ViFaI: A trained video face indexing scheme

Harsh Nayyar
hnayyar@stanford.edu

Audrey Wei
awei1001@stanford.edu

1. Introduction

With the increasing prominence of inexpensive video recording devices (e.g., digital camcorders and video recording smartphones), the average user's video collection today is increasing rapidly. With this development, there arises a natural desire to rapidly access a subset of one's collection of videos. The solution to this problem requires an effective video indexing scheme. In particular, we must be able to easily process a video to extract such indexes.

Today, there also exist large sets of labeled (tagged) face images. One important example is an individual's Facebook profile. Such a set of of tagged images of one's self, family, friends, and colleagues represents an extremely valuable potential training set.

2. Problem Statement

Use a labeled (tagged) training set of face images to extract relevant indexes from a collection of videos, and use these indexes to answer boolean queries of the form: "clips with 'Person 1' OP1 'Person 2' OP2 ... OP(N-1) 'Person N'", where 'Person N' corresponds to a training label and OPN is a boolean operand such as AND, OR, NOT, XOR, and so on.

3. Proposed Scheme

In this section, we outline our proposed scheme to address the problem we posed in the previous section.

For the purposes of this work, we define an *index* as follows: (person, video id, frame #).

We subdivide the problem into two key phases, the first "off-line" executed once (or upon additional data) and the second "on-line" phase initiated on each query.

We first outline Phase 1 (the "off-line" phase):

1. Use the labeled training set plus an additional set of 'other' faces to compute the Fisher Linear Discriminant (FLD) [1].
2. Project the training data onto the space defined by the eigenvectors returned by the FLD, then train a classifier (first nearest neighbour, then something more sophisticated like SVM if required) using the projected training set.
3. Iterate through each video, detecting faces [2] and adding an index if the detected face corresponds to one of the labeled classes from the previous step.

Now we outline Phase 2 (the "on-line" phase):

1. Key the indexes on their video id.
2. For each video, evaluate the boolean query for the set of corresponding indexes.
3. Keep videos for which the boolean query evaluates true, and discard those for which it evaluates false.

4. Evaluation Methodology

Our first requirement is a labeled (tagged) training set and an 'other' set. We propose obtaining the former using several friends from Facebook. We have identified several face sets available at [3] that we may select and use for our 'other' category.

We must also obtain a video collection. We will do so using an iPhone 4 as a representative popular video capture device. We will obtain a set of videos that will allow us to effectively test our query space. For example, we will collect a video with Person A and Person B, another with Person A and Person C, another with Person D, and so on. This will allow us to evaluate queries such as A & B, A, or NOT(C).

Obtaining the ground truth for the videos is relatively straight-forward and can be accomplished mechanically.

References

[1] P.N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. "Eigenfaces vs. Fisherfaces: Recognition using class specic linear projection," *IEEE Trans. Patt. Anal. Mach. Intell* 19, 711 720, 1997

[2] P. Viola and M. Jones, "Robust Real Time Object Detection, *IEEE ICCV Workshop Statistical and Computational Theories of Vision*, July 2001.

[3] Face Recognition Homepage. http://www.face-rec.org/databases/