Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 10, Issue, 06, pp.70033-70038, June, 2018

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

FACIAL EXPRESSIONS RECOGNITION VIA PATCH BASED DESCRIPTORS (PD-FER)

***Shafiq-ur-Rehman, Junaid Baber, Maheen Bakhtyar, Imam Dad, Anwar Ali Sanjrani**

University of Balochistan, Pakistan

ARTICLE INFO

Article History:
Received 24th March, 2018
Received in revised form
07th April, 2018
Accepted 20th May, 2018
Published online 28th June, 2018

Key words:
SIFT, Multiscale Descriptors,
Facial Expression Recognition.

ABSTRACT

Image classification has gained vital attention in recent years due to vast applications such as object recognition, face detection, face recognition, and facial expression recognition. Images are represented by robust and distinctive features such as SIFT and BIG-OH which are used for machine learning. The features are learned; specially SIFT which is gold standard, from local patches within the images. These local patches are mostly of the size of 41 x 41 pixels. These patches have shown significant low accuracy when the size of databases increase. In this research, variation of SIFT descriptor is proposed which has shown better accuracy when used on challenging dataset of facial expression recognition. The SIFT descriptor is computed from local patches on multiple scales, the similar approach has shown better performance in Image retrieval based applications. The SIFT obtains 55.6% accuracy on Fer2013 dataset using with bag of visual word model, whereas, the proposed extension obtains 58.3%.

****Corresponding author:***

Citation: Shafiq-ur-Rehman, Junaid Baber, Maheen Bakhtyar, Imam Dad, Anwar Ali Sanjrani, 2018. "Facial Expressions Recognition via Patch based Descriptors (PD-FER)", *International Journal of Current Research*, 10, (06), 70033-70038.

INTRODUCTION

Human Machine Interaction (HMI) systems are the kind of system which can deal and understand human emotions and feelings. Current HMI systems still have some drawbacks to reach full emotional and socially capable, necessary for robust and efficient interaction with human beings and need to be focused for making efficient HMI. Facial Expressions is one of the most important thing in social interaction and it is one the most crucial non-verbal channel via which HMI can recognize person's interactional emotions. The automatic facial expression recognition has recently gained wide attentions of the researchers around the world. There are six basic facial expressions reported in literature; sad, happy, anger, fear, disgust and surprise (Ekman, 1971). Facial Expressions Recognition (FER) is important in designing the human computer interaction and HRI i-e Human Robot Interaction systems (Mollahosseini, 2014). Many different annotated datasets (Gross, 2010; Pantic *et al*., 2005; Lyons *et al*., 1998) or faces captured spontaneously in an uncontrolled setting (Mavadati *et al*., 2013; Dhall, 2013), has been designed for evaluation of FER systems and many machine learning and computer vision algorithms has also been developed for automated FER systems.

FER systems are used to classify the faces on the bases of their emotions and facial expressions. Many traditional classifiers such SVM and to lesser extend Bayesian classifier, have been successful when classifying the faces being captured on controlled environment, many researches showed that these techniques do not perform well when classifying the images captured in spontaneous uncontrolled environment or when these techniques are tested on the dataset for which they were not designed (Mayer, 2014). This is due to the fact that many classifiers are only limited to the databases for which they have trained on and can only recognize the emotions similar to those in training databases. Moreover, obtaining the proper, accurate and complete database for all the kinds of emotions particularly for sadness and anger is difficult. Due to the progress in machine learning algorithms, innovation in many different approaches and availability of high computation power and larger dataset to work with, neural network takes a lot of attention of researchers and has gained much popularity in the field of object recognition, human pose estimation, face verification, and many more. In traditional approaches and methods, images are being described by the handmade features which do not perform well in many different situations, in this case neural network performs well in describing the objects and images and we can visualize much better results using Deep Neural Networks (DNN).

These networks extract undefined features from the database and able to extract more robust and distinguished features which can be used for complex problem of classification and many other purposes. These kind of network even perform well for the scenarios on which the network has not been trained on. In our case we have used SIFT which is of 128 dimension on multi-scale for the description of an image, proposed by (Baber, 2015) and further used these robust and scale invariant features for facial expression classification. The present paper presents the extension of the methodology proposed by Baber *et al*. (2015) which computes the patch based descriptors on multiple scales, as shown in Figure 1. We have modelled the multi-scale descriptors for facial expression recognition. Rest of the paper is organized in following sections. Section 2 consists of Related work and Section 3 comprised of the proposed methodology, and Section 4 discusses the experiments and results in detail.

