

# DETECTION OF FACES UNDER SHADOWS AND LIGHTING VARIATIONS

G. Iyengar, C. Neti  
IBM T. J. Watson Research Center  
Yorktown Heights, NY 10598, USA  
{giyengar,cneti}@us.ibm.com

Abstract - Successful face detection and facial feature extraction is crucial for for a variety of applications, including speech recognition and fatigue monitoring. Detecting and tracking faces in a moving automobile is challenging because of a variety of reasons. Among these reasons are changing poses, extreme lighting changes and shadowing. In this paper, we investigate two approaches for shadow compensation in such images. Together, these techniques offer a good improvement in the face detection accuracy on a data set captured in a moving automobile.

## INTRODUCTION

Many emerging multimedia applications strive to model audio-visual events for the purposes of understanding, indexing and managing content. Joint use of audio-visual information for a variety of speech related tasks recently attracted significant interest [1]-[3]. Much of this interest focuses on ways of combining the video channel information with its audio counterpart, in the quest for a combined system that outperforms audio-only systems.

Because of environmental conditions such as road noise and wind noise, acoustic-only speech recognition in a moving automobile is a very hard problem. In such situations, augmenting the acoustic stream with visual speech information can be used to successfully improve the speech recognition performance. Such a performance improvement depends on both the audio-visual fusion architecture, as well as on the visual front end, namely, on the extraction of appropriate visual features that contain relevant information about the audio-visual event.

In this paper, we concentrate on the task of face detection and facial feature extraction, which forms the visual front-end in such a system. Much of the prior work in face detection [4]-[8] has concentrated on mostly frontal facial poses under controlled lighting conditions. Detecting and tracking faces in a moving automobile is challenging because of a variety of reasons. Among these reasons are changing poses, extreme lighting changes and shadowing. In this paper, we study the effects of shadow compensation on the performance of two face detectors operating on faces captured in moving automobiles. As we see from experiments, these techniques offer good improvement in face detection accuracy. The novelty of the results is two-fold. First, shadow compensation improves face detection accuracy. Second, shadow compensation improves face detection accuracy even if it is used only in the modeling stage.



Figure 1: Example of a face captured in an automobile

## SYSTEM OVERVIEW

Figure 1 shows a snapshot of a video sequence captured in a moving automobile. Notice the significant shadows and lighting effects. We note here that we record both the audio and the visual streams for automatic speech recognition (ASR) but in this paper concentrate on the task of extracting a good visual speech representation. For work on the importance of good quality visual speech representations, please refer to our paper[13].

## FACE DETECTION UNDER SHADOWS

We use two approaches for face detection in this paper. The first approach uses a Fisher discriminant and eigenspace based face detection approach to extract the face and locate facial features from video[8]. In this approach, an image pyramid over a range of scales is used to search for face candidates. Every face candidate is given a score based on several features like skin tone, proximity to face space and the Fisher discriminant. Once a face is detected, an ensemble of facial feature detectors are used to extract important facial features, including the lip corners and mouth centers. Subsequently, a size-normalized mouth image of size  $45 \times 30$  pixels is extracted from the face image.

In the second approach, we extract DCT coefficients from training faces and non-faces and use these exemplars to build Gaussian Mixture Models for both these classes. Faces are then detected using a likelihood ratio test. As in the earlier approach, scale and rotation invariance is achieved by searching over an image pyramid and a set of rotations. Once a face has been detected, we use the above approach to locate facial landmarks and extract the mouth image.

Both these approaches for face detection perform very well on a large database of faces captured under controlled lighting conditions. Under these conditions, we obtain a face detection accuracy of greater than 99.5% using the Fisher discriminant and the GMM face detector. However, under real-world conditions such as in an

automobile, the performance of both the face detectors degrades quite severely as seen in Table 1.

