LBP-HF and HOG Descriptors with Voting Mechanism for Emotion Recognition in the Wild

Liang Li
School of Automation, Huazhong University of Science and Technology
Wuhan, 430074, China
LL_2016@foxmail.com

Hao Wang
School of Automation, Huazhong University of Science and Technology
Wuhan, 430074, China
hao.wang_1996@foxmail.com

Weiran Zhang
School of Automation, Huazhong University of Science and Technology
Wuhan, 430074, China
weiran0720@foxmail.com

ABSTRACT
In this paper, we present the method for our submission to the Emotion Recognition in the Wild Challenge (EmotiW 2016). We only take the sub-challenge of video-based emotion recognition. The challenge is to automatically classify the emotions acted by human subjects in movie clips under real-world environment. In our method, each movie clip is cut into lots of pictures from which we can get face images. After pre-processing, we zoom the face images to a same size of 128 × 128 pixels and 64 × 64 pixels. For features, we utilize HOG and LBP-HF descriptors. As for classification, we choose two different classifiers, SVM and Random Forest, to investigate the classification accuracy, and for further comparison. We combine LBP-HF and SVM, HOG and SVM, HOG and Random Forest as three methods to recognize the expression of the faces shown in images. In the end, we use these three methods to vote for a relatively accurate result. The final recognition accuracy achieved 48.06% on test set, 7.59% over the challenge baseline 40.47%.

Categories and Subject Descriptors
I.5.4 [**Pattern Recognition**]: Applications—computer vision, signal processing; I.4.m [**Image Processing and Computer Vision**]: Miscellaneous

General Terms
Algorithms, Experimentation.

Keywords
Emotion Recognition; HOG; LBP-HF; Random Forest; SVM; EmotiW 2016 Challenge

1. INTRODUCTION
Automatic emotion recognition in the wild is a popular but challenging problem. There are many factors in natural scenarios that create obstacles for emotion recognition, such as light intensity, angles of faces and so on. As a reference, in EmotiW2015, the champions reached only 53.8% classification accuracy [11]. In EmotiW 2016, the video-based emotion recognition challenge contains audio-video short clips labelled using a semi-automatic approach defined in. The task of EmotiW 2016, which is continuation from the challenge in EmotiW 2013 is to assign a single emotion label to the video clip from the six universal emotions (Anger, Disgust, Fear, Happiness, Sad & Surprise) and Neutral. However, the major change in this sub-challenge as compared to the earlier years is the introduction of reality TV data in the Test set.

With recent advances in deep learning and pattern recognition, we have too many kinds of features and classifiers to choose for an automatic emotion recognition project. In this work, we choose HOG, LBP-HF as the features, SVM and Random Forest as the classifiers. In order to improve the accuracy of our method, we combine LBP-HF and SVM, HOG and SVM, HOG and Random Forest as three methods to recognize the expression of the faces shown in movies. Finally, we use these three methods to vote for a relatively accurate result. We will detail the whole procedure in the following sections.

2. THE PROPOSED METHOD
2.1 Data Pre-processing
At first, we find that the videos provided by the committee become deformed after being decoded by our ffmpeg decoder, so we resize the images captured from videos to let them back to normal size. Secondly, we zoom the images to accelerate the detection speed, and then, we convert RGB images into gray images. After that, we use the face-detector based on histogram equalization, to get the face images sized 128 × 128 and 64 × 64 pixels (An example is shown in Fig.1).

Figure 1. An example of face detection

2.2 Image Feature
Apparently, one key step of this task is to extract features that are capable to describe the facial expression in the images accurately. In our work, we utilize the LBP-HF and HOG descriptors to represent the image features.

2.2.1 HOG
Histogram of Oriented Gradient (HOG) first appeared in SIFT, and it has the strong ability of image-description. We depict the process below:

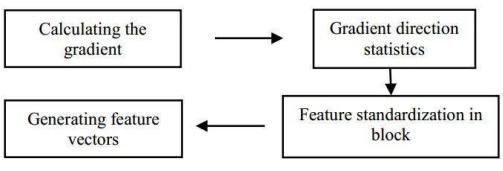

Figure2. The processing procedure of HOG

Conference'10, Month 1–2, 2010, City, State, Country.

As is known that the HOG feature for pedestrian detection has achieved impressive results, so we try to use HOG for recognition of facial expression. Facial expression recognition mainly includes feature extraction and classification.

The calculation process is that:

1. We use face images sized 128 × 128 and 64 × 64 without Gamma illumination pre-process, which are gray scale. The descriptor's block is 16 × 16, and its cell is 8 × 8.

