Face Replacement with HOG and SVM
CIS 581 - Final Project - Fall 2015

Vidur S. Bhatnagar
MSE Robotics,
University of Pennsyvania,
vidurb@seas.upenn.edu

Prerna Srivastava
MSE Computer and Information Science,
University of Pennsyvania,
prernasr@seas.upenn.edu

***Abstract*—This project is an attempt to completely replace faces in video frames using Support Vector Machines (SVM) as classifiers on Histogram of Oriented Gradients (HOG). We trained an SVM classifier to first identify a face in an image using HOG descriptors and then identify fiducial markers like eyes, nose and mouth in the identified face using Matlab's Cascade Object Descriptor. Once a convex hull was identified in the target image face, Thin Plate Spline (TPS) morphing, alongwith Gradient Domain Blending, was used to warp features from source image face onto the target frame face.**

I. Introduction

Detection of faces in an image has long been a topic of interest in the computer vision domain and serves several practical purposes. Face detection is used from automatic detection of social media platform users like Facebook to anonymizing faces for use in public data like Google Street View, etc. Our project is focused on developing a face detection algorithm and then replacing faces in a target video frame from a set of pre-determined replacement faces. This report is organized to discuss our approach which consists of the following sections. The rest of this section discusses the choice of HOG features. Section II states the project details in terms of what we achieved and what we missed. Section III discusses our approach and algorithm for the task and Section IV concludes with results on different video sets.

A. HOG Descriptors for Object Detection

[1] discusses at length the advantages of using HOG features over other descriptors like SIFT and shape context. They are computed on a grid of cells and normalized over overlapping blocks for improved performance [1]. The motivation behind using Histogram of Oriented Gradients as cited by the authors was that local object appearances and shapes can be captured by the distribution edge orientations without exact knowledge of the edge position [1]. HOG descriptors have been widely used in a plethora of applications for object detection like pedestrian detection, object detection and face detection. There are several properties that make using HOG features advantageous like invariance to local photometric and geometric transformations [1].

II. Project Details

Our team primarily focused on Option 1, but also implemented substantial portions of Option 2. Details are as follows:

- What We Achieved
  - Face detection with a linear SVM classifier using HOG descriptors
  - Face Replacement using Thin Plate Spline warping
  - Great results on easy set, Average results on medium set, poor/no results on hard/challenging set
  - Blending with Gradient Domain, Poisson and Laplacian Pyramid blending - final output generated with Gradient Domain Blending
- What We Missed
  - Fiducial marker (eyes, nose and mouth) detection with more SVM classifiers - resorted to using Matlab's Cascade object detector
  - Source image repository with multiple face poses - used a single frontal pose for replacement
  - Motion Compensation

III. Algorithm Details

Our entire approach to the task of face replacement is presented in Fig. 1. Separate parts of the process are explained below:

A. HOG Descriptor Extraction

The basic procedure outlined in [1] was used for generating the HOG descriptors. A 64 by 64 window was used for detecting faces. This window was then divided into rectangular cells of 8 by 8 pixels inside which image gradients were calculated, counted and binned into a 9 bin histogram. The histograms were then normalized over 16 by 16 overlapping blocks. For the purpose of gradient computation, we did not use any smoothing and decided to use [-1 1] as suggested in the paper. We tried using a host of variations on both smoothing and gradient filters but got marginal performance gains on no smoothing and [-1 1] filter. We decided to use unsigned orientations for the purpose of binning as it performed slightly better than the signed orientations. For block normalization, we used a stride of 1 cell and L2 norm. The extracted HOG descriptors were of length 1764.

B. Model Training and Prediction

A linear SVM was used for as a classifier to predict whether a patch of 64 by 64 pixels is a face. The SVM was trained on 8,721 cropped face images of size 64 by 64 which were collected from the datasets: Caltech 10,000 Web Faces [2], Closed Eyes In The Wild (CEW) [3] as positive samples. For the negative dataset, we used 64 by 64 images of random object and scenes from INRIA Person Dataset [4], Microsoft Image Understanding Dataset [5] and random images of half cut faces. A total of 14,194 images were fed into the classifier as negative samples. We achieved a test accuracy of 98.03% on a subset of images from Caltech 10,000 Web Faces [2]. For prediction of face in an image, a sliding window of 64 by 64 pixels with a stride of 10 pixels was used on the target image. The trained SVM thus predicted a face over 64 by 64 block as shown in Fig. 2a. The prediction was done on multiple scales as we needed to get the face in the image to an approximate size of 64 by 64 for positive prediction.

