Multiple Kernel Learning for Emotion Recognition in the Wild

Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort and Marian S. Bartlett

Machine Perception Laboratory
UCSD

EmotiW Challenge, ICMI, 2013

Task

- Emotion Recognition on the ‘Acted Facial Expression in the Wild dataset’- **AFEW**.

- Video clips collected from Hollywood movies.

- Classification into 7 emotion categories: Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise.

Challenges in AFEW

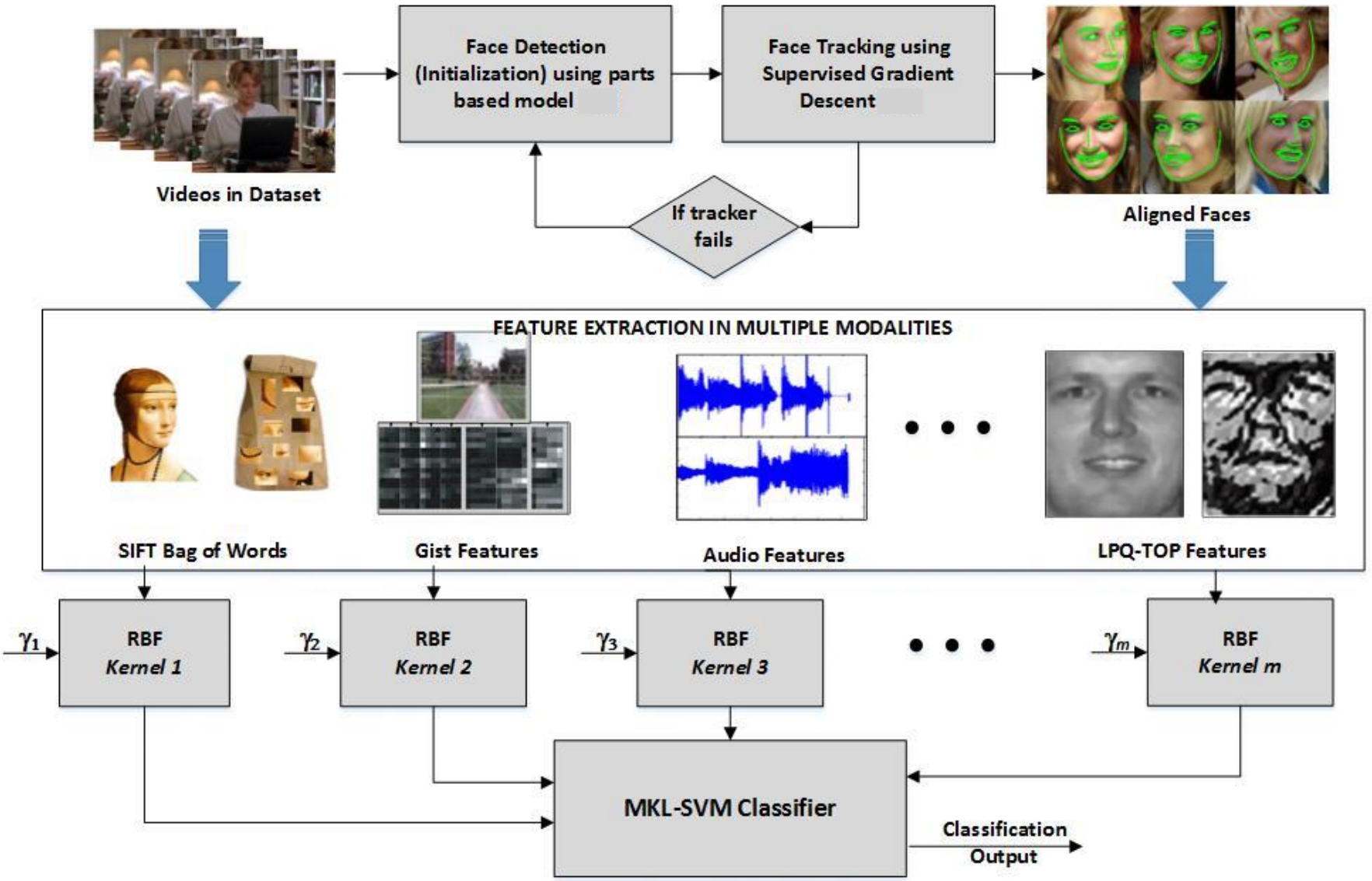
- Videos resemble emotions in real-world conditions.
- Others:
  - Pose Variations.
  - Occlusion.
  - Spontaneous nature of expressions.
  - Variations among subjects.
  - Small number of training samples given the complexity of the problem (~ 60 clips per emotion).

