competitive results, with an accuracy gain of approximately 10% above the challenge baseline.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; H.5.1 [**HCI**]: Multimedia Information Systems

General Terms

Machine Learning, Emotion Recognition

Keywords

Support Vector Machine, Multiple Kernel Learning, Bag of Words, Multimodal, Feature Fusion

1. INTRODUCTION

In this paper, we propose a method to automatically detect emotions in the wild, as part of the Emotion Recognition in the Wild Challenge, organized in conjunction with the ACM International Conference on Multimodal Interaction (ICMI 2013) [16]. The challenge dataset called Acted Facial Expression in the Wild (AFEW) [17] consists of short audio-video clips extracted from a set of Hollywood movies. Separate sets of videos were provided for training and validation and labeled as one of the seven emotion categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise. The task is to classify a sample audio-video clip into one of these seven categories. Our method uses Multiple Kernel Learning (MKL) to find an optimal combination of audio and visual features for input into a non-linear support vector machine (SVM) classifer. The visual features used were dense multi-scale SIFT BoW [40], LPQ-TOP [31], HOG [13], PHOG [6], and gist [30] features.

The paper is organised as follows. Section 2 gives a brief overview of existing related work, followed by Section 3 that introduces the dataset that was used for this challenge. Next, Section 4 describes the proposed method in detail, including descriptions of the various features that were explored in the audio and video modalities. This is followed by Section 5 that summarizes the results, after which a discussion is presented in Section 6. The paper is concluded in Section 7.

2. RELATED WORK

Automatic detection of human emotions from a scene finds numerous applications such as in human-computer interfaces, intent analysis and image retrieval. Most existing methods rely either on visual data or audio data for emotion recognition, although there is relatively little work on recognition of human emotions using audio data as compared to video data [38].

In one of the early works based on audio features [2], Chiu et al. deployed a multilayered neural network for automatic classification of emotions using five features that were extracted from speech. In [20], Chen et al. estimated pitch, intensity, and pitch contours as acoustic features, which were then classified into the following basic emotion categories using a rulebased approach: happy, sad, fear, anger, surprise and dislike. Scherer et al. [36] extracted a more exhaustive set of 29 audio features from speech and concluded, "Sadness and anger are best recognized by audio data, followed by fear and joy. Disgust is the worst."

Most vision-based emotion recognition studies focus on facial expression analysis, given the importance of the face in emotion expression and perception [51]. Ekman et al. [19] developed the Facial Action Coding System (FACS) to objectively measure facial activity for behavioral science investigations of the face. The FACS defined 46 Action Units, or AUs, corresponding to each independent motion of the face. A trained human FACS coder decomposes an observed ex-

