RECOGNIZING PEOPLE IN VIDEO SEQUENCES

RED DALY

1. Abstract

We confront the problem of identifying people by name using video data. This paper has two main contributions. (1) A face recognition algorithm based on SIFT features that is capable of high accuracy rates for some datasets. (2) A simple algorithm to be used in conjunction with a facial recognizer to smooth recognition in video sequences.

Date: November 13, 2008.

2. Introduction

Facial recognition systems have wide applicability in search, security, robotics, and other areas. In many of these fields, video data is already highly available. However, most facial detection and recognition technology exists to identify individuals using a single still image. We leverage these existing technologies in the current system use some additional information available .

There has been extensive work in facial recognition. Commonly used facial recognition algorithms to date include Eigenfaces (Turk, 1991) and Fisherfaces (Belhumeur, 1997). We base our approach to facial recognition on SIFT features. SIFT, as described in (Lowe, 2004) is the best known a family of local feature detection algorithms which also includes SURF and derivatives of SIFT.

3. Problem

We are provided n training pairs $(x^{(i)}, y^{(i)}); i = 1, ..., n$ where $x^{(i)}$ denotes the i^{th} training image and $y^{(i)} = l; l = 1, ..., N$ if the l^{th} image is a picture of the l^{th} person's face. Given a test movie x with m frames, we wish to label each frame $y_{i,l} = 1; i = 1, ..., m$ if the i^{th} frame contains the l^{th} person and $y_{i,l} = 0$ otherwise.

In order to constrain the problem, we assume the positive training examples show only the face region of a single person's body. We do not, however, assume that the eyes, nose, or other features of each face are aligned within a training set. Ideally the algorithm we use should be robust to changes in illumination and pose.

4. Method

We attempted to solve the above problem by first developing a facial recognizer that is capable of working on still frames. We use SIFT features as the basic unit of recognition, as opposed to approaches based directly on pixel values. Thus, for each image we processed the input image for SIFT features and were then able to discard the image data. A basic assumption we made was that recognition is based on facial features and does not not use other aspects of the image to train or label.

4.1. **Facial recognition based on (Lowe, 2004).** As our first approach, we used the technique for matching SIFT keypoints as described in (Lowe, 2004). Keypoints are extracted from all training images and the test image. Let $\boldsymbol{r_j^{(i)}}$ denote the j^{th} 128-point SIFT feature descriptor vector corresponding to training pair $(x^{(i)}, y^{(i)})$. For each keypoint descriptor from the test image, $\boldsymbol{r}$, we find its nearest neighbor in the set of feature descriptors, $\boldsymbol{r_*^{(i)}}$. We also find the second nearest neighbor, $\boldsymbol{r_{**}^{(j)}}$ such that $y^{(i)} \neq y^{(j)}$. The ratio of the Euclidean distance between $\boldsymbol{r}$ and $\boldsymbol{r_*^{(i)}}$ and $\boldsymbol{r}$ and $\boldsymbol{r_{**}^{(j)}}$ must be less than .8 (i.e. $\alpha = \dfrac{\left|\left|\boldsymbol{r - r_*^{(i)}}\right|\right|_{\boldsymbol{2}}}{\left|\left|\boldsymbol{r - r_{**}^{(i)}}\right|\right|_{\boldsymbol{2}}} < .8$), and $y^{(i)}$ must be 1 in order for the match to count. (This exclusion method uses the empirical result that 96% of $p(CorrectMatch|\alpha)$ is distributed where $\alpha < .8$ while only 10% of $p(IncorrectMatch|\alpha)$ is distributed where $\alpha < .8$.)

These preliminary matches are further filtered by performing a Hough transform using the scale, orientation, and position components of the matched SIFT features, and maximizing in the Hough space. We did not perform the final geometric step used in (Lowe, 2004), and instead accepted the results of the Hough transform to cluster matched keypoints. We also weighted bin entries in our Hough transform by $3(1-\alpha)$, using the insight that $(1-\alpha)$ roughly corresponds to how good a match is. We used the largest cluster of keypoints output from the Hough transform as the final match, accepting the match if there were more than 3 points in the cluster.

4.2. **SVM-based supplement.** Single-frame-based approaches to facial recognition on video often suffer from sporadic incorrect recognitions due to bad frames or deficiencies of the facial recognition algorithm. We therefore developed an algorithm based on SVMs that predicts, based on information available about previous frames, whether the current frame should be labelled as a match for a given person l.

The input to the SVM is a measure of confidence that the person appeared in each of C of the previous frames. We used a C of 15, corresponding to one second of video in our test set. The measure of confidence was exactly the size of the largest cluster of SIFT keypoints described the the basic facial recognition algorithm. The SVM thus made use of the trained SIFT-based face recognizer to perform its on training. In addition to the additional information used from video data, this algorithm also replaced the hand-set 3-point threshold with a value determined by the SVM to improve performance.

4.3. **Implementation.** SIFT features were extracted using the SIFT++ library (Vedaldi, 2006) and a script written to analyze its output. For SVMs, LIBSVM (Chang, Lin, 2008) was used in combination with a Common Lisp wrapper library (Melis, 2008). Image formats were interchanges using ImageMagick.

All other algorithms related to image processing and machine learning were written by the author.

5. Testing

First we tested our facial recognition algorithm against the AT&T face database, which is a database of frontal face images. There are 10 images each of 40 subjects. We only tested with 10 subjects due to the $O(nm)$ complexity of our unoptimized implementation.

We then tested the same algorithm against a self-made image database. Our database, unlike those commonly available, contains video data of subjects to accompany frontal facial images. The database contains three subjects, each with ten frontal facial images and two to 3 ten-second videos. The individual video frames are labelled for whether they contain the person in them or not.

5.1. **Results.** Our SIFT-based facial recognizer was able to achieve an accuracies of 94.2% and 70.1% by training on only 2 images per subject on the AT&T and in-house databases respectively.

The 70.1% accuracy on our own database is based on using the facial recognizer on each frame of video for each participant. When we used the SVM-supplemented facial recognizer on the video frames, we were able to achieve a higher accuracy rate of 82.3%. Most of the observed gains were due to smoothing of poor earlier results.

6. References

Belhumeur, Hespanha, et al. ”Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection.” 1997

Chang, Chih-Chung; Lin, Chih-Jen. 2008. http://www.csie.ntu.edu.tw/ cjlin/libsvm/
Melis, Gabor. cl-libsvm, 2008. http://www.cliki.net/cl-libsvm

Turk, M.; Pentland, A. ”Eigenfaces for Recognition.” 1991
Turk,
Vedaldi, Andrea. SIFT++, 2006. http://vision.ucla.edu/ vedaldi/code/siftpp/siftpp.html

Viola, Paul; Jones, Michael J. ”Robust Real-Time Face Detection.” 2003