Figure 3. An image processed by HOG

2. Then we use gradient operator to calculate horizontal and vertical gradient for each pixel, which are set as $H(x, y)$ and $V(x, y)$ respectively.
3. Calculate gradient directions and ranges for each pixel, and the calculating formulas are:

$$\theta(x, y) = tan^{-1}\left[\frac{V(x,y)}{H(x,y)}\right] \quad (1)$$

$$m(x, y) = [H(x, y)^2 + V(x, y)^2]^{1/2} \quad (2)$$

where $\theta(x, y)$ is gradient directions of pixel (x, y), and $m(x, y)$ is the gradient range of pixel (x, y).

4. $\theta(x, y)$ ranges from -90 to 90 degrees, we divide it into nine parts. Next, each cell vote by gradient direction with a weight of $m(x, y)$, so that each cell gets a nine-dimensional vector. String the vectors of four cells in the same block together, we can get the 36-dimensional vector of a block, At last, stringing all the vectors together, we can get the HOG feature vector of a detection window.

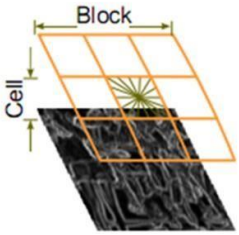

Figure 4. Computing HOG descriptors

For further process, firstly, we want to remove redundant information and the singularity of scatter matrix between classes, and global HOG feature dimensionality is reduced with principal component analysis (PCA). Next, we can get local feature vectors. However, after these vectors being trained by SVM, we find that the results rather terrible. We suppose that because the descriptor's dimension is too high, when processed by PCA, the error in the test may be amplified. Doing cluster analysis before PCA is also considered, but the results show that the gap among different emotions is further widened, and the emotion calculated almost does not change in different cases. Because of the bad results, we abandon the processing methods above.

2.2.2 LBP-HF

LBP (Local Binary Patterns) feature is one of the most popular representation schemes for face recognition. It is first described in 1994 ([1], [2]). In 2004, Ahonen et al ([3]) applied LBP for face recognition. Besides, it has also been used for facial expression recognition ([4]). The Basic LBP operator is firstly proposed by Ojala [2] et al. The pixels in the neighborhood are converted to binary code 0 or 1 by taking the gray value of the center pixel as threshold. All these codes form an ordered pattern of the center pixel.

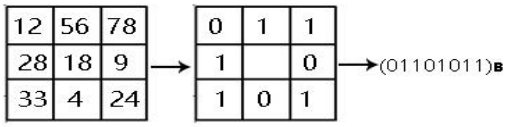

Figure 5. An example of basic LBP operator

Such coding is denoted as $LBP_{P,R}$ according to

$$LBP_{P,R} = \sum_{i=1}^{P-1} s(g_i - g_c) \cdot 2^i \quad (3)$$

$$s = \begin{cases} 0, & x > 0 \\ 1, & x \le 0 \end{cases} \quad (4)$$

where R represents the scale of the radius of neighborhoods, P represents the number of sampling points, g_i is the gray value of a pixel in the neighborhood and g_c is the gray value of the center pixel. When there are at most two instances of 0 to 1 in binary code of the center pixel, it is called uniform pattern [5]. In the computation of the uniform LBP histogram, every uniform pattern has a separate bin and all non-uniform patterns are assigned to a single bin. Given the number of the sampling points P, the number of the uniform patterns is $P^2 - P + 2$. In the challenge, we adopt LBP-HF (Local Binary Pattern Histogram Fourier) features that are proposed on the base of uniform LBP [6]. For 8 sampling points, there are 58 possible uniform patterns.

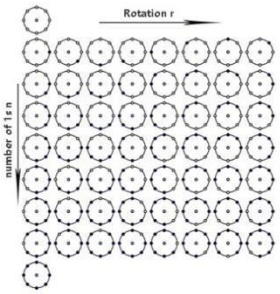

Figure 6. The 58 different uniform patterns in (8, R) neighborhood

Denote the specific uniform pattern by $U_P(n, r)$. In the formula, n is the number of 1-bits in the pattern and r is the rotation of the pattern. Consider the uniform LBP histograms $h_I\big(U_P(n, r)\big)$. The histogram value h_I at bin $U_P(n, r)$ is the number of occurrences of uniform pattern $U_P(n, r)$ in an image. If the image is rotated by an angle $\alpha = a \times 360°/P$, this rotation of the input image causes a cyclic shift in the histogram along each of the rows:

$$h_I^{\alpha}\big(U_P(n, r + a)\big) = h_I\big(U_P(n, r)\big) \quad (5)$$

Based on the property, a class of features that are invariant to the rotation of the input image are proposed [6]. Such features computed along the input histogram rows are invariant to cyclic shifts. These features can be constructed using Discrete Fourier Transform [6]. Let H (n,·) be the DFT of nth row of the histogram $h_I\big(U_P(n, r)\big)$, we have:

$$H(n,u) = \sum_{r=0}^{P-1} h_I\big(U_P(n,r)\big)e^{-i2\pi ur/P} \quad (6)$$

It can be proved that with any $1 \le n_1, n_2 \le P-1$ and $0 \le u \le P-1$, the features

$$LBP^{u2} - HF(n_1,n_2,u) = H(n_1,u)\overline{H(n_2,u)} \quad (7)$$

where $\overline{H(n_2,u)}$ denotes the complex conjugate of $H(n_2,u)$, are invariant to cyclic shifts of the rows of $h_I\big(U_P(n,r)\big)$ and the rotations of the input image[6]. The Fourier magnitude spectrum

$$|H(n,u)| = \sqrt{H(n,u)\overline{H(n,u)}} \quad (8)$$

can be consider a special case of these features.