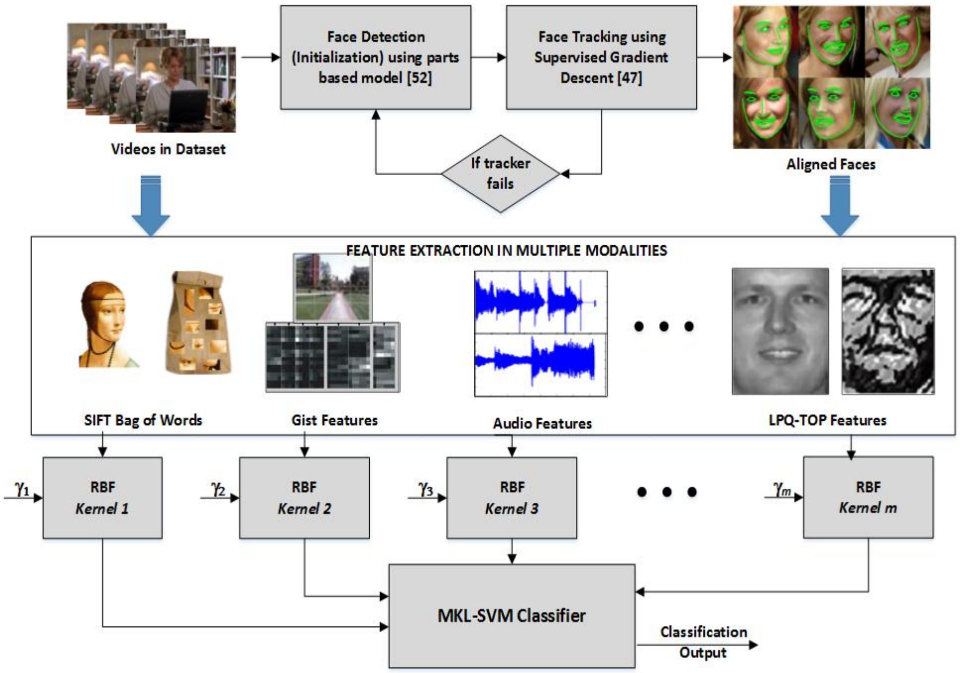


Figure 1: Classification pipeline of the proposed method. Once visual and audio features are extracted we construct a radial basis function (RBF) kernel from each descriptor. We then use MKL to optimally combine the feature kernels for input into an SVM classifier.

pression into the specific AUs that produced the expression [38]. [18] presented a comparative analysis of several techniques for automatic recognition of facial expressions using FACS, involving optical flow, principal component analysis, independent component analysis, local feature analysis, linear discriminant analysis, Gabor wavelet representations and local principal components. Lucey et al. [25] used Active Appearance Models for extracting facial features. Nearest Neighbor (NN) classifiers and Support Vector Machines (SVM) [41] were used for the classification of FACS units [19].

Yacoob et al. [50] proposed a method that involved tracking of facial parts and applying optical flow in high gradient regions to identify directions of rigid and non-rigid motions caused by human facial expressions. A high level facial expression classification was obtained using a mid-level representation of the flow direction. Another approach based on optical flow proposed by Black et al. [4] accounted for changes in image appearance by generalizing optical flow to provide a richer description of image events. Different parametric models were used to extract parameters from facial features and a nearest neighbor classifier was used for expression recognition.

In order to tackle the problem of expression recognition in profile face image sequences, Pantic et al. [32] proposed a method that segmented profile faces using connected component analysis in HSV space. A contour based method was further applied to extract 20 points, that were used for AU detection.

Further improvements to this work can be found in [33] and [43]. In [33], Motion History Images were used and temporal rules were applied to identify AUs, while in [43], wavelet based gentleboost template were used for tracking 20 facial landmark points, which were then used for constructing spatio-temporal features. Dhall et al. [15] proposed a method for automatic emotion recognition based on pyramid of histogram of gradients (PHOG) [7] and local phase quantisation (LPQ) [29] features. For classification, they used SVM [41] and largest margin nearest neighbour (LMNN) [5].

Psychological studies such as [3], [35], [38], have highlighted the importance of using multiple modalities to strengthen the accuracy of the emotion analysis. In [9], Carlos et al. analyzed the strengths and weaknesses of vision-only and audio-only based expression analysis systems. They also outlined approaches for fusing the two modalities, and it was shown that when these two modalities were fused, the performance and the robustness of the emotion recognition system improved measurably. Tawari et al. [42] presented a multimodal facial expression recognition framework using audio-visual information, showing that not only accuracy was improved by the integration of audio and video features, but there was also a reduction in computational cost.

3. DATASET

The Acted Facial Expression in the Wild (AFEW) dataset [17] consists of short video clips extracted from popular Hollywood movies. While automatic facial expression recognition has been an active area of research for decades, previous work has focused on datasets collected in uniform conditions with posed facial expressions removed from emotional context [24] [34] [26]. The goal of the AFEW dataset is to address the challenges in recognizing emotions in near real-world conditions. Each clip contains an actor expressing one of seven emotions: neutral, happy, sad, disgust, fear, anger, or surprise. The goal of the challenge is to correctly distinguish between the seven emotions. AFEW contains training, validation, and testing datasets, respectively consisting of 380, 396, and 312 video clips. For the final submission, use of the validation set is limited to setting model hyperparameters. Since the test set labels are currently unavailable in this paper we only report results on the validation set.

