Face Detection Based on Frequency Domain Features

B.H. Shekar(✉) and D.S. Rajesh

Department of Computer Science, Mangalore University, Mangalagangothri, Mangalore 574199, Karnataka, India
bhshekar@gmail.com, rajeshds1972@gmail.com

Abstract. In this paper we have developed a novel face detection method using the Stockwell and the log dyadic wavelet transform features, following the cascaded face detectors framework. Stockwell transform (ST) time frequency distribution of an image region is known for its excellent feature representational capabilities (due to the high resolution of the distribution). Log dyadic wavelet transform (LDWT) is capable of representing image patches with high accuracy. We have used the Stockwell transform and the log dyadic wavelet transform for representing the facial features effectively. Our face detection method consists of two stages. The first stage consists of a cascade of 4 face detectors constructed using discriminative facial ST features selected by the ADABOOST feature selection method. The second stage consists of a cascade of 4 more face detectors, each of them is a SVM classifier trained with face/nonfacial LDWT features. We have conducted our face detection experiments on the well known CMU-MIT and FDDB face detection datasets to verify the efficacy of our method.

1 Introduction

Face detection system involves location of human face regions and their size in a given digital image. A face recognition system relies heavily on a good face detection system. A face recognition system can recognize faces in an image only after a face detection system has identified regions of face in it. Recently Jun et al. [5] also developed a face detection system based on cascaded face detector framework. They used the ADABOOST algorithm [3] to select the best LBP features and their location on the face images. For each detector of the cascade, multiple LBP features were selected and incorporated in the detector until nearly 100% detection rate and around 50% false alarm rate was achieved on the test face/nonface images. They then repeated their work by replacing the LBP features by the LGP features which showed an improvement over their LBP based face detector. Further they repeated the same experiment using a hybrid of LBP, LGP and BHOG descriptors. In each stage of the ADABOOST feature selection based face detector system, they selected highly discriminative mixture of LBP, LGP and BHOG features on the face images. While LBP was global illumination invariant, LGP was local illumination invariant and BHOG

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 203–211, 2017.
https://doi.org/10.1007/978-3-319-69900-4_26

captured bigger facial parts like nose, eyes, etc. Because of all these features they could develop a more better face detection system than their previous one.

2 The Stockwell Transform and the Log Dyadic Wavelet Transform

Stockwell transform [1] based time frequency distribution (TFD) of a signal is found to be more accurate (Fig. 1 shows the results of our experiments conducted to demonstrate this fact) than the time frequency distribution obtained through other traditional transforms like the short time Fourier transform (STFT) and the Gabor transform. The log dyadic wavelet transform based representation of image signals [4] is more accurate than the representations obtained through the traditional discrete wavelet transform (Fig. 2 shows the results of our experiments conducted to demonstrate this fact). Hence we have used the ST and the LDWT in our face detection system for representation of image features.

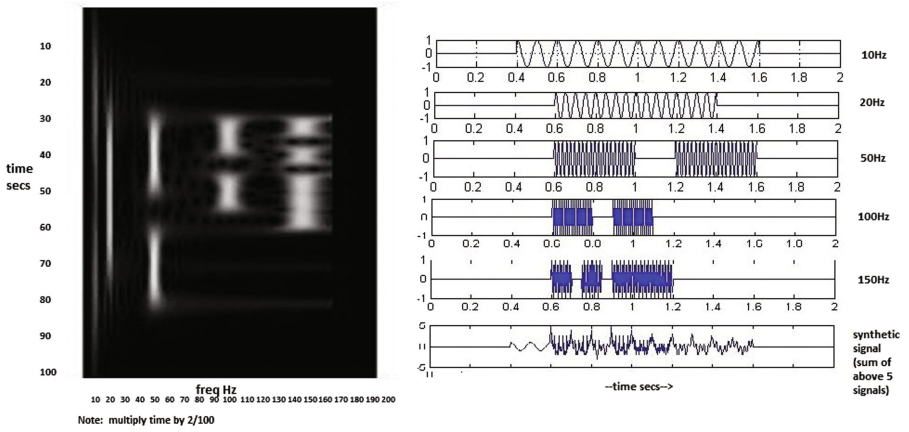

Fig. 1. (right) The last row is a synthetic signal obtained as a sum of the other 5 sinusoids above. (left) The Stockwell transform based time frequency distribution characterizing the synthetic signal accurately both in time and frequency axes.