Related work: The section of literature review comprises of two subsections in which we discussed state of the art methodologies on local keypoint descriptors and facial expression recognition. In the first subsection we have discussed local keypoint descriptors and in the second subsection we have presented the state of the art methodologies on Facial Expression Recognition (FER). Initially in the field of computer vision, global features were used, which describe image as a whole but later on switched on local features which describe the patch of the images also known as keypoints in the image. Both have their own pros and cons. Global descriptors have limitations which local features overcome with computational cost. We will discuss each of them in this section. Chang *et al*. (1998) introduced a novel approach and proposed a system known as RIME (Replicated Imaged Etector). This system was able to detect the images (which are pirated copies) on the internet using wavelets and colour spaces. This system was found efficient and accurate for some basic type of transformations. Kim (2003) proposed a new method for content based copy detection (CBCD) and they argued that colour does not play a vital role for copy detections whereas it is important for image retrieval (images are similar based on colour, texture, or objects). So they used discrete cosine transform (DCT) which is robust to distortions and many kind of changes in images, for copy detection in images. They first converted the images into the YUV format and used Y component of image in the proposed method. By using this methodology, the authors successfully detected the copies of the test images but they were not succeeded in the detection of copies with 90°or 270°rotation (Wan *et al*., 2008).

As stated above that global features have many limitations. The global features do not perform well and are not robust for severe type of transformations, its performance is not so good, e.g. in the matter of cropping, occlusion and aspect ratio change, global features fail to perform well, however it is good for simple type of transformations. For severe type of transformations in images, local descriptors perform well and have proven to be more robust and efficient than global descriptor. Many problems related to CBCD and image retrieval was proposed using SIFT and other kind of local descriptors because of their robust and transformations invariant nature (Chum, 2011; Nister, 2006; Philbin, 2007; Lost in quantization, 2011; Wu, 2009; Xu *et al*., 2011; Zhou *et al*., 2011; Zhou *et al*., 2010). Many researchers worked on CBCD using local features or descriptors. Xu *et al*. (2011) proposed a framework for CBCD using SIFT and spatial features, but their system performed poor in the presence of occlusions. They detected the circular patches from the images using the SIFT detector and later on computed the multi-resolution histograms as feature vectors of the images. Zhou *et al*. (2010) used Bag of Visual Word model for the partial image copy detection and proposed a framework for large scale applications. In their methodology authors quantized the SIFT descriptors in descriptors space and orientation space. They further used XMAP and YMAP strategy for encoding the spatial layout of keypoint, which helped them to eradicate the outliers. But their framework is sensitive to digital images error such as drifting or shifting of keypoints because of transformation and because of this, their methodology missed many true matches. Wu *et al*. (2009) proposed a framework in which authors have used group of keypoints rather than using single keypoints. The authors in their research have used SIFT along with the MSER regions for image descriptions. They have used SIFT keypoints with maximally stable extremal regions (MSER) keypoints, as MSER keypoints are affine covariant keypoints and have higher repeatability than that of SIFT keypoints. These keypoints are larger in scale but smaller in number in image. If these keypoints are used with SIFT keypoints, they perform very much better with high accuracy and higher discriminative power and perform well with 45% better accuracy in BoVW model, as compared to only simple SIFT BOW model. But the only limitation in their system is computational time. The system was not time efficient when compared with literature.

Patch based descriptors: There are many feature point descriptors in computer vision, but most of the successful keypoint descriptors are categorized into two types: one is based on gradient histograms (Bay *et al*., 2008; Ke, 2004; Lowe, 2004; Mikolajczyk, 2005), while the other one is based on local pixel intensity without explicitly gradient calculation (25)– (28). In our experiments we have used two different techniques, one is Scale-invariant feature transform (SIFT) (23) with multiscale (9), and other one is CSLBP (Center-Symmetric Local Binary Pattern) (27).