There were a number of challenges encountered while working with this dataset. Firstly, the range of poses in the dataset was quite vast. As a result of this, readily available face detectors such as the OpenCV Viola Jones [46] failed to give the required face initialization. This prompted us to explore advanced face detection methods, such as the one used in the proposed method. Secondly, there was significant variation in the way the same emotion was expressed by different subjects in different videos. In many cases, it was difficult for even human viewers to discern the emotion displayed in the video. Thirdly, most video samples consisted of multiple human subjects, making it challenging to isolate the primary candidate of interest. Lastly, the number of training samples was low, given the complex nature of the dataset. This in turn imposed a challenge for the prediction task.

4. METHOD

We created a multimodal classification system by combining audio and visual features using Multiple Kernel Learning (MKL). For visual features, we experimented with dense multi-scale SIFT BoW [40], LPQ-TOP [31], HOG [13], PHOG [6], and ”gist” [30] features. For audio features, we used a set of paralinguistic features provided by the challenge organizers [16]. We built a Radial Basis Function (RBF) kernel from each set of descriptors and use MKL to optimally combine them for input to a support vector machine (SVM). A visual overview is given in Figure 1. The following sub-sections describe each component of our process in more detail.

4.1 Face Extraction and Alignment

For extracting faces from the video frames we combined a state-of-the-art face detection method [54] with a recently proposed tracking method [49]. Recent work in face detection has focused on part-based deformable models including Active Appearance Models (AAMs) [27], and their extension Constrained Local Models (CLMs) which build global models on top of local part detectors [12]. For face detection we used Zhu and Ramanan’s deformable parts model (DPM) which achieved competitive results by fitting a mixture of trees model and then applying a shape model similar to AAMs and CLMs [54]. This model is able to handle non-frontal head pose which is especially important for successful face detection in the AFEW dataset. Following initialization, the facial landmarks were then tracked using the supervised descent method (SDM) [49]. SDM provided an efficient framework for fitting AAMs that outperformed discriminative methods through supervised linear estimation of the descent direction. The DPM model was used to re-initialize the facial landmarks whenever the tracker failed.

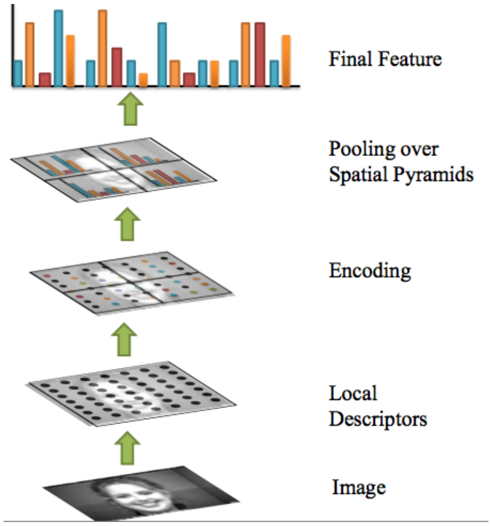


Figure 2: Pipeline for BoW extraction

4.2 BoW on Faces and Scene

We used a Bag of Words (BoW) model on top of multi-scale dense SIFT features (MDSF) [45], which has shown promise for application to automatic facial expression recognition [40]. The strength of this approach lies in the combination of dense feature sampling, implicit inclusion of spatial information through a multi-scale pooling step using spatial grids [22], and state-of-the-art feature encoding using Locality-constrained Linear Coding [47]. The approach is outlined in Figure 2.

First, we sampled multiscale dense SIFT features (MDSF) [23] [45] using a stride of 4 pixels. Four different scales were used by setting the SIFT spatial bins to 8, 12, 16, and 32 pixels. The codebook for BoW was generated using approximate K-means clustering, a clustering approach that employed data-to-cluster distances using the Approximate Nearest Neighbor algorithm. For clustering we used the code provided by the authors in [11] and a randomly selected subset of 1,000,000 features. Once the codebook of MDSF cluster centers was generated, each local MDSF was assigned to a codeword using Locality-constrained Linear Coding (LLC) [47]. LLC projected each descriptor to a local linear subspace spanned by a selection of the codewords using an optimization problem.

