To appear in In A. Fitzgibbon, S. Lazebnik, Y. Sato, and C. Schmid (Eds), European Conference on Computer Vision, Workshop on What's in a Face. Lecture Notes in Computer Science, Springer, 2012.

Exploring Bag of Words Architectures in the Facial Expression Domain

Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett

Machine Perception Laboratory, University of California San Diego
{ksikka,ting,josh,marni}@mplab.ucsd.edu

Abstract. Automatic facial expression recognition (AFER) has undergone substantial advancement over the past two decades. This work explores the application of bag of words (BoW), a highly matured approach for object and scene recognition to AFER. We proceed by first highlighting the reasons that makes the task for BoW differ for AFER compared to object and scene recognition. We propose suitable extensions to BoW architecture for the AFER's task. These extensions are able to address some of the limitations of current state of the art appearance-based approaches to AFER. Our BoW architecture is based on the spatial pyramid framework, augmented by multiscale dense SIFT features, and a recently proposed approach for object classification: locality-constrained linear coding and max-pooling. Combining these, we are able to achieve a powerful facial representation that works well even with linear classifiers. We show that a well designed BoW architecture can provide a performance benefit for AFER, and elements of the proposed BoW architecture are empirically evaluated. The proposed BoW approach supersedes previous state of the art results by achieving an average recognition rate of 96% on AFER for two public datasets.

1 Introduction

Automatic Facial Expression Recognition (AFER)[?,?,?,?] has been gaining momentum over the years due to its application in multiple domains such as human computer interaction, and analyzing human behavior, among others. This progress has been possible due to advancement in computer vision and machine learning, along with higher computing power.

Bag of words (BoW) models represent images as orderless collection of local features. These models have recently shown remarkable performance in multiple domains including scene recognition [?,?], object and texture categorization [?,?,?], and human activity recognition [?]. Applicability to multiple domains can be attributed to the simplicity of the BoW model along with significant research in feature extraction [?,?], codebook construction [?,?], application of fast but efficient kernels for discriminative classification [?,?,?], and encoding schemes [?,?]. Efficient methods for pooling and matching over spatial and temporal grids at multiple scales [?,?] has helped extend BoW to domains where spatial or temporal information is important.

A strength that BoW brings to the table is invariance. Feature values are pooled within a specific region of space (or time) without reference to exactly where in the window the feature occurred. This gives BoW models tolerance to small perturbations in the positions of image features, making them robust to variations in the shape of a cup, for example. In contrast, facial expression is a subordinate level classification problem defined by nonrigid deformations of a basic object shape. Fundamental differences have been described between face and object recognition, in which objects can be defined by the presence of component parts with substantial tolerance to metric differences in their positions, whereas faces consist of the same parts in approximately the same relations, and detection of metric variations in a holistic representation takes on much more importance [**?**]. Hence the invariance structure of the facial expression task differs from those of object recognition, potentially requiring for example, information on multiple spatial scales. It therefore stands to reason that an optimal BoW architecture suited for object recognition may also differ for the AFER task.

Recently BoW has been applied to problems involving discrimination of subordinate-level categories such as flowers [**?**] and breeds of cats and dogs [**?**]. These findings further motivated us to look into BoW for the problem of facial expression recognition.

Our contribution lies in a principled exploration of BoW architectures for the AFER domain, and investigating whether the resulting BoW architecture is competitive with state-of-the-art performance for current approaches for AFER. While a small number of previous papers employed BoW for an AFER task (e.g. [**?**]) this is the first to present a principled exploration of BoW design in this domain. Thus a clear picture has not yet emerged of a BoW architecture best suited for the AFER domain, and how it compares to current approaches to AFER. Specifically, this paper explores BoW for an appearance-based discriminative approach to AFER. Appearance based discriminative approaches have been highly successful for face detection [**?**] face identification [**?**] and expression recognition [**?**] [**?**], and can provide person-independent recognition performance in real-time [**?**]. State of the art AFER performance has been shown for Gabor energy filters [**?**] [**?**], and local binary patterns (LBP) [**?**]. Here we compare the performance of BoW framework to state-of-the-art AFER approaches based on Gabor and LBP features.