In the challenge, the image segmented from the videos contains 128 × 128 pixels. It is divided into 64 blocks (8 × 8 pixels) and 40 blocks are utilized for extracting LBP-HF features.

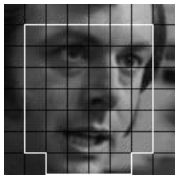

Figure 7. An image segmented from the video which are divided into 64 blocks and 40 blocks are used for extracting features

In each block, there are 64 pixels. First calculate the uniform LBP of every pixel and form a uniform LBP histogram given $P = 8$. Then use Discrete Fourier Transform so that the features can be saved in a 37-dimensional vector. Therefore, the features of the whole image can be saved in a 1481-dimensional vector including a label dimension. Further , we utilize SVM as the classifier on the training set.

2.3 Classifiers

2.3.1 SVM

SVM (Support Vector Machine) is a generally used classifier in pattern recognition. In the challenge, we choose C_SVC as the SVM type and POLY as the kernel type. We call the function train_auto so that the parameters can be optimized in the process of training.

2.3.2 Random Forest

Random Forest is an important ensemble learning method and it is widely used in data classification. Random Forest is a model of integrated learning that is based on decision trees. Voted by all of the decision trees in the Random Forest, the final result we get is able to avoid overfitting caused by decision trees. Besides, Random Forest is strong in robustness and practicability.

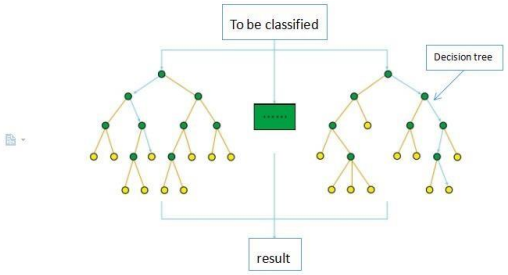

Figure 8. The frame of Random Forest

2.3.2.1 Decision Trees

The decision trees trained by the sample space is actually a division of this training sample space. The subset of the sample space corresponding to one class is considered a hypercube union, which is some of the rectangles under the two-dimensional case and boxes and boxes under the three-dimensional case.

Since the corresponding subset of the sample space of different classes is disjoint, these hypercube unions are disjoint, as well. All of the hypercube unions make up the whole sample space.

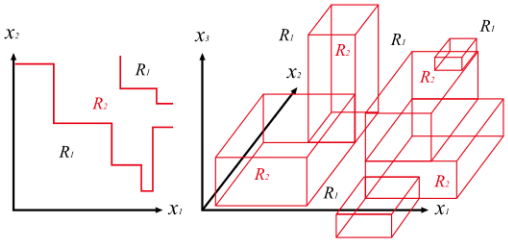

Figure 9. How the decision trees classify

2.3.2.2 Random Forest

Random Forest consists of K decision trees that play the role of basic classifiers, and has higher training speed than SVM. By ensemble learning, the decision trees make up a combined classifier. All of the decision trees in the Random Forest vote the results for the classification.

$\{\theta_k, k = 1,2,3,\dots,K\}$ is the number of decision trees. It is random, and it depends on two rules:

(1) Bagging: We get K training sets that have the same size as the original samples by sampling with replacement. $\{T_k, k = 1,2,3,\dots,K\}$ (About 37% samples of the original samples are not chosen each time).

(2) When dividing each node of the decision trees, we select part of the features. (Generally $\log_2 M + 1$ chosen features, where M is the number of features) Next, we select the most effective features from the part of features we choose to divide the node into two parts.

3. EXPERIMENTS

3.1 EmotiW2016Challenge

The Emotion Recognition in the Wild Challenge (EmotiW 2016) consists of some movie clips based emotion classification task that mimics real-world conditions. The major change in this sub-challenge as compared to the earlier years is the introduction of reality TV data in the test set which has showed close-to-real-world conditions. The task is to classify an audio-video clip into one of the seven emotion categories.