The traditional Bag of Words model is robust to spatial translation, but sacrifices spatial layout information during the histogramming process. Spatial Pyramid Matching (SPM) implicitly incorporates spatial information into the feature representation through histogramming within different subdivisions of the image [22]. For SPM each image was partitioned into 2^l x 2^l segments at multiple scales l = 1,2,4,8. The BoW representation was then computed within each of these segments, and all of the subsequent BoW histograms were concatenated into a single feature vector. Since each frame of the video produced a pyramid BoW feature vector, information from all frames of the video were combined using max spatial pooling, accomplished by taking the maximum of the pyramid BoW feature vectors over all frames [39].

Inspired by recent work in multiple dictionary classification [52] [1], we included multiple BoW kernels, each built using a different dictionary size. Each dictionary represents the structure of the data at a different resolution, and thus may contain complementary information. We experimented with dictionary sizes of 200, 400, and 600. Experiments by [40] showed that performance saturates at larger dictionary sizes.

In addition to extracting BoW descriptors from the aligned face images, we extracted BoW descriptors from the entire image. We reasoned that descriptors from the whole image may contain information about aspects important to emotion recognition, such as body posture, context, and scene information. BoW features for the entire image were computed using a dictionary size of 600.

4.3 LPQ-TOP

Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) are efficient representations of dynamic image texture [28], and successfully applied to facial expression recognition [53]. Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) is a variant of LBP-TOP which is additionally robust to image blur and uniform changes in illumination [31].

We used the code by the authors of [31] to extract LPQ-TOP features. First we calculated the short term Fourier transform (STFT) over neighborhoods of MxMxN pixels with neighborhood N_x centered at each pixel x, where M and N equal the sizes of the neighborhood in the spatial and temporal domains, respectively. We tested M and N at 5, 7, and 9 pixels. We then took the 13 lowest non-zero frequency points as these are more likely to contain blur-invariant phase information. We separated the imaginary and real components of the frequency points to get vectors of length 26 from each sampled pixel in the image x. This generated an excessive number of feature vectors, so PCA was used as a method to reduce the set to a smaller number of uncorrelated descriptors. We applied a decorrelation transform with spatial and temporal correlation coefficients $\rho_s = .2$ and $\rho_t = .2$, and projected the data using the largest $L = 8$ eigenvectors. The resulting descriptors were then quantized using a simple scalar binary quantization method. The quantized coefficients were mapped to integer value using binary coding. Finally, a histogram of dimensionality 2^L was generated from the resulting integer values at all pixel positions x. Separate histograms were generated for three orthogonal planes, similar to LBP-TOP, and then all three histograms were concatenated into one vector. The window settings of 5, 7, and 9 were respectively used for LPQTOP-5, LPQTOP-7, and LPQTOP-9.

4.4 HOG + PHOG

Histogram of oriented gradients (HOG) [13] features are commonly used in computer vision problems to describe shape information for object detection. HOG is based on the fact that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. HOG features have been used to describe the shape of human faces, in the context of static facial expression analysis. PHOG is a variant of HOG based on pyramids, which has been used for object detection [6] and facial expression analysis [15].

Descriptors	**Functionals**
PCM loudness	Position max/min
MFCC [0-14]	arith. mean, std. deviation
log Mel Freq. Band [0-7]	skewness, kurtosis
LSP Frequency [0-7]	lin. regression coeff 1/2
F0	lin. regression error Q/A
F0 Envelope	quartile 1/2/3
Voicing Prob.	quartile range 2-1/3-2/3-1
Jitter Local	percentile 1/99
Jitter consec. frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90

Table 1: Sound features: 38 low level descriptor along with their first regression coefficients, 21 functionals. Table reproduced from [37].