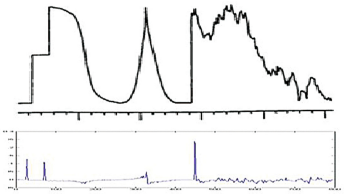

Fig. 2. (plot at the top) 1D Signal with sharp edges. (plot at the bottom) Edges detected accurately by convolving the signal with the 1D LDWT kernel.

3 Proposed Method

Our face detection system consists of two stages. The first stage is made up of a cascade of 4 face detectors each being constructed using highly discriminative Stockwell transform based feature classifiers and the second stage is made up of a cascade of 4 more face detectors each being a SVM classifier trained using LDWT coefficients of face/non face training images. Each face detector is constructed in such a way that they have 99.5% face detection rate and 50% false alarm rate. Given a sample image at the input of the face detection system, each detector rejects non face regions in it and forwards the probable face image regions to the next face detector in the cascade. At the output, our system is supposed to localize the face regions in the input image.

3.1 Construction of Stage 1 Face Detectors

Highly discriminative Stockwell transform based features (Stockwell transform of 3×3 and 5×5 size regions on face/non face training image samples) are selected as classifiers using the ADABOOST feature selection method, in constructing the face detectors of stage 1. We have set the parameters of the Stockwell transform in such a way that for a 3×3 size image signal, we obtain a 3×9 size TFD plot and for a 5×5 size image signal we obtain a 7×25 size TFD plot. Following are our Stockwell transform based features from which highly discriminative ones are selected as feature classifiers of the face detector.

1. Stockwell transform TFD of a 3×3 size image region, $ST3_{Fi}(x, y)$, around every pixel (x,y) of a face image sample is computed (at x = 2,3,4,.....21, y = 2,3,4.....23) for all face image samples i = 1,2,3.....16000. Figure 3 shows some of the 3×3 size image regions (at locations (2,2), (2,7), (2,21), (23,2) and (23,21) of an imaginary face image Fi), whose Stockwell transform TFDs $ST3_{Fi}(2, 2)$, $ST3_{Fi}(2, 7)$, $ST3_{Fi}(2, 21)$, $ST3_{Fi}(23, 2)$ and $ST3_{Fi}(23, 21)$ are computed.
2. Similarly Stockwell transform of a 5×5 size image region, $ST5_{Fi}(x, y)$, around every pixel (x,y) of a face image sample is computed (at x = 3,4,.....20, y = 3,4.....22) for all face image samples i = 1,2,3.....16000.
3. Stockwell transform of a 3×3 size image region, $ST3_{NFi}(x, y)$, around every pixel (x,y) of a non face image sample is computed at x = 2,3,4,.....21, y = 2,3,4.....23, for non face image samples i = 1,2,3.....16000.
4. Stockwell transform of a 5×5 size image region, $ST5_{NFi}(x, y)$, around every pixel (x,y) of a non face image sample is computed at x = 3,4,.....20, y = 3,4.....22, for non face image samples i = 1,2,3.....16000. Note that $ST3_{Fi}(x, y)$, $ST5_{Fi}(x, y)$, $ST3_{NFi}(x, y)$ and $ST5_{NFi}(x, y)$ are all vectors of either size 3×3 or 5×5.

Apart from these our method uses the following features during the face detector construction.

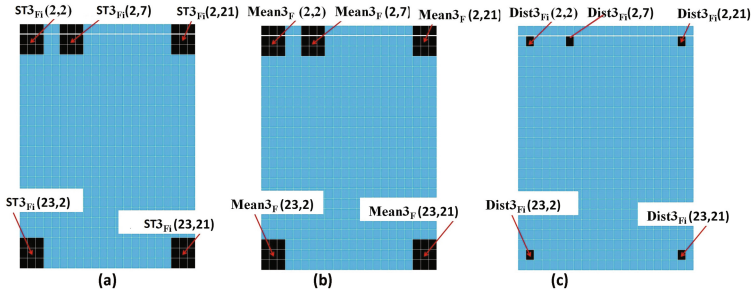

Fig. 3. The 3×3 size image regions used in computing ST based features.