C. *Face Landmarks Localisation*

After face detection, Cascade Object Detector that internally implements Viola Jones Feature Detector was used for getting bounding boxes of Left Eye, Right Eye, Nose and Mouth in the face image. First the nose was identified in the image, which was used as reference point for detecting the eyes and the mouth based on the golden ratio. The centers of these were used as correspondences for downstream warping. These were also used for discarding false positives that are predicted from our face detector model. Initially, we tried using HOG features for detection of face landmarks as well. We trained three instances of SVM's on images of eyes (32 by 64 pixels), nose (64 by 64 pixels) and mouth (32 by 64 pixels) detecting eyes, nose and mouth. The results were not encouraging and we decided to do away with the idea and use the cascade object detector instead.

D. *Thin Plate Spline Warping and Image Blending*

The correspondences identified in step 2 (Fig. 4d) of our algorithm are now used to warp the source image to the target frame. For this purpose, we create 2 morphed images - first, we extract the faces in both the source image and the target frame and create a source to target warping (resulting in Fig. 4e) and next, we warp the entire source image onto the target frame (resulting in Fig. 4f). The first warping is used to create a mask which then along with the second warping and the target frame is blended using Gradient Domain Blending [6]. We also tried Poisson [7] and Laplacian Pyramid blending [8], but the results with Gradient Domain Blending were far superior.

IV. Results

The results of our approach are presented in Fig. 4 and Fig. 5.

A. *SVM Detection With HOG Descriptors*

We achieved a test accuracy of 98.03% on a subset of images from Caltech 10,000 Web Faces [2]. A lot of spurious faces were detected in a few frames as shown in 2b. These were probably manifestations of incorrect HOG matches. Our classifier also returned multiple matches for the same face in the image as in 3a because of a small block stride.

B. *Comparison with Cascade Object Detector*

In most cases, our classifier was able to detect faces with similar accuracy as Cascade Object Detector. The bounding boxes for faces returned by Cascade Object Detector are more precise like in 3b and 3a. Also, Cascade Object Detector did not suffer from the problem of multiple hits and there were fewer instances of false predictions as compared to our linear SVM classifier.

C. *Face Replacement With Gradient Domain Blending*

With a linear SVM classifier and Gradient Domain blending approach, we were able to achieve fairly smooth results on the easy videos, but struggled with more complex video sets. This was partly due to the fact that we only had a single source image (we did not consider a set of source images with different poses) and can be partly attribute to the lack of robustness of our classifier against occlusion, etc.

While creating the final videos, we did not consider the frames where the algorithm could not detect faces. The loss of frames per video varied in number, due to several reasons as identified below:

- easy1.mp4 - missed around 20 frames, since the person on the right is looking down for the first few frames and the person on the left has an expression which our classifier could not match
- easy2.mp4 - missed around 10 frames, due to the speaking motion of the mouth for the right person
- easy3.mp4 - missed less than 5 frames
- medium1.mp4 - missed around 50 frames, since the lady in the center is looking away for the first few frames and the expressions of the man on the right were not picked by our classifier
- medium2.mp4, hard, challenging set - very poor results, did not generate the final output

Conclusion

This project proved to be a meaningful culmination of the several topics that were covered during the semester. However, there are various aspects of our algorithm which can be improved to get better results like motion compensation, consideration of different source image poses, etc.

Acknowledgment

The authors would like to thank Prof. Jianbo Shi and TAs, Jyh-Jing Hwang and Siddharth Mysore for their constant guidance throughout the course.