We apply both HOG and PHOG features for capturing the local shape information of the faces. Canny edge detector were first applied to the cropped face, after which the face was divided into spatial grids at each pyramid level. Next, a 3x3 Sobel filter was applied to the edge contours to calculate the orientation gradients, which were then combined at each pyramid level. We used an orientation range of [0-360] and set the number of pyramids to 4 and the number of bins to 9. To attain a fixed length representation of each video, we took a max-pooling of the HOG and PHOG features over all frames.

4.5 Gist

We hypothesized that apart from the humans in the scene, the higher level context of the scene may also contain discriminative information for the problem at hand. To capture the context, ”gist” features were used. Gist features were originally proposed by Aude Oliva et al. [30] as an approach to recognizing real world scenes without the need for either segmentation or recognition of individual objects or regions. The gist features were generated for different settings with orientations per scale set to 4, 8 and 16, and number of blocks set to 4 and 8, resulting in different lengths of gist descriptors. The descriptors from each frame in a video were then combined using a max-pooling operation.

4.6 Sound

In the previous sections we used multiple methods to represent the visual content of the dataset. In [14], experiments showed that humans alternatively rely on audio and visual information depending on which emotion was being expressed, suggesting the use of both audio and visual features for improving emotion recognition systems. This is supported by empirical findings that automatic systems which combine audio and visual features have higher accuracy of emotion identification than the equivalent monomodal systems [10] [48].

One successful approach to acoustic feature extraction for emotion recognition has been to extract timeseries of multiple paralinguistic descriptors and use pooling operations such as max or min on each timeseries to extract feature vectors. Such an approach was used as the baseline for the

2010 INTERSPEECH paralinguistic speech challenge and is described in detail in the paper [37]. A total of 1,582 speech features were generated for each video by taking 21 functionals of 38 low level descriptors and their first regression coefficients. Descriptors and functionals are detailed in Table 1. 16 zero-information features were dropped (e.g. min F0 is always zero) and an additional two features were added, F0 number of onsets and turn duration. These sound features were extracted and posted for use by the challenge authors [16].

4.7 Classification using MKL-SVM

Once we have extracted multiple feature representations of the data from both the audio and visual domains, we wish to combine the information contained in each of our descriptors in a way that increases the discriminatory power of our classifier.

Two common methods for fusing multiple feature representations are feature-level, where a single classifier is trained using all features as input, and decision-level, where a classifier is trained for each feature separately and a decision rule such as the sum rule or majority voting combines the classifier outputs. We followed a feature fusion approach based on Multiple Kernel Learning (MKL) [21]. Rather than using a single-feature kernel, MKL was used to find an optimal linear combination of base kernels which was used for training a SVM.

We tried the MKL implementation as given in [8] and [44]. Both of these implementations gave the same results, however we selected [8] since it poses the MKL as a convex-optimization problem promising an optimal solution. This work followed the one-vs-all multi-class classifier startegy, learning unique kernel weights for each class. This implemetation gave us better results as compared to one-vs-one multiclass classifier strategy as discusssed in Section 6.

We set N as the number of training examples and M as the number of features used for MKL. We define our feature sets X_m for $m \in \{1, ..., M\}$ and labels $y_i \in \{-1, 1\}$ for $i = 1, ..., N$. For each feature set we generated an RBF kernel $\mathrm{K}_m \in \mathbb{R}^{nxn}$ with the kernel function $\mathrm{k}(x_i, x_j) = \exp(-\|x_j - x_i\|/\gamma)$. To select the spread parameter γ for each kernel, we performed a grid search and selected the values which gave the best classification accuracy.

The dual formulation of the SVM optimization problem is then

$$\max_{\alpha,\beta} \left[\sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathrm{K}_{mkl}(x_i, x_j) \right] \quad (1)$$

$$\sum_{i}^{N} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

$$\mathrm{K}_{mkl} \succ 0$$

Where K_{mkl} is defined as the convex combination of all feature kernels

$$\mathrm{K}_{mkl} = \sum_{k}^{M} \beta_k \mathrm{K}_k \quad (2)$$

$$\sum_{k}^{M} \beta_k = 1$$

$$\beta_k \geq 0 \;\; \forall k$$

MKL learns both the kernel weights β and the SVM coefficients α. We used grid search to set the value of the regularization parameter C. The above equations are a generic formulation of the MKL-SVM problem and as specified earlier we used the method described in [8] to solve the optimization problem.