Centre-Symmetric Local Binary Pattern: (CSLBP) which is Centre-Symmetric Local Binary Pattern, is the extension of simple local binary pattern (LBP). The functionality of LBP is to compare the pixel p with each of its neighbour$N$ with radial distance R. The output of LBP will be considered 1 if and only if the pixel value is smaller than its neighbour and the output will be considered 0 if the pixel value is greater than its nearest neighbour. The length of the output for pixel p will be of N bits because each pixel p is going to be compared with N neighbours. Mathematically it is shown in Equation 1. The values of CSLBP parameters in our experiments is as follows; for N we have taken 8 and for R, the value is 1. The length of LBP histogram for each image is 2^N, while in CSLBP this histogram is quantized further. In CSLBP only centre-symmetric neighbours are compared with each pixel p.

$$CSLBP_{N,R,T}(p) \quad \sum_{i=1}^{\frac{N}{2}} s(|n_i| - |n_{i+\frac{N}{2}}|)2^{i-1},$$
$$s(j) \quad \begin{cases} 1 & j > T \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In CSLBP, after the step of quantization we get our final histogram which is of $2^{\frac{N}{2}}$ in length which is quite small than LBP.

The value suggested in research by many researchers for N is 8, R is 1 and for T is 0.01. For the computation of CSLBP, the CSLBP is computed for each cell and the given patch P or cell of an image is further divided into spatial grid of $G_x \times G_y$, and at the end all the histograms of all the cells are concatenated to form a single histogram. The length of CSLBP descriptor is $G_x \times G_y \times 2^{\frac{N}{2}}$, that is quite often the double of SIFT descriptor. For our experiments, the values for $CSLBP_{N, R, T}$ are $CSLBP_{8,1,0.01}$, and the highest efficiency is obtained by keeping $G_x = 4$ and $G_y = 4$ making $CSLBP$ the length of 256.

Scale-Invariant Feature Transform: In this section we have briefly describe the SIFT descriptor and its computation methodology. It is basically the representation of gradient orientation histograms. For computation of SIFT descriptor, image is divided into patches, then the given patch P is divided into grids of $G_x \times G_y$. Then for each pixel in each cell, the gradient magnitude $g(x,y)$ (SIFT) descriptor is the representation of gradient orientation histograms, orientation and $\theta(x,y)$ are computed. After this computation, each gradient orientations are quantized into 8 different directions and histogram of this quantized orientations are computed. After this, each sample which is added to the histogram is weighted by their Gradient magnitude and Gaussian weight. A circular region or window which is approximately 1.5 times larger than that of scale of keypoint is taken (Lowe, 2004). The Gaussian weight is used to give more preference to the pixel which are more nearer to the centre. Finally, at the end all the cells are concatenated into a single vector. The maximum efficiency of SIFT can be achieved by keeping the $G_x = 4$ and $G_y = 4$. That is why the length of the SIFT descriptors is of 128 lengths ($8 \times 4 \times 4$).

For Gaussian weight, circular window with a σ that is 1.5 times that of the scale of keypoint is taken (Lowe, 2004). The Gaussian weight is used to give more preference to those pixels that are near to centre. Finally, gradient orientation histograms of all cells are concatenated to single vector. The maximum efficiency of SIFT is also obtained by keeping $G_x = 4$and $G_y = 4$. Therefore, the SIFT descriptor is of 128 lengths ($8 \times 4 \times 4$). The FER algorithm consists of few steps which needs to be completed for accurate classification of expressions. In the first step, we have to locate the faces from the images using some set of landmark points during face localization and face detection. Then these detected faces are further normalized geometrically for face registration. After this we have the step of feature extraction from the detected faces. In literature we have so many kinds of local and global descriptors. These features can be *geometric features* such as facial landmarks (Kobayashi, 1997), *appearance features* such as pixel intensities (Mohammadi, 2014) etc. The state of the art researchers like (Zhang, 2015; Zhang, 2014), join many kind of descriptors using multiple learning algorithms, but in our case it's not needed and we do not have to fuse multiple features using multiple kernel learning algorithms for better representation as it is very time consuming and memory inefficient. In our method we have extracted a single descriptor on multiple scales which gives equally robust features with memory and time efficiency. (Cohn *et al*., 2007; Ekman, 2002), used two different approaches for studying the facial behaviour: *Message-based approaches* and *Sign based approaches*. Both of these facial behaviour techniques are used to categorize facial expressions, Sign-based are used to describe facial actions/conjuration regardless of the actions meaning, whereas Message-based approaches categorize facial behaviours as the meaning of expressions. The Facial Action Coding System (FACS) is well known sign-based approach (Ekman, 2002). FACS used Active Units (AUs) for facial movements. Some of the AUs are given below. FACS is a tool developed by (Ekman, 2002) which describes every muscle activity using AUs and each AU is comprised of set of certain components of facial muscles movements. Many frameworks have been proposed for FER (Mohammadi, 2014), (Kumbhar *et al*., 2012; Londhe, 2012; Cheon, 2009). Gabor filters are used as features for FER (Kumbhar, 2012) and PCA is used for reduction of the dimensions. Facial expressions are categorized into seven categories and performance is evaluated on JAFFE database along with Convolution Neural Network. Additionally, Affine Moment invariants for features extraction are also used from the faces and then used in CNN for classification of the facial expressions (Londhe, 2012). The accuracy was 93.8% on the JAFFE database (Londhe, 2016). The differential-AAM features are also used with K-nearest neighbour sequence technique for classification with the accuracy of 86.4% (Cheon, 2009).