1. Mean Stockwell transform feature $mean3_F(x, y)$, of 3×3 size image regions of face image samples at location (x,y) is computed as $\sum ST3_{Fi}(x, y)/16000$, where the summation is over i = 1,2,3.....16000. This mean is computed at all locations x = 2,3,4,.....21, y = 2,3,4.....23, using face image samples. Figure 3 shows some of the mean features computed at locations (2,2), (2,7), (2,21), (23,2), (23,21) i.e. $mean3_F(2, 2)$, $mean3_F(2, 7)$, $mean3_F(2, 21)$, $mean3_F(23, 21)$ and $mean3_F(23, 2)$
2. Mean Stockwell transform feature $mean3_{NF}$(x,y), of 3×3 size image regions of non face image samples is computed at location (x,y) as $\sum ST3_{Nfi}(x, y)/16000$ where i = 1,2,3.....16000. This mean is computed at all locations x = 2,3,4,.....21, y = 2,3,4.....23, using non face image samples.
3. Similarly $mean5_F$(x,y) and $mean5_{NF}$(x,y) are computed using $ST5_{Fi}$(x,y) and $ST5_{NFi}$(x,y) features respectively (at x = 3,4,.....20, y = 3,4.....22). Note that $mean3_F(x, y)$, $mean5_F(x, y)$, $mean3_{NF}(x, y)$ and $mean5_{NF}(x, y)$ are all vectors.
4. $Dist3_{Fi}(x, y) = \frac{chi\ square\ distance\ of\ ST3_{Fi}(x,y)\ from\ mean3_F(x,y)}{chi\ square\ distance\ of\ ST3_{Fi}(x,y)\ from\ mean3_{NF}(x,y)}$ is the chi square distance between the Stockwell transform feature of face image Fi, $ST3_{Fi}$(x,y) and mean facial Stockwell transform feature $mean3_F$(x,y) over the mean non facial Stockwell transform feature $mean3_{NF}$(x,y).
5. $Dist3_{NFi}(x, y) = \frac{chi\ square\ distance\ of\ ST3_{NFi}(x,y)\ from\ mean3_F(x,y)}{chi\ square\ distance\ of\ ST3_{NFi}(x,y)\ from\ mean3_{NF}(x,y)}$ is the chi square distance between the Stockwell transform feature of non face image NFi, $ST3_{NFi}$(x,y) and mean facial Stockwell transform feature $mean3_F$(x,y) over the mean non facial Stockwell transform feature $mean3_{NF}$(x,y). Also $Dist5_{Fi}$(x,y) and $Dist5_{NFi}$(x,y) correspond to 5×5 features. Note that $Dist3_{Fi}$(x,y), $Dist3_{NFi}$(x,y), $Dist5_{Fi}$(x,y) and $Dist5_{NFi}$(x,y) are all scalars.

A low value of $Dist3Fi(x, y)$ indicates that this Stockwell transform feature at (x,y) is a dominant feature of face images and also a recessive feature of non face images. A large value of $Dist3NFi(x, y)$ indicates that this Stockwell transform feature at (x,y) is a dominant feature of non face images and also a recessive feature of face images. Locations (x,y) which have a small $Dist3Fi(x, y)$ and a large $Dist3NFi(x, y)$ (over all i = 1,2,3...16000) are the best candidates to be a feature classifier (and hence are capable of distinguishing face image regions

from non face image regions). This is what our ADABOOST feature selection method does during face detector construction

Construction of First Face Detector of Stage 1: The cascade of face detectors framework based on ADABOOST feature selection method (followed by Viola and Jones [2]) is used in our method. Each of the face detectors in the cascade is capable of performing face detection with around 99.5% detection rate and 50% false alarm rate. Here we explain the procedure followed in constructing the first face detector (out of the cascade of face detectors) of our face detection system. Algorithm 1 along with Algorithms 2 and 3 is used in constructing the face detector. Algorithm 1 goes through several iterations (calling Algorithms 2 and 3 in each iteration) selecting the next most discriminative Stockwell transform feature classifier in each iteration and constructs the face detector using these classifiers. The number of feature classifiers thus selected must be capable of performing face detection with a 99.5% face detection rate and 50% false alarm rate. Using the Stockwell transform features explained above, Algorithms 1, 2 and 3 construct face detector as follows.