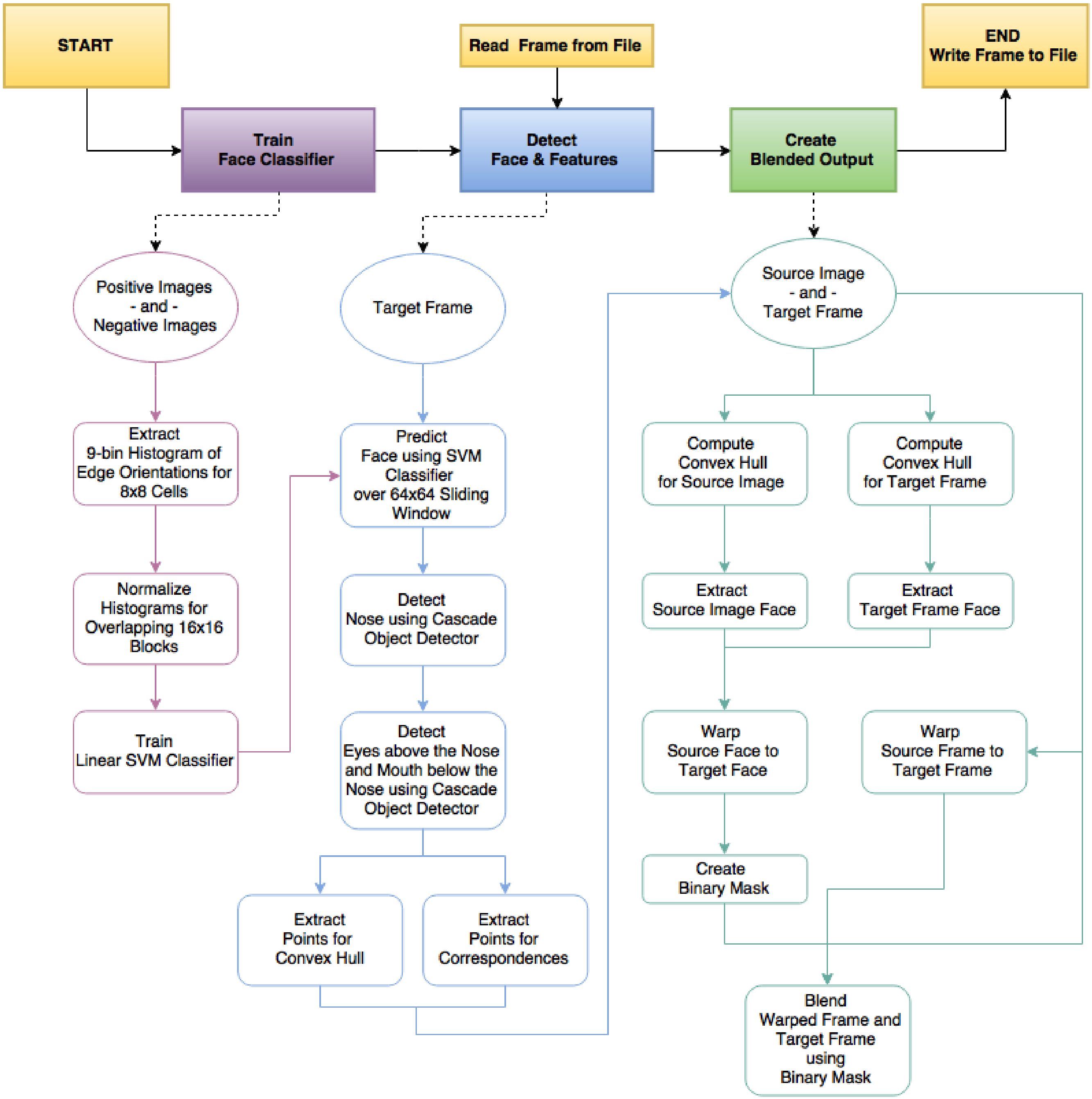
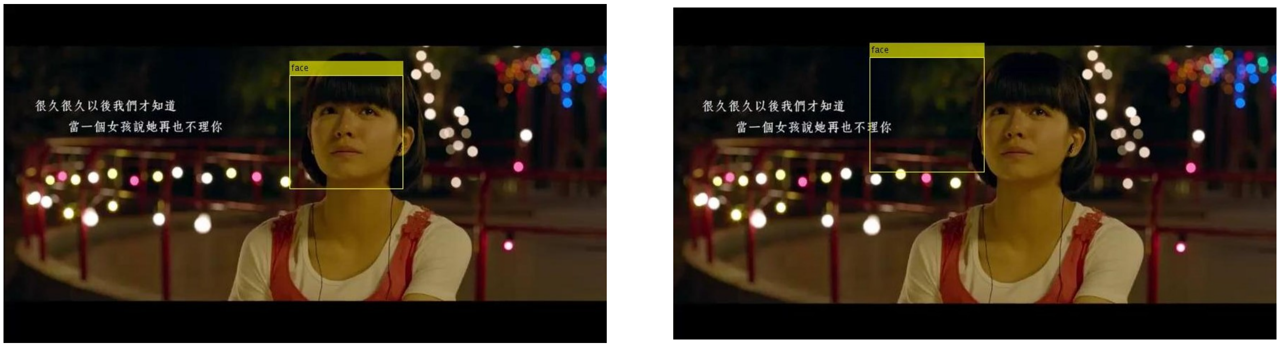