5. RESULTS

We wanted to highlight the importance of using multiple feature combinations for this particular task. This point has been shown using the results presented in Table 2, which shows the validation set accuracy for BoW features alone, 4 different feature pairs, and our final method which combined 11 features. The average accuracy for combining two features was approximately 34%. Our final submission was prepared using a combination of 11 feature kernels: HOG with 4 and 8 bins, PHOG with 4 bins, BoW on extracted faces with dictionary sizes of 200, 400, and 600, BoW computed on the entire image, LPQ-TOP with block sizes of 5, 7, and 9, and sound. The final submission to the challenge gave a classification accuracy of 37.08% on the validation set. For videos where no face was detected, we relied on an alternate classification pipeline built with the sound kernel only. Since 7 MKL-SVM classifiers were trained in a one-vs-all approach, the mean and standard deviation of the kernel weights across different classifiers has been shown in Table 3.

The pre-computed baseline performances have also been included in Table 2. The video-only baseline was calculated by the challenge authors using DPM [54] to localize the faces. These faces were then aligned and LBP-TOP [53] features were extracted from non-overlapping blocks of dimension 4x4. For classification a non-linear SVM was then learned on top of the concatenated LBP-TOP histograms. The audio-only baseline was computed using the same sound features as described earlier in this paper, but with a linear kernel SVM. For the audio-visual baseline the LBP-TOP and sound features were simply concatenated and used to train a non-linear SVM. The accuracy of the baseline method on the validation set for visual-only, audio-only, and visual+audio was 27.27%, 19.95%, and 22.22%, respectively.

The confusion matrix on the validation data for our final method is shown in Figure 3. Our classifier was most successful in handling Anger and Happiness, with respective accuracies of .71 and .68, and least successful in handling Fear and Disgust, which gave accuracies of .02 and .08, respectively. It can also be seen that a high percentage of other emotions such as Sad, Surprise, Disgust and Fear were wrongly classified as Neutral.

Method	**Accuracy**
BoW-600	33.16%
BoW-600 + LPQTOP-5	33.94%
BoW-600 + Sound	34.99 %
LPQTOP-5 + Sound	34.99 %
BoW-600 + gist	34.99 %
Final submission	**37.08** %
Baseline video	27.27
Baseline sound	19.95
Baseline video+sound	22.22

Table 2: The overall accuracy of our MKL-SVM pipeline using multiple descriptor combinations. Baseline accuracies are listed for comparison.

Kernel Name	**Mean Weight (Std)**
HOG-4	.5008 (.1167)
BoW-200	.2024 (.0614)
BoW-400	.1186 (.0544)
BoW-600	.1112 (.0230)
LPQTOP-5	.0252 (.0212)
Sound	.0184 (.0088)
HOG-8	.0177 (.0061)
LPQTOP-9	.0028 (.0029)
LPQTOP-7	.0008 (.0009)
BoW-FullScene	.0006 (.0010)
PHOG-4	4.4e-05 (.0001)

Table 3: List of all 11 descriptors used in our top-accuracy MKL-SVM framework along with their learned kernel weights, sorted in order of descending kernel weight. The mean and standard deviation of the kernel weights are calculated across classes.

6. DISCUSSION

The BoW descriptors on their own performed well above the video-only baseline, suggesting that these features are better-suited to the current task than LBP-TOP. In addition, BoW/LPQTOP+Sound and BoW+gist performed better in comparison to BoW+LPQTOP. Since the kernels combined in the BoW+LPQTOP method were both calculated from features extracted over aligned faces, the higher performance from using gist and Sound can be attributed to their encoding multimodal information beyond facial expression.