Algorithm 1

1. Assign weights to face training samples as $Wf(i)$=1/(2*number of face training samples) and and non face training samples $Wnf(i)$=1/(2*number of non face training samples) for all i = 1,2,3....16000.
2. Using Algorithm 2 select the next most discriminative Stockwell transform feature classifier
3. Construct a face detector using the feature classifiers selected so far and evaluate its classification performance (the detection and false alarm rate) using the classification method given by Algorithm 3.
4. If the classification performance of 99.5% detection rate and 50% false alarm rate is achieved, the face detector construction is complete. Otherwise go to step 2 to include the next most discriminative feature classifier into the face detector under construction, after the following procedure:
   - update (reduce) the weights of training samples which were correctly classified in this round (in Algorithm 2 during step 2) as follows, $Wf(i)$= $Wf(i)*Feat(iter).beta$ where i = index of correctly classified face samples, $Wnf(i)$=$Wnf(i) * Feat(iter).beta$ where i = index of correctly classified non face samples. By doing so the algorithm will not reselect a feature classifier already selected in the previous rounds.
   - Find $weight\_sum$, sum of weights of all samples. Normalize the weights as follows: $Wf(i) = \frac{Wf(i)}{weight\_sum}$ where i = 1,2,3,...$fnum$ and $Wnf(i) = \frac{Wnf(i)}{weight\_sum}$ where i = 1,2,3,....$nfnum$ such that the sum of sample weights is again 1. Proceed to step 2 to incorporate a new feature classifier.

Algorithm 2

1. The minimum possible classification error corresponding to location (x,y) (assuming the Stockwell transform feature at location (x,y) as the classifier) and scale 3 (i.e. 3×3 size features) is computed. For this computation classification is done over all training samples (face/non face images). Minimum possible classification error at location (x,y) and scale 3 is computed as follows:

- Sort the array $Dist3_{Fi}$(x,y) where i = 1,2,3.....16000, in ascending order and store it in Dist_3xy. Also sort $Dist3_{NFi}$(x,y) where i = 1,2,3.....16000 and concatenate it to Dist_3xy.
- **for i =1:16000**
  set **thresh**=Dist_3xy(i);
  Classify the training samples whose values in Dist_3xy are less than **thresh** as faces and the rest as non faces. Comparing the classification result with the ground truth compute the classification error. The sum of weights of the misclassified samples is the classification error of iteration i, err3[i].
  endfor

- min3(x,y)=minimum(err3) is the minimum possible classification error at position (x,y) and scale 3.

2. Similarly we can find the min3(x,y) at all positions x = 2,3,4,.....21, y = 2,3,4.....23. Similarly, using $Dist5_{Fi}$(x,y) and $Dist5_{NFi}$(x,y) we can find min5(x,y) at all positions x = 3,4,.....20, y = 3,4.....22. The minimum among all min3(x,y) (i.e. at positions x = 2,3,4,.....21, y = 2,3,4.....23) and min5(x,y) (at all positions x = 3,4,.....20, y = 3,4.....22) gives us the position (X,Y) and scale (which is 3 if minimum is from min3(x,y) or else 5 if minimum is from min5(x,y)) of the most discriminative feature classifier of this round.

3. Now the best classifier of the current iteration is at location (X, Y). Record classifier as $Feat.loc$=[X, Y]; $Feat.scale$=3 if minimum was obtained in err3 and $Feat.scale$=5 if minimum was obtained in $err5$; Let $\beta = \frac{min\_err}{1-min\_err}$ and $Feat.beta$=β, $Feat.thresh$=**thresh** value at (X, Y) corresponding to the minimum classification error, $Feat.confidence$=$log(\frac{1}{\beta})$.
return $Feat$

Algorithm 3: Tries to build a classifier of 99.5% detection and 50% false alarm rate

val=0; **conf_sum**=0
for i = 1:no of feature classifiers selected so far.
conf_sum=**conf_sum** + $Feat(i).confidence$
end

Repeat

for each given test sample

1. **for i**=1:no of feature classifiers selected so far
   - The ratio of sample-to-mean_face by sample-to-mean_nonface distance at $(x, y) = Feat(i).Loc$ i.e.

$$Dist_{sample}(x, y) = \frac{chi\_square\_dist(ST3_{sample}(x, y), Mean3_F(x, y))}{chi\_square\_dist(ST3_{sample}(x, y), Mean3_{nF}(x, y))} \tag{1}$$

     computed. Use scale 5 features if the classifier belongs to scale 5.
   - If this ratio is less than $Feat(i).threshold$ it implies that the i^{th} feature classifier of current iteration has decided the sample to be a face and the classifier confidence is accumulated as **val**=**val**+ $Feat(i).confidence$.
     end
2. **val** contains the confidence values of all the feature classifiers that decided the sample as face. if **val** > **par*****conf_sum** the sample is decided as face else as non face.

end

Compute the detection and false alarm rate using the decisions in step 2.