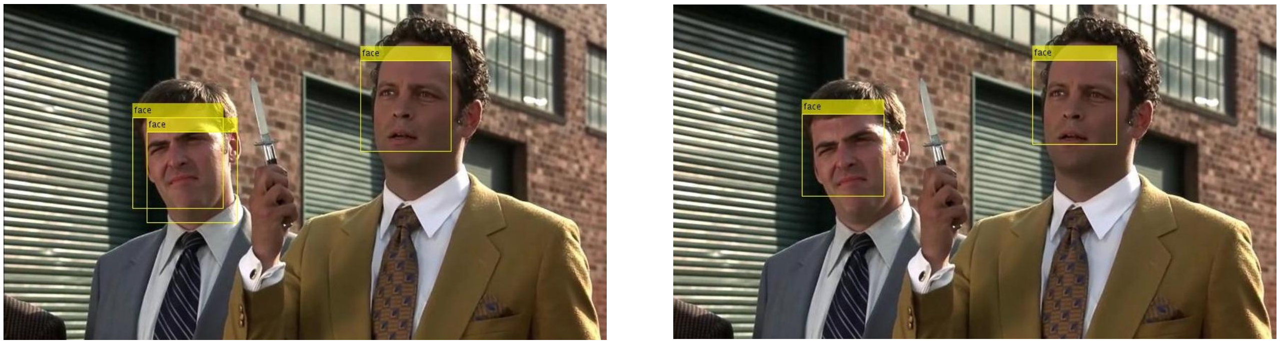
Fig. 1: Face Replacement Algorithm



(a) Successful Face Detection

