FACE DETECTION BASED ON MULTI-SCALE ENHANCED LOCAL TEXTURE FEATURE SETS

Zhe Wei[†]*, Yuan Dong*[†‡]*, Feng Zhao*[‡]*, Hongliang Bai*[‡]

[†]Beijing University of Posts and Telecommunications,100876, P.R.China
[‡]France Telecom Research & Development - Beijing, 100190, P.R.China
weizhebupt@gmail.com,yuandong@bupt.edu.cn
{feng.zhao,hongliang.bai}@orange.com

ABSTRACT

This paper presents a distinctive rectangle feature Multi-Scale Local Ternary Patterns (MS-LTP) for face detection. The MS-LTP is a generalization of the Local Ternary Patterns (LTP) [1] and is able to capture larger scale structures of faces. It's less sensitive to noise and more discriminative that can reduce the number of weak classifiers for the AdaBoost learning algorithm to construct a strong face/non-face classifier. The size of the MS-LTP feature set is also medium for the AdaBoost learning algorithm to select a proper set of features. Our experimental results on the CMU-MIT frontal face test set show that the MS-LTP outperforms Haar, Local Binary Patterns (LBP) under noisy conditions and the MS-LTP based face detector works more rapidly.

Index Terms— MS-LTP, AdaBoost, face detection

1. INTRODUCTION

Face detection has become one of the most important research topics for a wide range of applications such as natural human-computer interaction (HCI), video surveillance and automatic face recognition [2]. Although the face detection task is trivial for human beings, it is a great challenge for computers due to many variations in orientation (including pitch, yaw and roll), facial expression, lighting conditions, etc.

Hundreds of approaches to face detection have been reported and have made significant progress to face detection in the past decades [3]. In particular, the seminal work by Viola and Jones [4] has made face detection practically feasible in real world applications. Their work contributed three great ideas: Haar features with integral image for rapid feature extraction, a modified AdaBoost learning algorithm boosting weak Haar classifiers to a powerful detector and the attentional cascade to speed up the detection. Since this breakthrough, substantial improvement has been made to this field in three aspects by recent research.

The first improvement is the selection of features. A set of rotated Haar-like features was introduced in [5]. These features enriched the origin Haar-like feature set and can reduce the false alarm rate on average by 10%. The joint Haar-like features based on co-occurrence of multiple Haar-like features were proposed in [6]. This feature captured the structural similarities within the face class, which made it possible to construct an effective classifier. LBP was introduced as a descriptor for face detection in [7][8] since the LBP is robust to illumination variations with low computational cost. The sparse features in granular based on heuristic search was presented in [9] for multi-view face detection.

The second improvement is the development of learning algorithms. RealBoost [10] substituted the Discrete AdaBoost by using confidence-rated predictions and GentleBoost [11] was applied when the real world data suffers from heavy noise [12]. FloatBoost Learning [13] was also proposed to select fewer but more discriminative features for face detection.

The third is the structures of the face detector. Multiple detectors are connected for multi-view faces that have widened the application of face detection. A parallel cascade [14] is used for more accurate detection. And a coarse-to-fine search pyramid [15] is used to get better balance between accuracy and efficiency.

In this paper, we will focus on the feature level for face detection. The main advantages of MS-LTP are: First, it's less sensitive to noise. Second, the MS-LTP is an extension of LTP that is able to capture larger scale facial structures. Third, the MS-LTP feature can be used to build more discriminative weak classifiers. So fewer weak classifiers are need to construct a face detector that makes the face detector more rapid.

The remaining parts of this paper are organized as follows. The MS-LTP features is introduced in Section 2. Section 3 describes the learning algorithm for MS-LTP based face detection. Section 4 presents the experimental results and Section 5 comes to the conclusion and future work.

2. MULTI-SCALE LOCAL TERNARY PATTERN FEATURES

LBP features only consider the intensity of the central pixel and its neighborhood so they are robust to monotonic changes

of illumination by design. However since they threshold at exactly the value of the central pixel, they tend to be sensitive to noise, especially in near-uniform image regions. To improve the robustness of LBP, LTP was extended from LBP. The main idea is to define a user-specified threshold $t(t > 0)$ to make the thresholding more tolerable to noise at the expense of slightly weakening the robustness to monotonic changes of illumination. At a given pixel position (x, y), the decimal form of the LTP code can be expressed as follows:

$$LTP(x,y) = \sum_{n=0}^{7} o(i_n - i_c)3^n \tag{1}$$

where i_c corresponds to the grey value of the center pixel (x, y) and i_n are grey values of the 8 surrounding pixels. Function $o(x)$ is defined as:

$$o(x) = \begin{cases} 0 & if \quad x \le -t \\ 1 & if \quad |x| < t \\ 2 & if \quad x \ge t \end{cases} \tag{2}$$

The LTP encoding procedure is illustrated in Fig 1. Here the user-specified t is set to 5 to make a tolerable interval.