Our **contributions** are as follows:

1. We identify inherent challenges in appearance based approaches to AFER and argue that most of these can be addressed through a suitable BoW pipeline. We highlight reasons why AFER differs from domains where BoW has been applied successfully in the past, and which would influence the architecture of our BoW model.
2. In order to address the demands of AFER task, we propose a BoW approach that combines highly discriminative Multi-Scale Dense SIFT (MSDF) features with spatial pyramid matching (SPM). To the best of our knowledge, this is the first work employing MSDF features for AFER. The spatial pyramid representation provides spatial information to BoW to maintain both

holistic and local characteristics. This is augmented with the recently proposed locality-constrained linear coding (LLC) [**?**] along with max-pooling, leading to a representation that is powerful even with a linear classifier.

3. We compare the BoW performance to two of the most successful algorithms for AFER: Gabor wavelets and LBP, as input to Support Vector Machines (SVM). Experiments on two public datasets demonstrate that the proposed BoW pipeline is highly effective for AFER, giving significantly better performance than approaches relying on Gabors or LBP features.
4. Experiments further demonstrate the contributions of specific components of the proposed BoW pipeline. We show that MSDF features compare favorably with Gabor and LBP features when input directly to a SVM, however a significant hike in performance is obtained by integrating them with proposed BoW architecture. We further demonstrate the advantages of employing multi-scale features, compared to single scale features, and LLC coding, compared to traditional hard coding, in our BoW pipeline.

2 Related Work

Appearance based approaches can extract features densely at every face pixel, as in gabor wavelets [**?**], or can pool features in certain regions of support. For instance in the case of LBP, the face is divided into regular overlapping grids [**?**], while for the approach in [**?**] features are pooled over 4 segmented facial regions around eyes, nose etc. It is possible to represent many AFER approaches through a general block-diagram as shown in Fig. 1. Strategies for each block for different methods used in this paper are given in Table. 1.

Although the Gabor energy representation possesses some shift invariance, their spatial invariance is low relative to other image features explored here. This could make them less robust to alignment errors that can result from variations such as head pose, ethnicity etc. LBP based methods achieve some invariance by employing a rectangular grid as a region of support. Selecting an apt grid pattern is a non-trivial problem for LBP based methods.

A previous paper [**?**] constructed a BoW representation by pooling features over 4 facial regions, which are obtained through a segmentation step. This intermediate segmentation step poses the problem of using a non-standard support region for feature pooling, and the final representation may critically depend on a good segmentation. Since the algorithm is evaluated only on a single dataset, it is unclear if the proposed segmentation works well across multiple datasets. Next, separate dictionaries were learned for each of these sub-regions, which may lack sufficient features for robust codebook construction. Moreover, the BoW representation in [**?**] did not achieve strong performance alone, and required the addition of pyramid histogram of gradients (PHOG) features to obtain appreciable results. This left the question open as to whether BoW itself provides coding advantages in the AFER domain. This work attempts to address this question.

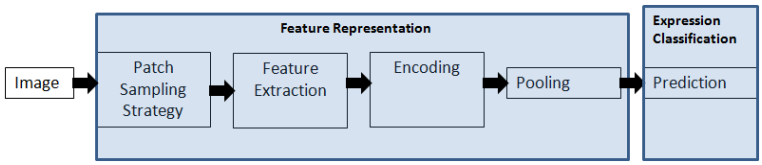

Fig. 1. General Block diagram for AFER

3 Proposed Bag of Words Approach

The AFER community follows a common practice of separating the face detection and alignment task from the task of expression recognition, factoring out rigid deformations in order to make the problem more tractable. Hence the AFER task does not typically entail invariance to transformations such as scale, translation, or in-plane rotations. The goal is to detect the appearance or texture variations in face images that result from facial expression, while allowing invariance to factors such as differences in the positions of the lip corners for a low versus high intensity smile, or differences in facial wrinkling patterns due to age or facial physiognomy, or variations due to small residual rotation and alignment errors. The invariance task therefore may be on a different scale or along different feature dimensions than for many object recognition tasks. We address these goals by exploring a BoW architecture with suitable extensions, which are mentioned in a sequential order below. The ordering adheres to the general block diagram in Fig. 1.