(b) Spurious Face Detection

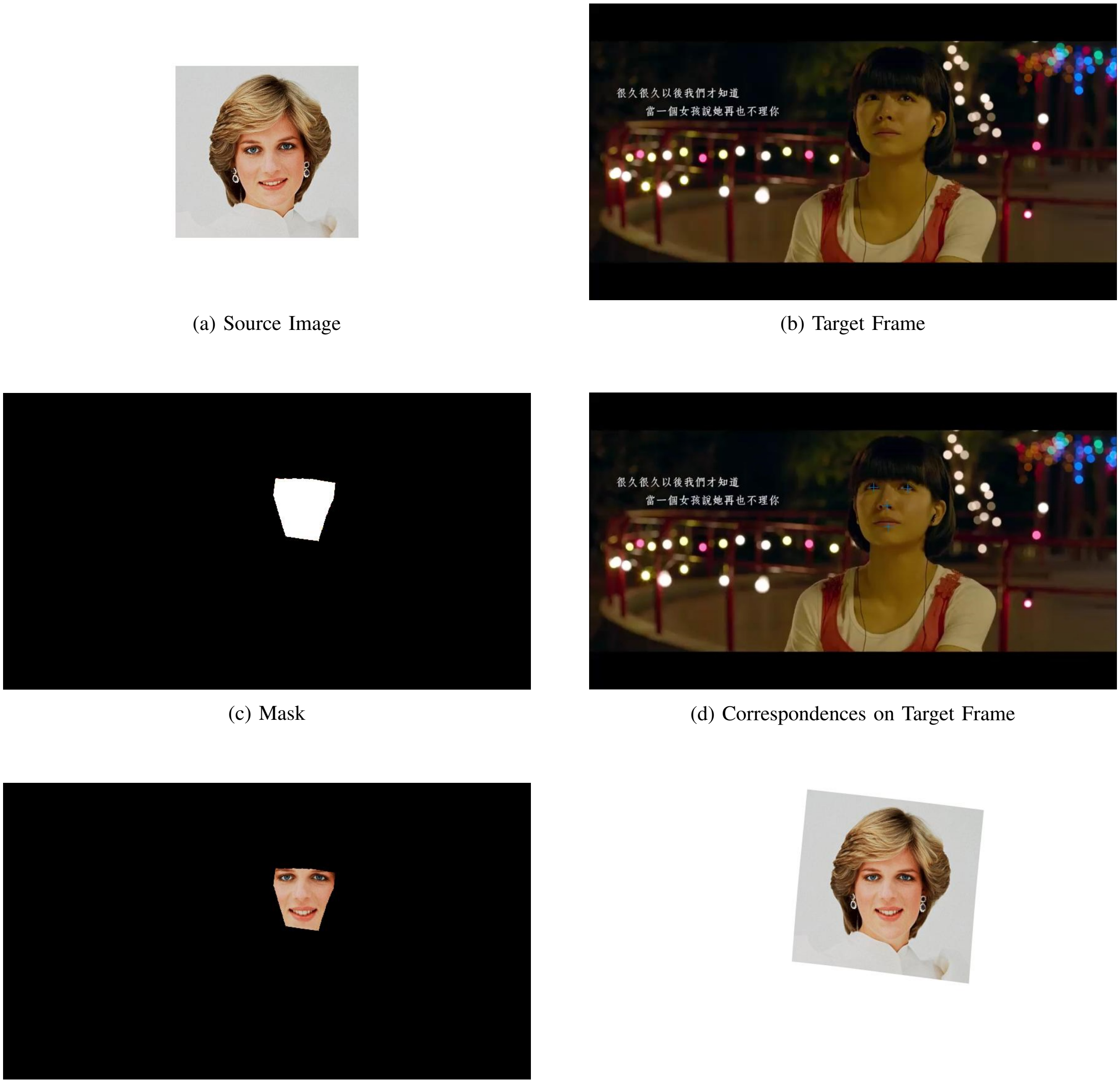
Fig. 2: Face Detection Results using SVM Classifier