PROPOSED METHODOLOGY

In this section we have described the methodology for descriptors computation and classification of facial expressions. We have used the methodology for computation of descriptors from different images as proposed by (Baber, 2015). The descriptors are computed by fusing more information from the space around each keypoint patch using the two specific ways done by (Baber *et al*., 2015).

Preliminaries: For descriptor computation we have also used the Harris affine detectors (Mikolajczyk *et al*., 2014) for the detection of the keypoints. Each keypoint has an elliptical region consist of its scale, gradient angle and second moment matrix. The elliptical region is first mapped to circular area which is then standardized into 41 × 41 pixels in Cartesian framework (Mollahosseini *et al*., 2014; Gross, 2010). At the end the Cartesian grid is additionally isolated into 4 by 4 grid. and then used SIFT as a local descriptor for the description of the keypoints. Each set of sift descriptor is consisting of the following parameters which is represented as $q = (x, y, \theta, \sigma, P_q, d_q^l)$, where x and y are the coordinates, θ is considered to be a dominant orientation, σ is a scale for the patch of the descriptor or keypoint aka default scale, P is defined to be the 2D affine area of size 41 × 41 centred around q, and d^l is a descriptor vector around the keypoint and $l \in \{CSLBP, SIFT, PCA{-}SIFT...\}$. In this research, we have only focused on $l \in \{SIFT\}$ only.

Multi-scale Technique: In order to insert the data around each keypoint we increased the size of patch by multiple scales and then computed the SIFT descriptors on each scale and join them together to make the descriptor more invariant and robust.For each and every keypoint q in an image we have computed descriptors from P_q, where $P_q = P_q^1, P_q^2, P_q^3, P_q^1 ... P_q^N$, P_q^i is computed on various different scale with a predefined value ω, which can be gradually increased or decreased by some value N. In our experiments we have used SIFT on multi scale and we used different values of $N \in 1,2,3$. Patch of an image is estimated on each scale by different parameters i-e gradient angle and moment matrix (Mikolajczyk, 2004), after this the descriptor

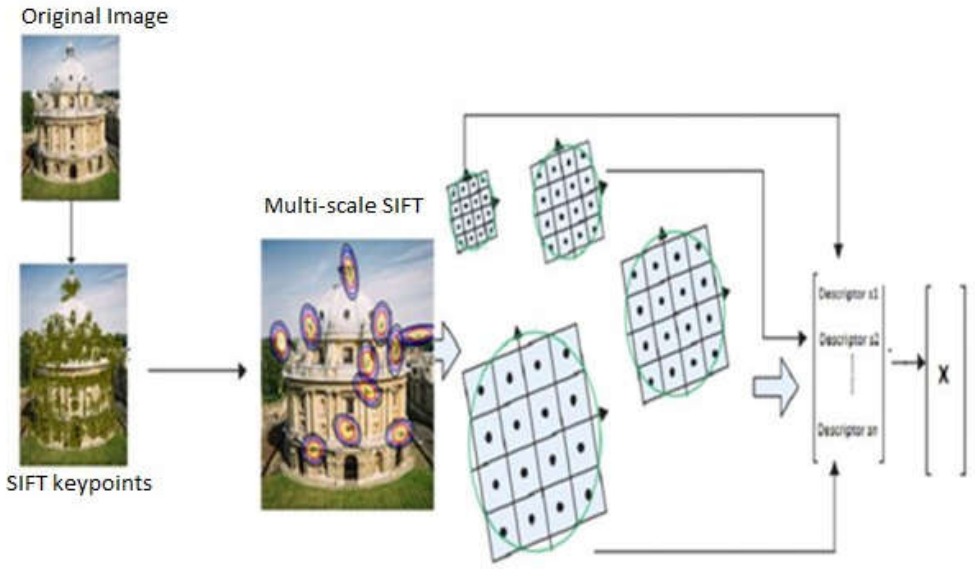

Fig. 1. Image classification using Multi-scale SIFT

Table 1. Facial Expression Annotated Images in Fer2013

Label	Number of images
Neutral	3501
Happy	7130
Sad	3128
Surprise	1439
Fear	1307
Disgust	702
Anger	2355