	Angry	Disgust	Happy	Fear	Neutral	Sad	Surprise
Train	129	72	145	80	144	115	73
Val	64	40	61	46	63	61	46

Table 1. The number of samples for each emotion category in the training and validation sets

3.2 Parameter Setting

At first, we simply use the aligned face images provided by EmotiW 2016 organizers. However, after a few failures, we realized that we had better use the images cut by the face detector we use. All of the images detected by our face detector are resized to 128×128 pixels and 64×64 pixels. Two kinds of image features are employed on the aligned faces: HOG and LBP-HF.

For HOG, we divide each image into 15×15 = 225 overlapping blocks with the size of 128 × 128 pixels (i.e. the strides are 8 pixels in both horizontal and vertical directions). The descriptor is applied by computing histograms of oriented gradient on 2 × 2 cells in each block, and the orientations are quantized into 9 bins, which results in 2 × 2 × 9 = 36 dimensions for each block and 36 × 225 = 8100 dimensions for the whole image.

3.3 Results Comparisons

With the purpose of increasing diversity, we use three methods to vote for the result, which are Submission No.2, 3, 4. As is shown in Table 2, our method reaches an overall accuracy of 48.06%. To be specific, angry, happy, and neutral expressions can be recognized at a relatively high accuracy, while it is difficult for our method to identify fear and disgust emotions. In addition, our classifier tends to identify other emotions as angry, fear, and sad.

Submission No.	Validation (%)	Test (%)	Method
1	44.76	48.06	Voted by method 3rd, 4th, 5th.
2	40.71	44.01	SVM trained with HOG. (face scale: 128 × 128 pixels)
3	39.40	44.18	SVM trained with HOG. (face scale: 64 × 64 pixels)
4	38.58	39.29	Random Forest trained with HOG. (face scale: 128 × 128 pixels)
5	37.03	37.94	SVM trained with LBP, while extracting 40 blocks among 64 blocks in an image. (face scale: 128 × 128 pixels)

Table 2. Our submissions on the AFEW 6.0 validation and test sets

Source	Results							Accuracy
	angry	disgust	fear	happy	neutral	sad	surprise	
angry(83)	**42**	1	2	7	18	10	3	50.60%
disgust(36)	3	**6**	3	8	8	8	0	16.67%
fear(66)	10	1	**15**	4	21	11	4	22.73%
happy(135)	19	0	3	**87**	17	7	2	64.44%
neutral(174)	20	1	14	13	**94**	25	7	54.02%
sad(71)	9	0	5	4	20	**29**	4	40.85%
surprise(28)	6	0	3	2	5	0	**12**	42.86%
total(593)	**48.06%**							
	38.53%	66.67%	35.71%	69.60%	51.37%	32.22%	37.50%	

Table 3. Our best submissions

4. CONCLUSIONS

In this paper, we propose a method for facial expression recognition in the wild. The database provided by the committee is collected from different movies. Each emotion video clip is regarded as a static image, and then we use HOG and LBP-HF to calculate feature vectors of each image. At last, SVM and Random-Forest classifiers are trained based on these feature vectors. The method does not achieve very outstanding results on validation and test data. As for the cause, we guess one decisive factor is that we fail to implement face alignment, so that our method takes into consideration the angles of faces, which are not critical and an interfering source for the task. Besides, we did not spend enough time on parameter tuning, which led to either overfitting or underfitting. In addition, CNN and its variants are highly capable for image recognition, the ability of which greatly surpasses that of some algorithms that have been proposed many years ago. In the future, we will try to add face alignment into the task and search for other related algorithms that are more effective to improve the performance.

5. REFERENCES

[1] T. Ojala, M. Pietikäinen , and D. Harwood (1994), Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR 1994), vol. 1, 582-585.

[2] T. Ojala, M. Pietikäinen, D. Harwood, A Comparative Study of Texture Measures with Classification Based on Featured Distributions, Pattern Recognition, 1996, 29(1), 51-59.

[3] T. Ahonen, A. Hadid, and M. Pietikäinen, Face Recognition with Local Binary Patterns, Proc. 8th ECCV, 2004, 3021, 469–481.

[4] Shan, C., Gong, S., and McOwan, P.W. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27, 6, 803-816.

[5] Ojala, T. Pietikäinen, M. Mäenpää, T. Multire solution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002).

[6] T. Ahonen, J. Matas, C. He, M. Pietikäinen. Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features. Image Analysis, 2015, 5575(4):61-70.

[7] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[8] Quinlan J R. Induction of decision trees [J]. Machine Learning, 1986, 1, (1): 81-106.

[9] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24 (2): 123-140.

[10] Richard O D. 模式分类 [M]. 李宏东，姚天翔等译. 第二版. 北京：机械工业出版社, 2003: 32.

[11] A. Yao, J. Shao, N. Ma, Y. Chen, Capturing AU-Aware Facial Features and Their Latent Relations for Emotion Recognition in the Wild. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. Pages 481-458.