Until the best possible detection and false alarm rate is obtained by varying the "par" value

After the construction of the first face detector, the non face training samples are classified (**Algorithm 3** is used for classification) and unclassified samples are collected. For the construction of the second face detector of the cascade of stage 1, we use the full set of face training samples and these misclassified non face training samples. Like this 4 detectors of first stage are constructed.

3.2 Stage 2

Log dyadic wavelet transform (LDWT) features of the training samples are used in constructing the 4 face detectors of stage 2. A SVM classifier is trained using the LDWT features until the classifier shows a performance 99.5% detection and 50% false alarm rate on the training samples. During the estimation of this performance, the misclassified training non face samples are collected and along with the full set of the training face samples the construction of the second face detector of this stage continues. This method is followed until the 4 face detectors of the stage is constructed.

4 Experiments and Data Set Preparation

Using face recognition dataset samples of ORL, LFW, FERET, ABERDEEN, PIE and MIT datasets we have formed our training and testing face samples (the

cropped face images include variations like scaling, rotation, poor illumination, poor resolution in them). 16000 each of training and testing face samples were obtained. 32000 non face samples were collected from the MIT dataset, of which 16000 were reserved for training and the rest for testing. We have conducted face detection experiments on the CMU-MIT dataset (both rotated and normal version) and the FDDB dataset, the qualitative and quantitative experimental results can be seen in Figs. 4 and 5. Given a test face sample, the first four face detectors of stage 1 removed the easier non face regions from the sample and forwarded the remaining regions to the next face detector of the cascade.

Fig. 4. The detection performance on the (a). FDDB (top two rows) and (b). CMU-MIT dataset (middle row). (c). shows the output of each of the 8 face detectors of the cascade of the face detection system, on a rotated CMU-MIT image. Also it shows the detected image regions.

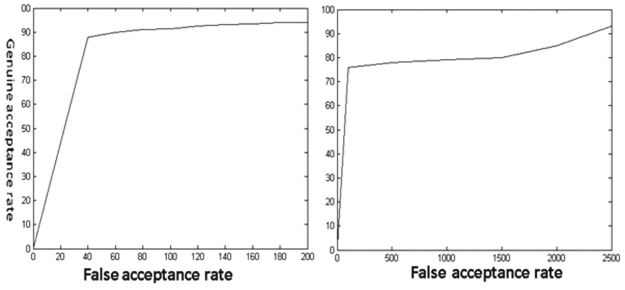

Fig. 5. The ROC curves of the performance of the proposed method on CMU-MIT (left) and FDDB (right) datasets

The tougher image regions (that contained nearly face-like-looking non-face regions) were removed from the test sample by the next 4 face detectors of stage 2. Each face detector of the cascade has checked the presence of face regions at a given location on the test sample at 8 different scales.

5 Conclusion

We have developed a face detection system following the cascade of detectors model. We have used Stockwell transform and log dyadic wavelet transform feature representation of images. Using these features we have built the face detection system to classify face and non face samples. Our experiments on the FDDB and CMU-MIT face detection datasets have shown comparable performance with the state of the art methods.

References

1. Todorov, T.I., Margrave, G.F.: Variable factor S-transform seismic data analysis. CREWES research report, vol. 21 (2009)
2. Viola, P., Jones, M.: Robust real-time face detection. In: Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, vol. 2, p. 747 (2001)
3. Freund, Y., Shapire, R.E.: A short introduction to boosting. J. Jpn. Soc. Artif. Intell. **14**(5), 771–780 (1999)
4. Tu, G.J., Karstoft, H.: Logarithmic dyadic wavelet transform with its applications in edge detection and reconstruction. Appl. Soft Comput. **26**, 193–201 (2015)
5. Jun, B., Choi, I., Kim, D.: Local transform features and hybridization for accurate face and human detection. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1423–1436 (2013)