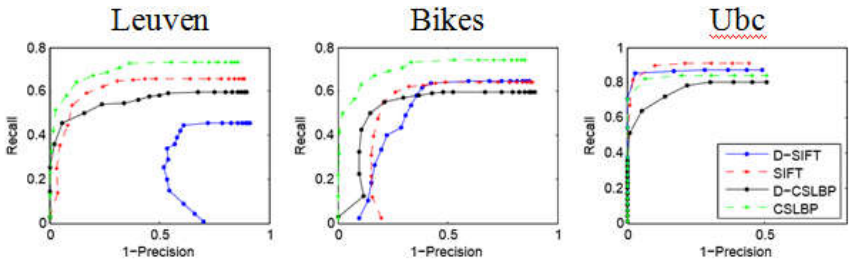

Fig. 2. Keypoint matching accuracy using double scale descriptors Reprinted from Baber *et al.* (2015)

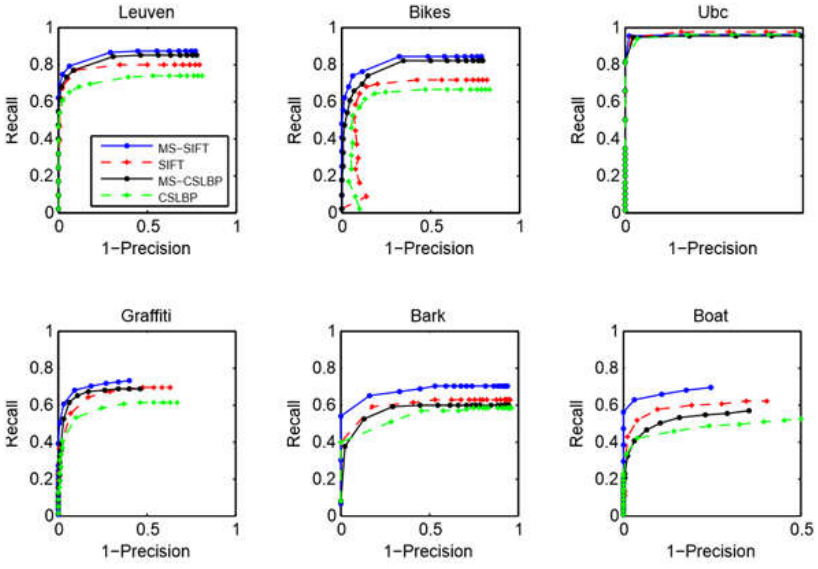

Fig. 3. Keypoint matching accuracy using multi-scale scale descriptors Reprinted from Baber *et al.* (2015)

of each patch on different scale is then computed and then all the vectors of different patch descriptors are concatenated to form a single vector. Each single image is represented by set of local features, i.e., on average there are 2000 to 3000 keypoints/image. Image representation by set of local feature make it infeasible for very large applications. To make image representable by single feature which should be robust and distinctive, bag of visual word is widely used. We have used the same configuration as suggested in our previous work (Baber, 2015).