1. **Sampling strategy and features:** BoW represents images as collection of independent local patches. These patches can either be sampled as sparse informative points, i.e. ones extracted using various interest point detectors [**?**,**?**], or as dense patches at different locations and scales. Work in [**?**] pointed out that, (1) interest points based methods often saturate in performance since they can only provide limited patches, and (2) the minimum-scale of patches has a considerable influence on results because the vast majority of patches or interest points typically occur at the finest few scales. Thus, based on these findings we compute SIFT descriptors (constant orientation) densely on the image by extracting them with a stride of 2 pixels. In this paper we explore a multi-scale dense SIFT representation (MSDF) [**?**], which pools over multiple region sizes, retaining more spatial information at the smaller scales. Here we employed 5 different scales, defined by setting the width of the SIFT spatial bins to 4, 8, 12, 16, and 24 pixels. Empirical comparisons demonstrated that multi-scale SIFT features indeed lead to better performance than their single-scale counterpart. SIFT features have been extracted using code from author's website in [**?**].
2. **Codebook Construction (for encoding):** The vocabulary is constructed by quantizing SIFT descriptors randomly selected from the training images using approximate K-means clustering. This clustering approach is based on

calculating data-to-cluster distances using the Approximate Nearest Neighbor (ANN) algorithm. We used the implementation provided by authors in [**?**]. The size of the dictionary was set based on empirical experiments to 800. We have provided a brief discussion in Sec. 5 on the effect of vocabulary size on recognition rates.

3. **Encoding and Pooling:** The original BoW model ignores the spatial order of local descriptors. However spatial information is necessary for AFER. To address this particular limitation, we have incorporated a particular extension of BoW [**?**] called *spatial pyramid matching* (SPM), that has shown significant success on a number of tasks [**?**,**?**]. For our spatial pyramid representation, we partition an image into $2^l \times 2^l$ segments in different scales $l = 0, 1, 2, 3, 4$, compute the BoW histograms within these 341 segments and finally concatenate the histograms into a single feature vectors. This strategy of pooling over spatial pyramids is more structured and eliminates the need to select the right grid pattern, as was the case for LBP and the approach in [**?**]. Experimental results demonstrate that spatial pyramid matching contributes substantially to performance for AFER.
   Recent research shows that the choice of encoding and pooling has a significant effect on the classification accuracy [**?**]. Hence this work employs the recently proposed LLC encoding along with max spatial pooling [**?**] to construct the final multi-scale representation. LLC encoding projects each descriptor to a local linear subspace spanned by some codewords by solving an optimization problem [**?**]. This approach is known to be more robust to local spatial translations and captures more salient properties of visual patterns as compared to the original simple histogram spatial encoding [**?**]. Most importantly a linear kernel is sufficient to achieve good performance with LLC encoding, thus avoiding the computational expense of applying non-linear kernels [**?**], as is the case with spatial histogram based encoding [**?**,**?**]. The two parameters for LLC encoding were (1) M, the number of nearest visual words to be considered, was set to 5, and (2) parameter β used in the computation of the projections was set to 10^{-4}, as done in [**?**].

4 Experiments

4.1 DATASETS

These algorithms were evaluated on two public datasets, the Cohn-Kanade+ (CK+) [**?**] and Amsterdam Dynamic Facial Expression Set (ADFES) [**?**]. Of these, CK+ is the most widely used dataset for AFER, while ADFES as an evaluation dataset for AFER task has been introduced in this paper. The rationale behind using these two datasets is that they both consist of directed facial expressions, validated by the Facial Action Coding System (FACS).