(a) Face Detection using SVM

(b) Face Detection using Cascade Object Detector

Fig. 3: Face Detection Comparison



(a) Source Image

(b) Target Frame

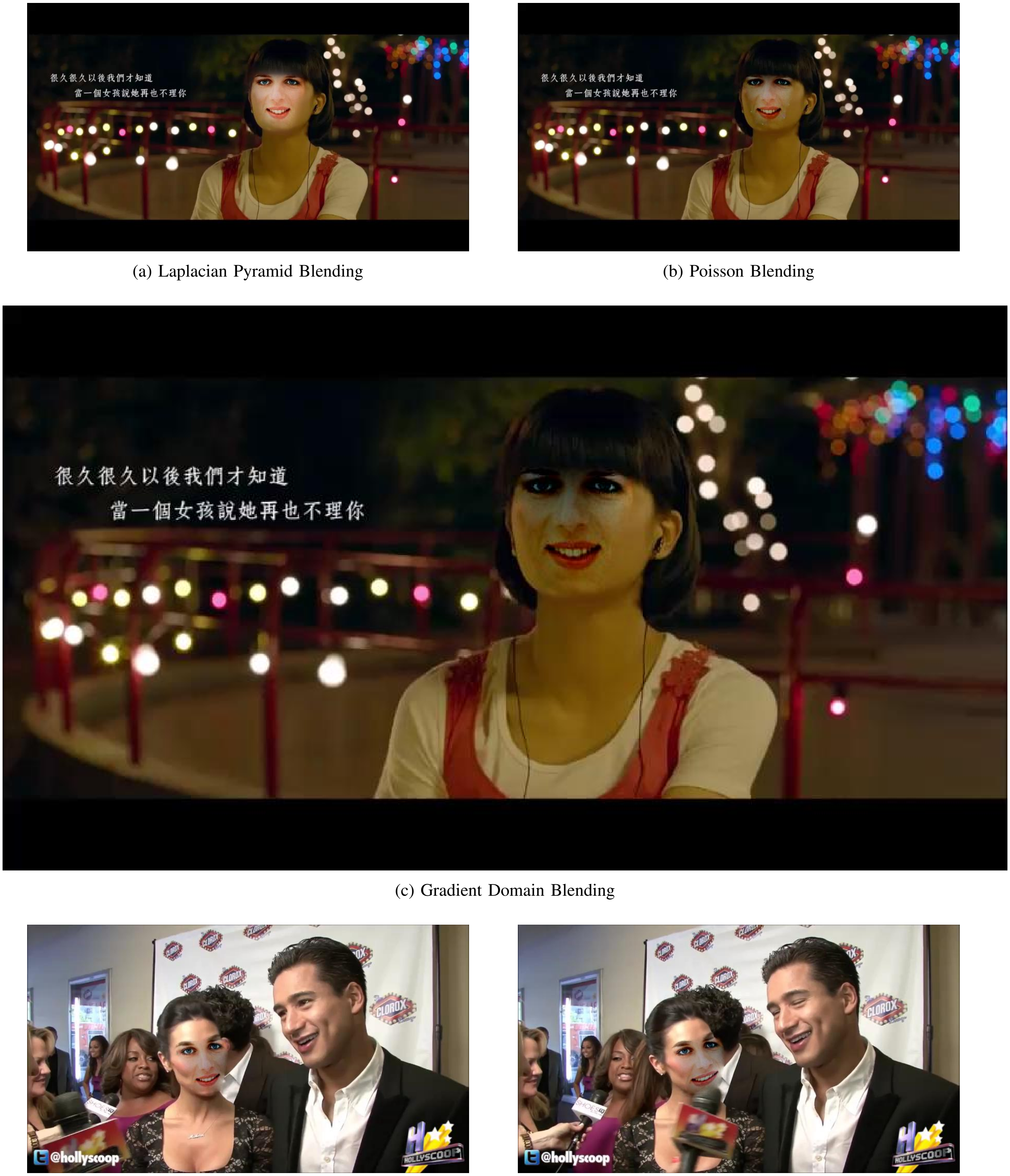
(c) Mask

(d) Correspondences on Target Frame

(e) Source Face warped to Target Face

(f) Source Frame warped to Target Frame

Fig. 4: Thin Plate Spline Warping Results



(a) Laplacian Pyramid Blending

(b) Poisson Blending

(c) Gradient Domain Blending

(d) Replacement Failure - pose mismatch

(e) Replacement Failure - occlusion

Fig. 5: Blending Results

REFERENCES

[1] Navneet Dalal and Bill Triggs, *Histograms of Oriented Gradients for Human Detection*, 2005

[2] Caltech 10, 000 Web Faces *http://www.vision.caltech.edu/Image_Datasets/Caltech_10K_WebFaces*

[3] Closed Eyes In The Wild (CEW) *http://parnec.nuaa.edu.cn/xtan/data/ClosedEyeDatabase*

[4] INRIA Person Dataset *http://pascal.inrialpes.fr/data/human/*

[5] Microsoft Image Understanding Dataset *http://research.microsoft.com/en-us/projects/objectclassrecognition/*

[6] Kecheng Yang *http://www.cs.unc.edu/ yangk/photog/a2p1.html*

[7] Masayuki Tanaka *http://www.mathworks.com/matlabcentral/fileexchange/37224-poisson-image-editing/content/PoissonEdiitng20151105/PoissonJacobi.m*

[8] Ray Phan *https://github.com/rayryeng/laplacianBlend*