Facial expression recognition: We have used Fer2013[1]dataset obtained by Kaggle[2]. Fer2013 is challenging dataset. Table I shows the number of images in Fer2013 dataset for each facial expression (Dhall *et al*., 2013). Each image in Fer2013 is 48 × 48 pixel (single channel). We extracted dense sift and dense multiscale-sift from each image and then quantized into visual word. To train visual words, we took 50K images and extracted 300 million features. for learning keeping clusters (visual word) approximately 1 million. We have used VLFEAT[3]library for training and learning. Since, the vocabulary size is too huge and classical K-mean cannot learn that huge clusters, therefore, we used Hierarchical K-mean clustering, as suggested by Zhu *et al*. (2012). We have used K-NN and SVM classifiers to recognize the facial expression of given face. There are 7-classes denoted by C = {Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger} which can be rewritten as C = {c_1, c_2..., cm}, respectively. K-NN can easily handle multi-class classification, though it is very slow at prediction phase. The SVM can easily handle only binary classification, for multi-class classification, there are two variations of SVM, 1) one-vs-one classification, 2) one-vs-all classification. In one-vs-all classification, the m binary classifiers are trained. For c_i, the positive samples are the samples of c_i and all other classes are considered as negative samples. In case of one-vs-one, there are $\frac{m(m-1)}{2}$ binary classifiers trained for m-way multi-class problem; each class receives the samples of the pair of remaining classes. In evaluation, voting schema is applied to decide the class of given face. We have used one-vs-all approach for facial expression recognition.

RESULTS

As stated in our previous work that the patch based descriptors show limited performance in case of very big corpus. To overcome the above mentioned problem, we can increase the information within the descriptors during computation. This can be done in several ways; either increase the patch scale so that it can capture more information around the keypoint, or compute the descriptors at multiple scales and concatenate at the end. Figure 2 shows the performance of descriptors when computed on double scales. It can be seen that it doesn't increase the performance. This experiment is done on VGG dataset which is widely used for keypoint matching accuracy. It has severe affine transformed images, the transformations include blurring, scale, illumination, rotation, and viewpoint change.

[1] https://www.kaggle.com/c/challenges-in-representation-learning-facialexpression-recognition-challenge/data
[2] https://www.kaggle.com
[3] https://www.vlfeat.org

The set of protocols used for this experiments can be seen in our previous work (Baber, 2015). Figure 3 shows the performance of multi-scale descriptors for keypoint matching experiment. It can be seen that multiscale descriptors, both SIFT and CSLBP, increase the performance of matching. In this paper, same technique is applied in classification task, facial expression recognition, and it outperforms consistently. It gives 58.3% accuracy whereas SIFT only gives 55.6% on Fer2013 dataset using K-NN classifier. In case of SVM, one-vs-all, it gives 57.1% with SIFT and 61.3% with multiscale SIFT. It can be seen in Table I that some facial categories have large number of instances which make the training data imbalance for some categories such as category '*Disgust*'.

Conclusion

In this paper, we have extended the multi-scale descriptors from image retrieval to image classification. For classification, we have used facial expression recognition problem. The Kaggle dataset, known as Fer2013, is used for experiments. Table I shows the number of images for each facial category. Multi-scale SIFT has improved the performance of descriptors in all cases: image-to-image matching, query based image retrieval, and image classification. For image classification, we have used K-NN and SVM, the performance is increased by 2.7% and 4.1%, respectively.

REFERENCES

Baber, J., Fida, E., Bakhtyar, M. and Ashraf, H. 2015."Making patch based descriptors more distinguishable and robust for image copy retrieval," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*. IEEE, pp. 1–8.

Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L. 2008. "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346– 359.

Calonder, M., Lepetit, V., Strecha, C. and Fua, P. 2010. "BRIEF: Binary Robust Independent Elementary Features," in *ECCV*.

Chang, E. Y., Wang, J. Z., Li, C. and Wiederhold, G. 1998. "RIME: A Replicated Image Detector for the World-Wide Web," in *Storage and Retrieval for Image and Video Databases*.

Cheon, Y. and Kim, D. 2009. "Natural facial expression recognition using differential-aam and manifold learning," *Pattern Recognition*, vol. 42, no. 7, pp. 1340–1350.

Chum, O., Philbin, J. and Zisserman, A. 2011. "Near duplicate image detection: Min-hash and tf-idf weighting," *Proc. BMVC, 2008*.

Cohn, J. F., Ambadar, Z. and Ekman, P. 2007. "Observer-based measurement of facial expression with the facial action coding system," *The handbook of emotion elicitation and assessment*, pp. 203–221.

Dhall, A., Goecke, R., Joshi, J., Wagner, M. and Gedeon, T. 2013. "Emotion recognition in the wild challenge 2013," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, pp. 509–516.

Ekman, P. 2002. "Facial action coding system (facs)," *A human face*.

Ekman, P. and Friesen, W. V. 1971. "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124.

Gross, R., Matthews, I., Cohn, J., Kanade, T. and Baker, S. 2010. "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813.

Heikkila, M. and Pietik¨ ainen, M. 2006. "A texture-based method for modeling¨ the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662.