CK+: The CK+ dataset consists of 123 subjects between the age of 18 to 50 years, of which 69% female, 81% Euro-American, 13% Afro-American, and 6% other groups. Subjects were instructed to perform a series of 23 facial displays, six of which were based on description of prototypic emotions. For our experiments,

we used the 327 peak frames which had emotion labels validated by trained professionals as provided by the authors. The faces were cropped and resized as mentioned in Section 4.2. A salient feature of this dataset is that the authors have provided a standard evaluation procedure [**?**] so that it is possible to compare results from different works directly. The algorithms are tested on a leave-one-subject-out cross-validation experiment where an expression is classified into one of the 7 classes, namely- angry, contempt, disgust, fear, happy, sadness and surprise.

ADFES: The ADFES dataset consists of 22 subjects aged from 18 to 25 years old, of which 10 are Mediterranean and 12 are North-European, and 10 are females and 12 are males. The subjects were instructed to perform nine emotional states of which six popular *basic emotions* [**?**] are being used for our experiments. Again the faces were registered into 96×96 patches as mentioned in Section 4.2. The algorithms are evaluated on 5 fold cross-validation experiment where an expression is classified into one of the six classes

4.2 Methods

Pre-processing: Faces were detected automatically by a variant of the Viola and Jones detector [**?**] and normalized to 96×96 patches based on the location of the eyes.

The first baseline method was based on Gabor wavelets. Features were extracted using 72 Gabor filters spanning 8 orientations and 9 scales, whose parameters were selected as in [**?**]. The images were first convolved with each filter in the filter bank and the output magnitudes were concatenated into a feature vector. The second baseline method was LBP. Implementation of LBP features was obtained from the authors' [**?**] website. Normalized uniform LBP histograms were extracted with an exhaustive combination of parameters (namely neighborhood parameters (P, R), number of horizontal and vertical grids and overlap ratio between blocks) and the best results were reported. The rationale behind running the experiment with an exhaustive combination of parameters was to realize the degree to which the performance of the algorithm depends on the chosen parameters. Our experiments revealed that the best parameters for LBP were different for both datasets indicating that this method might be parameter dependent.

Classifier: As mentioned earlier a linear support vector machine (SVM) [**?**] was used as the classifier for the proposed BoW method as well as for the Gabor baseline. For LBP and simple BoW, SVM with a polynomial kernel of degree 2 and a histogram intersection kernel was used for consistency with previous papers employing these methods [**?**] [**?**]. To avoid over-fitting we applied a double cross-validation method. Double cross-validation is a method to estimate separate training, test, and validation set performance in small datasets. Here, the hyper-parameters for the SVM (or kernel) were obtained by selecting the parameters with best performance on a 25% subset of the training samples. For each experiment we have reported the average percentage accuracy and standard error measure.

Most of the details of the proposed BoW algorithm are described in Sec 3. Since experiments for CK+ are run in a leave-one-subject-out fashion, we chose to prevent re-clustering of data for each fold by constructing a dictionary once from samples in the ADFES dataset and using the same for each fold. This strategy also tests the generalization capability of the BoW approach, where a different dataset is used for codebook construction.

Method	Pixel Sampling	Features	Encoding	Pooling Strategy
Gabor	1 Pixel	Gabor filter outputs (9 scales & 8 orient.)	No	Concatenation
LBP	1 Pixel	LBP	No	Sum-pooling over regular grids
MSDF	2 Pixel	SIFT (5 scales)	No	Concatenation
Simple BoW	2 Pixel	SIFT (5 scales)	Hard encoding	Sum-pooling over Spatial Pyramids
SS-SIFT+BoW	2 Pixel	SIFT (1 scale)	LLC	Max-pooling over Spatial Pyramids
Proposed Method (MSDF+BoW)	2 Pixels	SIFT (5 scales)	LLC	Max-pooling over Spatial Pyramids

Table 1. Methods compared in this paper for AFER. All methods below the double line are based on BoW. Those above the double line are passed directly to an SVM.