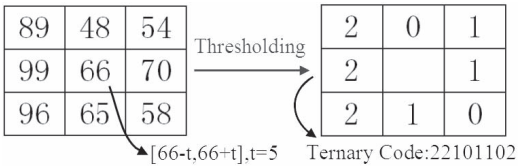

Fig. 1. Illustration of LTP operator. Here the operator thresholds at two value 61 and 71 making a interval.

The LTP features are defined for a 3×3 neighborhoods so can not be applied for other scales and this results in two limitations. One is in a common 20×20 window, only 324 (omitting the boundary pixels) LTP features can be provided for the AdaBoost learning algorithm to select a proper set of features while the number of Haar-like features is more than 100 thousand. When the candidate features are few, it may be insufficient for the AdaBoost to learn a strong classifier. Another limitation is the size of 3×3 neighborhood. So the LTP is not able to capture larger scale structure that may be dominant features of facial structures.

Inspired by the work in [8], we further improved LTP to a multi-scale operator to capture larger scale facial structures for face detection. The encoding procedure of MS-LTP is similar to LTP. The MS-LTP thresholds at the intensity sum of pixels in a block while the LTP at a single pixel. The MS-LTP at block (b_l, b_m) of scale s can be expressed as:

$$MS - LTP(b_l^s, b_m^s) = \sum_{n=0}^{7} o(b_n^s - b_c^s)3^n \tag{3}$$

where b_c corresponds to the center block (b_l^s, b_m^s), b_n^s to the 8 surrounding blocks. The b_n^s are sums of gray value of pixels in the 8 surrounding block at scale s. Scale s can be expressed by width times height of one block and Fig 2 shows some of MS-LTP features at different scales.

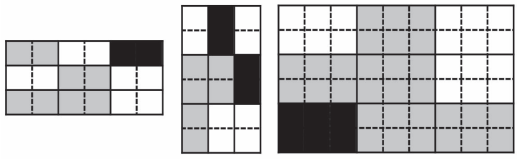

Fig. 2. The MS-LTP feature at different scales. Left:$s = 2\times1$. Middle: $s = 1 \times 2$. Right: $s = 3 \times 2$.

As the block size changes, the t should be multiplied by the magnitude of scale s to make the t fit for other scales. Fig 3 shows the MS-LTP encodes in a $s = 2 \times 2$ block with a new t value set to $t \times 4 = 20$. The MS-LTP features can also be calculated through the integral image to make a rapid computation. Compared to LBP, the LTP is encoded by a $3^8 = 6561$ value codes that the number of codes are much greater than the $2^8 = 256$ for LBP. This will provide more information for building weak classifiers. After the extension of multi-scale of LTP, in a 20×20 window the total number of MS-LTP features increases from 324 to 3969. And this can provide much more candidate features.

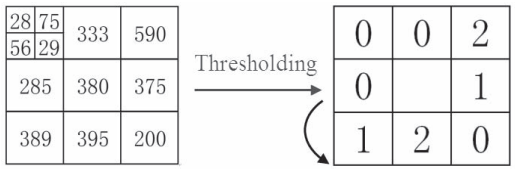

Fig. 3. The MS-LTP encodes by block intensities. The up-left block contains 4 pixels and the sum of gray value is 178. Since scale of this block is 4, the t is 20 so that the ternary code is 01201200.

3. LEARNING METHOD FOR MS-LTP BASED FACE DETECTION

Our motivation of this work is to make the face detection more robust to noise while inheriting most of the robustness to illumination variations from LBP and reduce the detection time by a more distinctive feature set. The structure of our detector basically follows the Viola-Jones object detection framework [4]. The main difference between ours and Violas is in the feature level, that is, the MS-LTP features rather than the Haar-like features.

The total number of MS-LTP features is 3969 which is much smaller than the Haar-like features. However the MS-LTP feature set contains some redundant information. To

eliminate redundant information the GentleBoost learning method [11] is used to select significant features and to construct a face/non-face classifier. To control the complexity of weak classifiers, each weak classifier is constructed by one of the MS-LTP features. Since the value of the MS-LTP features is non-metric, it is impossible to use threshold-based function as weak classifier. Decision trees are adopted as the structure of the weak classifier that will have at most 6561 output leaves. The weak classifier is defined as:

$$c_k(x) = \frac{\sum_i w_i y_i \delta(x_i^k = j)}{\sum_i w_i \delta(x_i^k = j)} \tag{4}$$

where x_i^k is the kth MS-LTP pattern of the ith training sample x_i and $j \in (1, \ldots, 6561)$ is the jth pattern of the MS-LTP. w_i and y_i are the weight and the label of the ith training sample. In each round t of the GentleBoost, one weak classifier $f_t(x)$ is chosen to minimize the weighted squared error:

$$f_t(x) = \arg\min_k \sum_i w_i (y_i - c_k(x_i))^2 \tag{5}$$

It deserves to be noted that more output leaves will endow more distinctive power to the weak classifier so that sometimes this may result in over fitting.