Our Approach

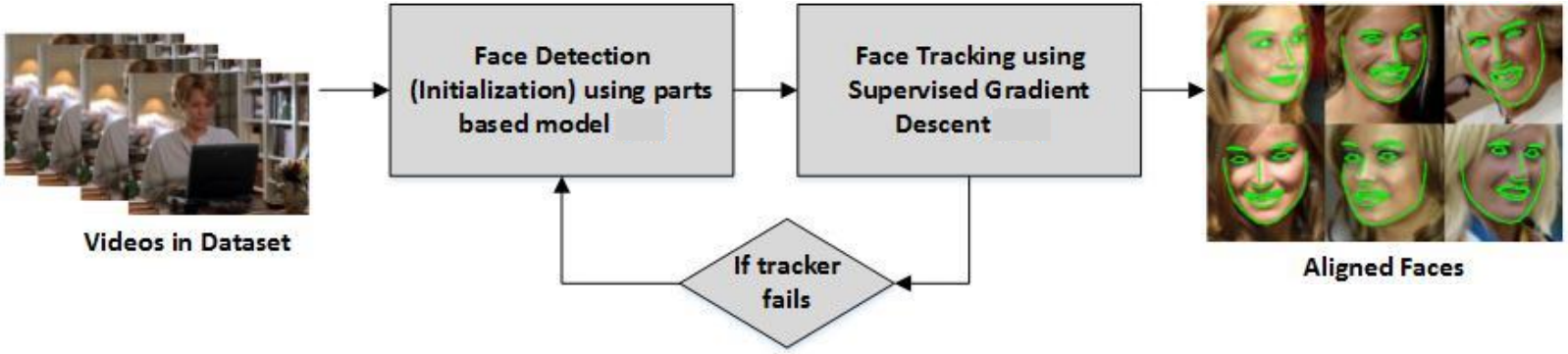
Multimodal classification system comprising of:

1. Face Extraction and Alignment.
   - Handle non-frontal faces.
2. Feature Extraction.
   - Visual and audio features.
3. Feature fusion using Multiple Kernel Learning.

Our Approach



Face Extraction and Alignment



- Combined state-of-the-art face detection method with state-of-the-art tracking method.

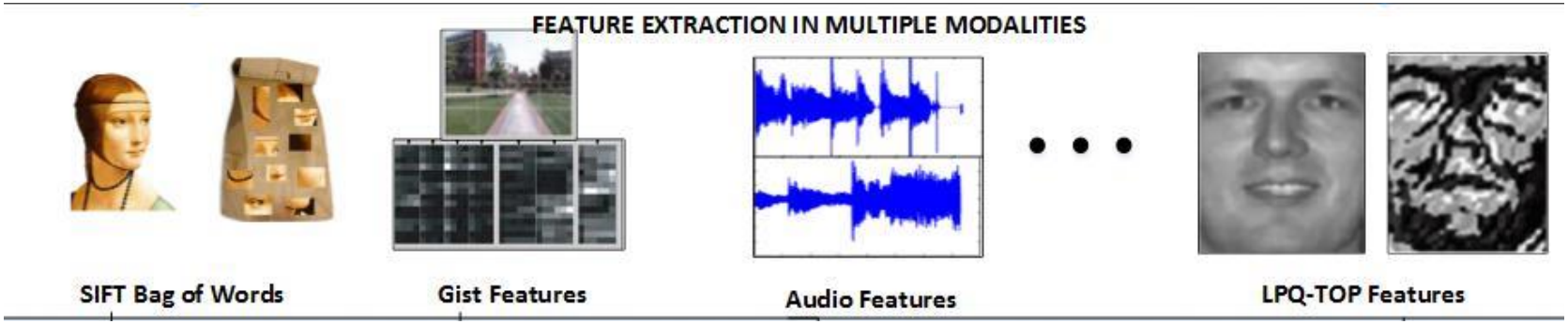
- **Face Detection**:
  - Deformable part-based model by Ramanam et al (CVPR'12).
  - Employs a mixture of trees model with shape model.
  - Ability to handle non-frontal head pose: critical for faces in AFEW.