## Shadow Compensation By Dynamic Range Compression

This simple technique is based on the assumption that shadow and sunlight affect only the luminance and not the color components. Shadow regions are darker than average and the sunlit regions in an image are brighter than average. This approach brightens all pixels below a threshold  $T_1$ . Similarly, to compensate for excessive brightness regions, we darken all pixels above a threshold  $T_2 > T_1$ . In order to ensure a smooth transition for pixels that lie between these two thresholds, we linearly compensate the pixels that lie between these two thresholds. This process reduces the dynamic range of the pixels in the original image. This is a simple, low-computation approach to shadow compensation. Experiments indicate that it performs reasonably. We now detail a more sophisticated scheme for shadow compensation that orders the image into layers of contrast depths.

## K-Factor Shadow Removal

K-factorization decomposes a normalized image (pixels taking values between 0 and 1) into a product of factor images[10]. For example, an image  $I(x, y)$  at pixel co-ordinates  $(x, y)$  can be factored as

$$I(x, y) = \prod_{n=1}^{\infty} f_n(x, y) \quad (1)$$

where the factors  $f_n(x, y)$  are defined as

$$f_n(x, y) = \frac{1 + k^n g_n(x, y)}{1 + k^n} \quad (2)$$

with  $g_n(x, y)$  some chosen basis function that takes binary values. The chosen value of  $k$  orders the image factors into different contrast depths  $D_n = \frac{k^n}{1+k^n}$ . We define a residual image  $I_n(x, y)$  as the image that is left after the  $n$ th factor is removed from the  $n - 1$ th residual image. That is

$$I_{n-1}(x, y) = I_n(x, y) \times f_n(x, y) \quad (3)$$

and  $I_0(x, y)$  is the original image.

Whenever the basis function  $g_n(x, y)$  is unity at a pixel location, the contribution of the factor  $f_n$  at that pixel to the original image is non-informative. Also, if at a particular pixel of the residual image  $I_{n-1}(x, y)$  which is above the threshold  $\frac{1}{1+k^n}$ , the basis function is chosen to be zero, then the product factorization dictates that we can never reach the original image value  $I(x, y)$  at that pixel by such a choice. This gives us a simple rule to choose the basis functions:  $g_n(x, y) = 1$  whenever  $I_{n-1}(x, y) > \frac{1}{1+k^n}$ . At other locations, we choose  $g_n$  to be zero.

Once an image is factored, we find that the majority of the shadows are concentrated in the first few factors. The factor,  $f_n$  identifies pixels in the image that are below  $\frac{1}{1+k^n}$ . At all other pixel locations, the factor is unity. If we divide the original image by a quantity proportional to this factor, it is akin to boosting the original image at locations with shadows. While this approach brightens dark, possibly shadow



Figure 2: Example ground-truth marked-up faces

pixels, it also brightens naturally dark pixels such as facial hair, pupil regions etc. In the automobile scenario, there are excessive lighting in addition to strong shadows as evidenced in Fig. 1. This algorithm does not address this excess brightness which also has a deleterious effect on the detection accuracy.

## Experimental Results

In addition to the data captured under controlled lighting conditions, we collected about 30 minutes of audio-visual data of people asking directions, reciting telephone numbers in a moving automobile. To evaluate face detection performance, we sub-sampled these video sequences at regular intervals and extracted 876 still images. Each of these images were hand annotated to mark locations of facial features as seen in the examples in Fig. 2. These ground truth images are split into a test and training set of 561 and 246 images respectively. The training set of 246 images were combined with the set of facial images captured in controlled lighting conditions to achieve better generalizability. Together with this set, we have a total of 1767 training images, all marked up as in Fig. 2 spanning more than 210 different subjects with mixed race and gender. We note that such approaches has been adopted in training speech recognition systems[12]. Henceforth, we will refer to the test set as AUTO561 and the training set as MS1767, reflecting both the nature of the facial images and the number in each set. In addition, the controlled lighting test set comprises of 421 facial images and will be referred to as VVAV421. The Fisher discriminant is computed on a 118 dimensional subspace of a  $11 \times 11$  size-normalized face template. The DCT GMM model is a 10 mixture model trained on the first 50 DCT coefficients of the  $11 \times 11$  window.

We perform two sets of experiments to illustrate effectiveness of shadow compensation. In the first set, we test and train the system identically. That is, we train the face detector using uncompensated images and test using uncompensated images. Likewise, we train the detector using shadow compensated images and test using shadow compensated images. In the second set, the training set is compensated and the test set is not compensated for shadows. Table 1 shows the results for the first set where the training and test cases are matched and the mismatched set results are detailed in Table 2.