4. EXPERIMENTAL RESULTS

The CMU-MIT frontal face test set [16] is used to evaluate the performance of the proposed method. This set consists of 130 greyscale images with 511 labeled frontal faces. However the original labels are not rectangles of faces so we use the x and y coordinates of two eyes to form corresponding rectangles by the rules illustrated in Fig 4. Correct detection must follow the two rules [12]:

- The Euclidian distance between the center of a detected face and ground truth must be less than 30% of the width of the ground truth rectangle.
- The width of the detected face must be within 50% of the width of the ground truth rectangle.

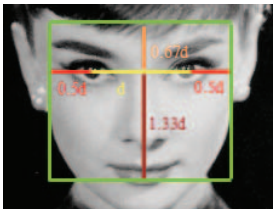

Fig. 4. Rules to generate rectangle from centers of two eyes. d is defined to be the distance between two eyes. Horizontal, produce lines from two eyes by $0.5d$. Vertically, produce lines up(down) by $0.67d(1.33d)$. Finally generates a $2d \times 2d$ facial rectangle region for evaluation.

Other detections are considered to be false alarms.

For the training procedure, about 40,000 frontal face images are collected and aligned by two labeled eyes. Then the original face images are derived to 120,000 face images by random rotation $\pm 15°$, random scaling $\pm 10\%$, random mirroring and random shifting $\pm 5\%$. Non-face images are randomly collected from the internet without face.

To make a full comparision, five kinds of features sets are included to evaluate their performances for face detection:

1. Haar: This Haar feature set contains the original 4 kinds of prototype Haar feature proposed in [4] and the extended set proposed in [5].
2. LBP: The LBP feature in 3×3 neighborhood.
3. LTP: The LTP feature in 3×3 neighborhood.
4. MB-LBP: The extended LBP feature sets based on multi-block proposed in [8].
5. MS-LTP: The proposed LTP feature sets.

To reduce disturbance from other factors the same training parameters (including training databases, training strategies, number of layers) are set to all five feature sets. In our experiment the user-specified t of LTP and MS-LTP are set to 1.

A ROC curve showing the performance of five kinds of feature sets is shown in Fig 5. The MS-LTP is superior to Haar, LBP, LTP but slightly backward to MB-LBP. To evaluate the resistance of noise, Gauss noise with parameters $\mu = 0, \sigma = 0.05$ was added to the face images in the CMU-MIT test set to construct a noisy test set and the results are shown in Fig 6. When noise added, the MS-LTP performs the best of all the five feature sets demonstrating its robustness to noise. The LBP feature performs the worst in this experiment, however, the MB-LBP still outperforms Haar and LTP under noisy condition. This may suggest that the resistance to noise of MS-LTP is not only because the t makes it tolerable to noise but also benefit from captures of larger scale facial structures.

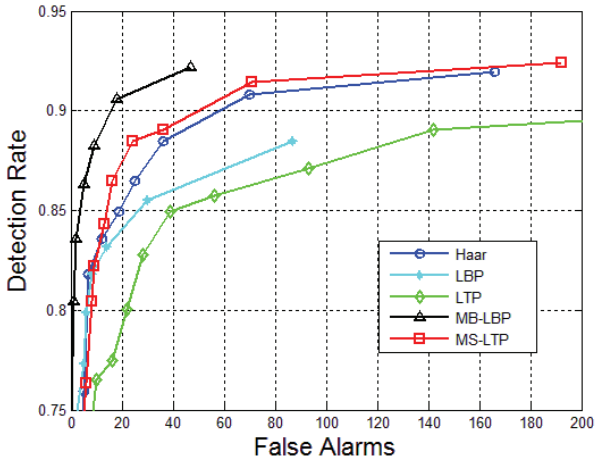

Fig. 5. ROC curve for five feature sets on CMU-MIT test set.