5 Results

The performance statistics for prescribed experiments on CK+ and ADFES are shown in Table. 2. It is evident that the proposed BoW approach outperforms previous state of the art AFER approaches, Gabors and LBP, on the two recognition tasks. Moreover to the best of our knowledge these are the best publishable results for AFER on both CK+ and ADFES. This observation supports our contention that a suitable BoW architecture can indeed provide a performance benefit for AFER.

Regarding dictionary size, we found that increasing the dictionary size improved performance for small dictionary sizes, saturated at 800, and began to decrease for larger dictionary sizes. Performance started decreasing significantly at dictionary size beyond 2000.

Regarding image features, we note that raw MSDF features performed better than both Gabor and LBP on CK+, although not on ADFES. In order to support the argument that the proposed BoW architecture is responsible for boost in performance and not MSDF features alone, we extracted MSDF features as described in Section 3, concatenated all the features in an image into a single vector and passed it straight to an SVM, which is similar to that used in Gabor based methods [**?**]. This was compared to the same MSDF features in the pro-

DATASET	ADFES	CK+
Gabor	94.59 ± 2.61	91.81 ± 1.94
LBP	94.96 ± 1.96	82.38 ± 2.34
MSDF	92.59 ± 3.41	94.34 ± 1.62
Simple BoW	94.09 ± 2.32	92.67 ± 1.93
SS-SIFT+BoW	93.3 ± 1.13	93.28 ± 1.76
Proposed Method	$\mathbf{96.30 \pm 1.08}$	$\mathbf{95.85 \pm 1.40}$

Table 2. Comparison of the proposed BoW pipeline with other approaches- Gabor wavelets [**?**], LBP [**?**], MSDF consisting of vectorized MSDF features passed to a linear SVM, Simple BoW, and Single-Scale SIFT+BoW as described in Section 4.2.

posed BoW architecture. Table. 2 shows that the proposed method contributes performance benefit beyond MSDF features alone.

Next, to highlight the advantage of employing multi-scale features compared to single scale features, we also ran our proposed pipeline, except with Single Scale SIFT (SS-SIFT) features instead of MSDF features. For SS-SIFT we used a spatial bin size of 4 pixels and a stride of 2 pixels. Table. 2 shows that the multiscale features in the **Proposed Method** give about a 3 percentage point advantage over the same model implemented with single scale features.

We also evaluated coding strategies. We compared the LLC encoding with max-pooling to the simple histogramming strategy [**?**], comprised of traditional hard quantization and sum-pooling. We refer to this as "simple BoW". By comparing the **Proposed Method** to simple BoW, we observe that selecting LLC encoding along with max pooling in our proposed method leads to consistent improvement in performance for both datasets.

The most substantial performance benefit was provided by the spatial pyramid matching. Removal of the SPM from the BoW model, eliminating all spatial information, reduced classification performance on CK+ from 95.9% to 83.1%($\pm$2.5).

6 Conclusion

This paper explores the applications of Bags of Words, a technique highly successful in object and scene recognition community, to AFER. We first highlighted reasons that had hindered the success of BoW for AFER, and then proposed ways to tackle these issues based on recent advances in computer vision. Elements of the proposed BoW architecture were empirically evaluated.

Spatial information at multiple scales is crucial for AFER as compared to object and scene classification tasks. Hence spatial pyramid matching is recommended for the BoW architecture to preserve spatial information during matching, and we showed that performance drops substantially without it. We also discovered that multi-scale features are necessary for AFER as is the case for most object and scene categorization tasks. In particular BoW allowed us to effectively employ highly discriminative multi-scale SIFT features. We also showed

that the performance benefit was not from the MSDF features alone. The BoW architecture itself contributed further to performance. Advantages of employing novel encoding and pooling strategies as compared to standard histogram of quantized descriptors was also shown for the task of AFER.

State of the art results on two public datasets highlight the effectiveness of this approach on AFER. In conclusion our findings support the claim that a principled BoW architecture can provide a performance benefit for the AFER task.

Acknowledgment. Support for this work was provided by NSF grant IIS-0905622. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.