Detector	Train set	Test set	Shadow Tech.	Acc. (%)
Fisher	MS1767	AUTO561	None	62.10
Fisher	MS1767	VVAV421	None	99.5
Fisher	MS1767_k	AUTO561_k	k-factor	66.08
Fisher	MS1767_r	AUTO561_r	range comp.	68.17
DCTGMM	MS1767	AUTO561	None	77.20
DCTGMM	MS1767	VVAV421	None	99.7
DCTGMM	MS1767_k	AUTO561_k	k-factor	82.3
DCTGMM	MS1767_r	AUTO561_r	range comp.	80.1

Table 1: Results for matched training and testing. The suffixes after the dataset name indicate the shadow compensation process used.

Detector	Train set	Test set	Shadow Tech.	Acc. (%)
Fisher	MS1767_k	AUTO561	k-factor	65.06
Fisher	MS1767_r	AUTO561	range comp.	63.35
DCTGMM	MS1767_k	AUTO561	k-factor	79.92
DCTGMM	MS1767_r	AUTO561	range comp.	66.57

Table 2: Results for training with compensated images and testing on uncompensated images.

As can be seen from Table 1, both the face detectors perform extremely well under controlled lighting conditions. The DCT Gaussian Mixture Model face detector has a better performance compared to the Fisher detector for the adverse lighting conditions. This is possibly because of the ability of the mixture to model arbitrary probability density functions whereas the Fisher discriminant implicitly assumes a single Gaussian model for each class. For both the detectors, shadow compensation offers significant improvement in face detection accuracy over uncompensated images. The best scenario is when the training images and the test images are both compensated. However, as seen from Table 2, the k-factor approach works reasonably even if the training and test sets are mismatched. We make two observations here. First, even in the matched case, shadow compensation improves performance. It is not obvious why this should happen. Secondly, compensation during training helps even if the face detectors are used with uncompensated images. A possible explanation is that the Fisher discriminant and the Mixture models benefit from clean training images and model the underlying “true” face space better when the training images are compensated for shadows.

## CONCLUSIONS

In this paper, we looked at two face detectors and two approaches for compensating for shadows in facial images captured in automobiles. These techniques offer significant improvement in face detection accuracy compared to uncompensated im-

ages. We are currently investigating robust tracking algorithms that can be used in conjunction with these shadow compensation techniques for robust detection and tracking of faces in automobiles.

## References

- [1] T. Chen and R. R. Rao, "Audio-Visual Integration in Multimodal Communication", Proc. IEEE, Vol. 86, pp. 837-852, 1998.
- [2] C. Bregler and Y. Konig, "Eigenlips" for Robust Speech Recognition", Proc. ICASSP, Adelaide, 1994.
- [3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Final Workshop 2000 Report, CLSP, Johns-Hopkins, Baltimore, 2000.
- [4] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," IEEE CVPR, 1997.
- [5] K. Sung and T. Poggio, "Example-based Learning for View-based Human Face Detection," IEEE TPAMI, Vol. 20(1), pp 39-51, January 1998.
- [6] H. A. Rowley, S. Baluja and T. Kanade, "Neural Network-Based Face Detection," IEEE TPAMI, Vol. 20(1), January 1998.
- [7] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition," IEEE CVPR, 1998.
- [8] A. W. Senior, "Face and feature finding for face recognition system", 2nd AVBPA, Washington, pp 154-159, March 1999.
- [9] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuro Science, Vol. 3:71-86, 1991.
- [10] J. L. Johnson and J. R. Taylor, "K factor image factorization," SPIE Optical Pattern Recognition X, pp 166-174, Orlando, FL, 1999.
- [11] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, B 39(1):1-38, 1977.
- [12] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Kluwer Academic Press, Boston, MA 1993.
- [13] G. Iyengar, G. Potamianos, C. Neti, T. Faruquie, and A. Verma. "Robust Detection of Visual ROI for Automatic Speechreading," To appear in IEEE Workshop on MMSP, Cannes, France, 2001.