As opposed to using MKL for feature fusion, the baseline accuracies provided by the organizers were calculated by simply concatenating together multiple descriptors prior to training the non-linear SVM [16]. The concatenation method for feature fusion was clearly inappropriate for the AFEW dataset as it resulted in lower performance of the audio-visual system vs. classifiers trained using vision or sound only. This drop in performance could be attributed to use of a single value of the γ parameter for all features when constructing the RBF kernel. In contrast, MKL can control the relative complexity of the different feature kernels through distinct γ parameters for each feature. Furthermore, MKL performs feature selection by learning a convex combination of the kernels. While the concatenation method represents an equal confidence in each descriptor, MKL successfully handles discrepancies in the discriminative power of different features by assigning lower weights to less discriminative feature kernels. This is evidenced in the small weight of the sound kernel as compared to the visual kernels (see Table 3). From both the baseline results and our experiments it is clear that sound features are less discriminative for emotion recogntion in the AFEW dataset, and MKL sets the weights accordingly. Similarly, the weight for features extracted over the entire scene, .0006, is relatively small. Since we did not remove the face from these images, it could be possible that no information relevant to emotion recognition is added through analysis of the entire scene.

Another component of our MKL+SVM method was the use of one-vs-all as opposed to one-vs-one multi-class classification, due to lower performance observed in experiments using one-vs-one classification. To account for the observed decrease we point to the smaller number of negative training examples available to each classifier under the one-vs-one scheme. For one-vs-one classification the number of negative training examples is approximately 50, while for one-vs-all classification this number is about 300. We reason that a greater number of training examples is especially important to our task given the complexity of the AFEW dataset and the sensitivity of SVM classifiers to outliers.

Another improvement in our method resulted from combining DPM [54] for face detection with SDM [49] for tracking. Face detection without tracking sometimes results in facial alignment errors from one frame to the next; while many feature extraction methods, such as Bag of Words, are robust to these shifts in alignment, we still observed small improvements in emotion recognition when we used a tracking method in conjunction with DPM face detection.

7. CONCLUSION

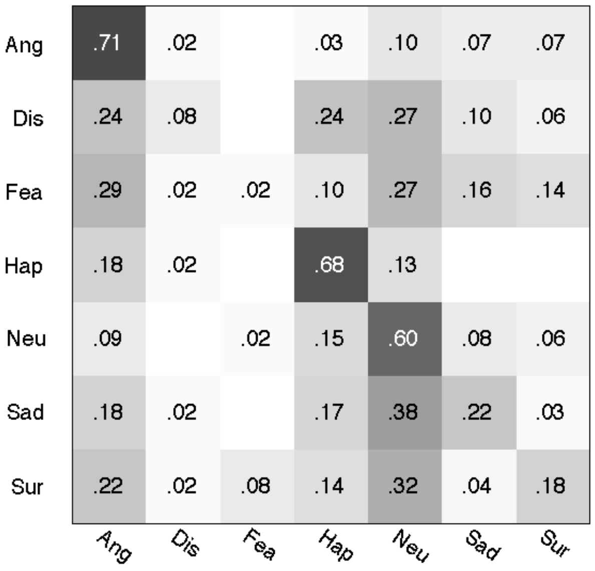
We proposed a novel method for multimodal emotion recognition in unconstrained near real-world conditions. We used a novel combination of DPM and SDM to yield a robust face initialization and tracking. Our pipeline of multiple kernel learning and support vector machine classification gave an optimal performance of 37.08% on the validation set, which was well above the highest challenge baseline of 27.27%. While our MKL-SVM method showed improvement above the baseline method, possible limitations include high computation costs for feature extraction and grid search. This could be mitigated through offline parameter selection using parallelization.

8. ACKNOWLEDGEMENTS

Support for this work was provided by NSF grants IIS 0905622 SOCS, SBE-0542013, and CNS-0454233, and by the Temporal Dynamics of Learning Center (TDLC) at UCSD.

9. REFERENCES

[1] M. Aly, M. Munich, and P. Perona. Multiple dictionaries for bag of words large scale image search. *Probe*, 1(D1):P1, 2011.