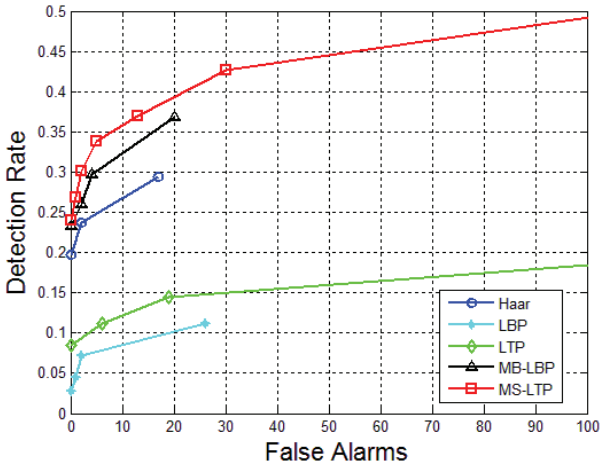

Fig. 6. ROC curve for five feature sets on CMU-MIT test set with Gauss noise $N(0, 0.05)$.

Because the output of the AdaBoost strong classifier depends on a linear combination of responses of weak classifiers so obtaining the responses of weak classifiers will cost most of the time in face detection. In our method each weak classifier is built by one feature, so fewer weak classifiers may mean consuming fewer time. With a more distinctive representation of the MS-LTP feature set, the MS-LTP based detector only needs 290 weak classifiers that is much fewer than MB-LBP while it achieves a similar performance. And this makes the speed of the MS-LTP detector the most rapid. Details are listed in Table 1. The tests run on an Intel® XEON® 2.0GHz CPU. They are repeated 10 times and the results are averaged.

Table 1. Time factors of five kinds of feature sets.

(a) Numbers of all weak classifiers

Haar	LBP	LTP	MB-LBP	MS-LTP
732	1160	390	702	**290**

(b) Detection time on CMU-MIT face test set in (s)

Haar	LBP	LTP	MB-LBP	MS-LTP
62.90	115.45	89.44	67.49	**45.49**

5. CONCLUSIONS AND FUTURE WORK

In this paper, a new feature set MS-LTP for face detection is presented. Comparisons are made among Haar, LBP, LTP, MB-LBP and MS-LTP as feature sets for face detection. The experimental results on the CMU-MIT face test set show that MS-LTP is comparable with MB-LBP on common application but is superior to four other feature sets under noisy conditions and much more rapid in detection speed.

We plan to carry out more experiments on the user-specified parameter t and figure out how it affects the performance. Also, we will try to let the machine learn a proper t during training instead of manually specifying the value to make the MS-LTP feature set more discriminative.

6. ACKNOWLEDGEMENTS

This work is supported by collaborative Research Project (SEV01100474) between Beijing University of Posts and Telecommunications and France Telecom R&D, and The National Natural Science Foundation of China (90920001).

7. REFERENCES

[1] Xiaoyang Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. on Image Processing*, vol. 19, pp. 1635 –1650, 2010.

[2] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja, "Detecting faces in images: A survey," *IEEE Trans. on PAMI*, vol. 24, no. 1, pp. 34–58, 2002.

[3] Cha Zhang and Zhengyou Zhang, "A survey of recent advances in face detection," Tech.Rep., Microsoft Research, 2010.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001, vol. 1, pp. I–511 – I–518 vol.1.

[5] Rainer Lienhart and Jochen Maydt, "An extended set of haar-like features for rapid object detection," in *IEEE ICIP*, 2002, pp. 900–903.

[6] T. Mita, T. Kaneko, and O. Hori, "Joint haar-like features for face detection," in *ICCV*, 2005, vol. 2, pp. 1619 –1626.

[7] Hongliang Jin, Q. Liu, H. Lu, and X. Tong, "Face detection using improved lbp under bayesian framework," in *Image and Graphics*, 2004, pp. 306 – 309.

[8] Lun Zhang, Rufeng Chu, Shiming Xiang, Shengcai Liao, and S Li, "Face detection based on multi-block lbp representation," *Advances in Biometrics*, vol. 4642, pp. 11–18, 2007.

[9] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao, "Learning sparse features in granular space for multi-view face detection," in *Automatic Face and Gesture Recognition*, 2006, pp. 401 –406.

[10] Robert E. Schapire and Yoram Singer, "Improved boosting algorithms using confidence-rated predictions," *Maching Learning*, vol. 37, pp. 297–336, 1999.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, pp. 2000, 1998.

[12] Rainer Lienhart, Er Kuranov, and Vadim Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *In DAGM 25th Pattern Recognition Symposium*, 2003, pp. 297–304.

[13] Stan Z. Li and ZhenQiu Zhang, "Floatboost learning and statistical face detection," *IEEE Trans. on PAMI*, vol. 26, pp. 1112–1123, 2004.

[14] Shengye Yan, Shiguang Shan, Xilin Chen, and Wen Gao, "Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection," in *CVPR*, 2008, pp. 1 –7.

[15] Stan Z. Li, Long Zhu, Zhenqiu Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *ECCV*, 2002, pp. 67–81.

[16] H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. on PAMI*, vol. 20, no. 1, pp. 23 –38, 1998.