Face Extraction and Alignment

- **Fiducial-point Tracker**:
  - Based on supervised gradient descent by Torre et al. (CVPR'13).
  - Returns 49 fiducial-points.

- Output from detector is fed to tracker.
- Re-initialization using detector if the tracker fails.
- Faces aligned with a reference face using affine transform.

Multimodal Features

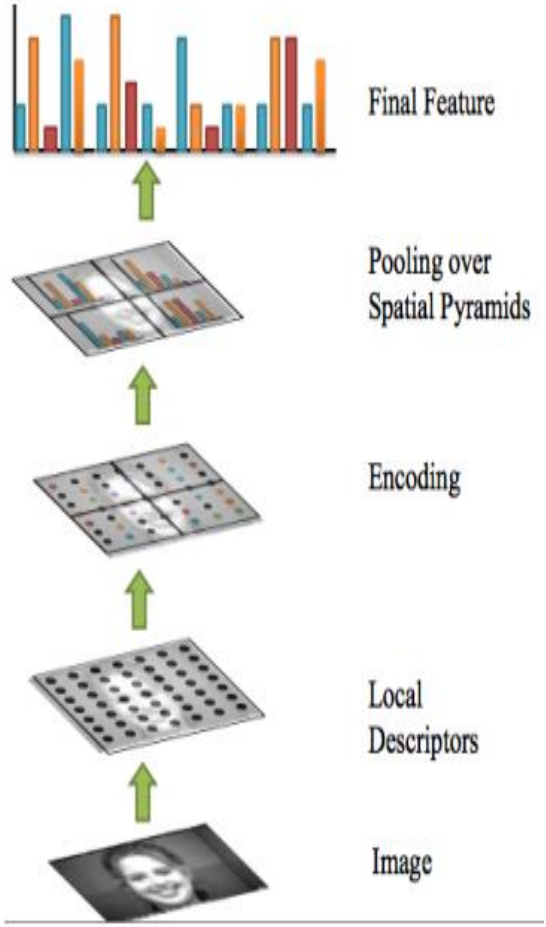


3 feature modalities:

- Facial features like BoW, HOG.
- Sound features.
- Scene or context features like GIST.

Facial Features

1. **Bag of Words (BoW)**:
   - State-of-art pipeline for static expression recognition by Sikka et al. (ECCV'12).
   - Based on multi-scale dense SIFT features (4 scales).
   - Encoding using LLC*.
   - Spatial information encoded using pooling over spatial pyramids.



* LLC- Locality constrained Linear Coding (Wang et al. 2010)

Facial Features

- Video features obtained by max-pooling over frame BoW features. (Sikka et al., AFGR'13).
- Robust compared to Gabor and LBP.
- Included multiple BoW features- constructed using different dictionary sizes (200, 400, 600).
- Motivated by recent success in multiple dictionary classification*.

*e.g. Aly, Munich, & Perona 2011, Zhang et al. 2009

Facial Features

2. LPQ-TOP*
   - Local Phase Quantization over Three Orthogonal Planes.
   - Texture descriptor for videos.
   - Robust variant of LBP-TOP.
   - Three set of features extracted with different window sizes of 5, 7 and 9.

* Päivärinta et al. 2011

Facial Features

3. HOG
   - Histogram of gradient features.
   - Describe shape information of objects using distribution of local image gradients.
   - Used for object detection and static facial expression analysis.

4. PHOG
   - Variant of HOG based on pyramids.

- Video features obtained by max-pooling over frame features.

Sound features

- Audio features improve performance of expression recognition systems (AVEC challenge).
- Employed paralinguistic descriptors from audio channel
  - Ex: MFCCs, fundamental frequency
- Summarized using functionals like max, min etc.
- 38 low-level descriptors + 21 functionals.
- Features provided by organizers.