Heikkila, M., Pietikainen, M. and Schmid, C. 2009. "Description of interest regions with local binary patterns," *Pattern Recognition*.

Ke Y. and Sukthankar, R. 2004. "Pca-sift: A more distinctive representation for local image descriptors," in *Proc. CVPR*, pp. 511–517.

Kim, C. 2003. "Content-based image copy detection," *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 169–184.

Kobayashi, H. and Hara, F.1997. "Facial interaction between animated 3d face robot and human beings," in *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., IEEE International Conference on*, vol. 4. IEEE, 1997, pp. 3732–3737.

Kumbhar, M., Jadhav, A. and Patil, M. 2012. "Facial expression recognition based on image feature," *International Journal of Computer and Communication Engineering*, vol. 1, no. 2, p. 117.

Londhe, R. and Pawar, V. 2012. "Facial expression recognition based on affine moment invariants," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 6, p. 388.

Lost in quantization: Improving particular object retrieval in large scale image databases," *Proc. CVPR, 2008*, 2011.

Lowe, D. G. 2004. "Distinctive image features from scale-invariant keypoints," *IJCV*.

Lyons, M., Akamatsu, S., Kamachi, M. and Gyoba, J. 1998. "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, pp. 200–205.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. Cohn, and J. F. 2013. "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160.

Mayer, C., Eggers, M. and Radig, B. 2014. "Cross-database evaluation for facial expression recognition," *Pattern recognition and image analysis*, vol. 24, no. 1, pp. 124–132.

Mikolajczyk, K. and Schmid, C. 2004."Scale & affine invariant interest point detectors," *International journal of computer vision*, vol. 60, no. 1, pp. 63–86.

Mikolajczyk, K. and Schmid, C. 2005. "A performance evaluation of local descriptors," *IEEE Transactions on PAMI*.

Mohammadi, M., Fatemizadeh, E. and Mahoor, M. H. 2014. "Pca-based dictionary building for accurate facial expression recognition via sparse representation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1082–1092.

Mollahosseini, A., Graitzer, G., Borts, E., Conyers, S., Voyles, R. M. Cole, R.and Mahoor, M. H. 2014. "Expressionbot: An emotive lifelike robotic face for face-to-face communication," in *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*. IEEE, , pp. 1098–1103.

Nister, D. and Stew´ enius, H. 2006. "Scalable recognition with a vocabulary´ tree," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

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

Philbin, J., Chum, O., Isard, M., Sivic, J. and Zisserman, A. 2007. "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition*.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. 2011. "Lost in quantization: Improving particular object retrieval in large scale image databases," *Proc. CVPR, 2008*, 2011.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G. 2011. "Orb: An efficient alternative to sift or surf," in *ICCV*, 11/2011 2011.

Wan, Y., Yuan, Q., Ji, S., He, L. and Wang, Y. 2008. "A survey of the image copy detection," sep. pp. 738 –743.

Wu, Z., Ke, Q., Isard, M. and Sun, J. 2009. "Bundling features for large scale partial-duplicate web image search," in *Computer Vision and Pattern Recognition*.

Xu, Z., Ling, H., Zou, F., Lu, Z. and Li, P. 2011. "A novel image copy detection scheme based on the local multi-resolution histogram descriptor," *Multimedia Tools and Applications*.

Zhang, X., Mahoor, M. H. and Mavadati, S. M. 2015. "Facial expression recognition using {l} _ {p} l p-norm mkl multiclass-svm," *Machine Vision and Applications*, vol. 26, no. 4, pp. 467–483.

Zhang, X., Mollahosseini, A., Boucher, E., Voyles, R. M., Nielsen, R., Mahoor M. H. *et al.*2014. "ebear: An expressive bear-like robot," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, pp. 969–974.

Zhou, W., Li, H., Lu, Y. and Tian, Q. 2011. "Large scale partial-duplicate image retrieval with bi-space quantization and geometric consistency," *Proc. ICASSP, 2010*.

Zhou, W., Lu, Y., Li, H., Song, Y. and Tian, Q. 2010. "Spatial coding for large scale partial-duplicate web image search," in *Proceedings of the international conference on Multimedia*.

Zhu, C.Z. and Satoh, S. 2012. "Large vocabulary quantization for searching instances from videos," in *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, ser. ICMR '12. ACM, pp.52:1–52:8.

****