	Ang	Dis	Fea	Hap	Neu	Sad	Sur
Ang	.71	.02		.03	.10	.07	.07
Dis	.24	.08		.24	.27	.10	.06
Fea	.29	.02	.02	.10	.27	.16	.14
Hap	.18	.02		.68	.13		
Neu	.09		.02	.15	.60	.08	.06
Sad	.18	.02		.17	.38	.22	.03
Sur	.22	.02	.08	.14	.32	.04	.18

Figure 3: Confusion matrix for the best performing method on the validation set.

[2] H. Atassi, A. Esposito, and Z. Smekal. Analysis of high-level features for vocal emotion recognition. In *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on*, pages 361–366, 2011.

[3] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in manmachine interaction systems. In *in Proc. MLMI, LNCS 3361*, pages 318–328, 2004.

[4] M. Black, D. Fleet, and Y. Yacoob. A framework for modeling appearance change in image sequences. In *Computer Vision, 1998. Sixth International Conference on*, pages 660–667, 1998.

[5] J. Blitzer, K. Q. Weinberger, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

[6] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, 2007.

[7] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.

[8] S. Bucak, R. Jin, and A. K. Jain. Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition. In *Advances in Neural Information Processing Systems*, pages 325–333, 2010.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, ICMI '04, pages 205–211, New York, NY, USA, 2004. ACM.

[10] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[11] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. *BMVC*, 2011.

[12] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 17, pages 929–938, 2006.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, 2005.

[14] L. C. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, volume 1, pages 397–401. IEEE, 1997.

[15] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883. IEEE, 2011.

[16] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ACM International Conference on Multimodal Interaction*, 2013.

[17] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. A semi-automatic method for collecting richly labelled large facial expression databases from movies. *IEEE Multimedia*, 19:34–41, 2012.

[18] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(10):974–989, Oct. 1999.

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

[20] L. S. hsien Chen. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. Technical report, University of Illinois at Urbana-Champaign, 2000.

[21] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[22] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178, 2006.

[23] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[25] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. De La Torre, and J. Cohn. Aam derived face representations for robust facial action recognition. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 155–160, 2006.

[26] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.

[27] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[29] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *Image and Signal Processing*, volume 5099 of *Lecture Notes in Computer Science*, pages 236–243. Springer Berlin Heidelberg, 2008.

[30] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[31] J. Päivärinta, E. Rahtu, and J. Heikkilä. Volume local phase quantization for blur-insensitive dynamic texture classification. In A. Heyden and F. Kahl, editors, *Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 360–369. Springer Berlin Heidelberg, 2011.

[32] M. Pantic, I. Patras, and L. Rothkruntz. Facial action recognition in face profile image sequences. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 37–40 vol.1, 2002.

[33] M. Pantic, I. Patras, and M. F. Valstar. Learning spatio-temporal models of facial expressions, 2005.

[34] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[35] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54(1):329–349, 2003.

[36] K. R. Scherer. Adding the affective dimension: A new look in speech analysis and synthesis, 1996.

[37] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, 2010.

[38] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Multimodal approaches for emotion recognition: a survey. In S. Santini, R. Schettini, and T. Gevers, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5670 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 56–67, Dec. 2004.

[39] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning.

[40] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Computer Vision ECCV 2012. Workshops and Demonstrations*, volume 7584 of *Lecture Notes in Computer Science*, pages 250–259. Springer Berlin Heidelberg, 2012.

[41] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[42] A. Tawari and M. Trivedi. Audio-visual data association for face expression analysis. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1120–1123, 2012.

[43] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pages 149–149, 2006.

[44] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.

[45] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM, 2010.

[46] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[47] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[48] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig. Low-level fusion of audio, video feature for multi-modal emotion recognition. In *VISAPP (2)*, pages 145–151, 2008.

[49] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. CVPR, 2013.

[50] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 70–75, 1994.

[51] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[52] W. Zhang, A. Surve, X. Fern, and T. Dietterich. Learning non-redundant codebooks for classifying complex objects. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1241–1248. ACM, 2009.

[53] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.

[54] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.