Scene or Context features

- Investigated if scene information is relevant to recognition on AFEW.
- Two sets of features:
  1. BoW features extracted over entire image instead of just faces.
  2. GIST features (Oliva et al.)
     1. Output of bank of multi-scale oriented filters + PCA.
     2. Popular to summarize scene context.

Feature Fusion

- Multiple features encode complementary information discriminative for a task.
- Combining features -> improves classification accuracy.
- Techniques for fusing features:
  1. Feature concatenation.
  2. Decision (classifier) level fusion.
  3. Multiple Kernel Learning (MKL) strategy.
- MKL is more principled since it can be coupled with classifier learning, e.g. with a SVM.

Multiple Kernel Learning

- Used Multi-label MKL (Jain et al., NIPS'10).
- Estimates optimal convex combination of multiple kernels for training SVM.
  - Formulates MKL as a convex optimization problem.
  - Globally optimal solution.
- Unique kernel weights are learned for each class.

Our Approach

- Our approach fused different features using MKL.
- Referred to as **All-features + MKL** in results.
- RBF kernels used as base kernels for all features.
- Employed one-vs-all multi-class classification strategy instead of one-vs-one in SVM.
  - More training data per classifier.
  - Showed improvement in results.
  - Class assignment based on maximum probability across the per-class classifiers.

Experiments

- Kernel and SVM hyper-parameters obtained by cross-validation on validation set.
- Performance metric is classification accuracy on the 7 classes.

Results
Validation Set

Features	Accuracy
Baseline video (LBPTOP)	27.27%
Baseline sound	19.95%
Baseline video + sound	22.22%

- Baseline-performance on validation set.

Results
Validation Set

Features	Accuracy
Baseline video (LBPTOP)	27.27%
BoW-600	33.16%

- BoW shows an advantage of 5% compared to LBPTOP used for baseline.
- Performance boost attributed to both (1) better face alignment + (2) more discriminative BoW features.

Results
Validation Set

Features	Accuracy
Baseline video (LBPTOP)	27.27%
Baseline sound	19.95%
Baseline video + sound (**Feature concatenation**)	**22.22%**
BoW-600	33.16%
BoW-600 + Sound (**MKL**)	**34.99%**

- Fusion method ‘feature concatenation’ leads to fall in performance for baseline features.
- However, performance rises for feature fusion using MKL.
- Highlights advantage of MKL.

Final Results

Validation Set	
Method	Accuracy
Baseline video (LBPTOP)	27.27%
BoW-600 + Sound + MKL	34.99%
All features + MKL	**37.08%**

Test Set	
Method	Accuracy
Baseline video (LBPTOP) + audio	27.56%
All features + MKL	**35.89%**

- Best accuracies are reported for baseline approaches.
- All-features + MKL is the **proposed approach**.
- Using **multiple features** gives significant improvement over just BoW-600 and sound features.

Kernel Weights

Visual features

Sound features

Context features

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)

- Mean and standard deviation are calculated across kernel weights learned for each class.

Kernel Weights

- Visual features are more discriminative compared to sound features.
- Highest weights are assigned to HOG and BoW kernels.
- Context based features:
  - BoW over entire scene (including faces) weight of .0006.
  - Information from this BoW kernel could come from both face and scene information.
  - GIST features not included in final features because they did not improve performance.
  - Scene information might not be discriminative.

Insights

- MKL works better than naïve feature fusion using feature concatenation.
- MKL allows separate γ for each RBF feature kernel leading to better discriminability.
- Fusion of visual and soundˢ features leads to improvement in results (multimodality).
- Found improvements in result using one-vs-all multi-class strategy.

Conclusion

- Proposed an approach for recognizing emotions in unconstrained settings.
- Our method of **combining multiple features using MKL** shows significant improvement over baseline on both test and validation set.
- Highlighted advantage of using both (1) multiple features, and (2) MKL for feature fusion.
- Investigated learned kernel weights to show the contribution of different kernels.

Thanks

- Pl. forward any questions to ksikka@ucsd.edu

- Thanks to our Presenter Yale Song, Graduate Student, Multimodal Understanding Group, MIT.