Tagging face in images

Mehul Parsana
Advisor: Dr. Chiranjib Bhattacharya
13-11-2006

Abstract

Digitized multimedia data do not have any structure in raw form and so is not amenable for information extraction by a machine. Consider the problem of creating structure over images by tagging it with labels which can easily be processed by machines and meaningful to humans. Faces are the most important part of images, we consider problem of tagging them. As an initial step, various face recognition methods including face detection, feature extraction and face recognition are studied and implemented. We are also studying the applicability of descriptive local features of person’s face for face tagging.

1 Introduction

Internet’s rapid growth has resulted in a huge increase in the amount of information generated and shared. Large portion of this data belongs to unstructured data like webtext, images and videos. Unstructured data can be seen as mass of computerized information which does not have structure in itself or in raw form. It cannot be used for information extraction by a machine. New techniques for search, classification and knowledge discovery are required to process these data. The problem of processing unstructured text is reasonably well understood. However for images and videos, there are no solution which can scale up to requirement. Internet site like Flickr1, Yahoo! Album and Google are having millions images and videos. With development of cheaper and higher quality digital cameras, increase in number of personal images on disk of computer is unavoidable. But there are no efficient ways to index these images, except manual tagging which is expensive. Criteria like date, time etc. are not enough to build meaningful search. We are trying to automate tagging process. So search can be made more efficient using these tags. Human faces are most important part of images, first we worked on tagging of human faces present in images, so queries like search person’s image or who are present in image can be answered. Immediate application of this work can be searching and managing family albums.

Problem Definition Given a set of images and set of tagged faces, build a system which can locate face given in labeled set from images and tag image appropriately. We try to construct system which can localized face and map it relevant tag/label.

Tagging faces require face detection and matching. We use current approaches in face detection and recognition literature. In section 2, we give survey of some representative approaches in literature. We have studied and implemented some of the approaches. In section 3, we give detail with results. Conclusion and future work is presented section 4.

2 Literature Survey

The solution to the problem of tagging faces involves segmentation of faces (face detection) from cluttered scenes, feature extraction from the face regions, recognition or verification as shown in figure(1). Research on automatic machine recognition of faces really started in the 1970s by Kelley[1] and after the seminal work of Kanade 1973. Over the past 30 years extensive research has been conducted by psychophysicists, neuroscientists, and engineers on various aspects of face recognition by humans and machines.

However the multitude of sources for face images under varying conditions makes the location and recognition of faces in images a complex problem. These variations differ as broadly as the sources from which the images were obtained. Factors such as visibility and hence image quality and resolution play a large role in the introduction of noise. Variation in illumination conditions is another key factor in facial image acquisition. As all images are a product of a

1www.flickr.com
Lambertian reflection of light on an object, changes in the type and direction of the illuminating source greatly affects the objects appearance. In depth orientation (pose) changes of the subject are also very important variations along with expression changes, aging, partial occlusions and general changes of the subjects appearance. Over the past 15 years, research has focused on how to make face recognition systems fully automatic by tackling problems such as localization of a face in a given image or video clip and extraction of features such as eyes, mouth, etc.

According to T. Poggio [2], face detection methods are generally classified as knowledge based, feature invariant based, template matching and statistical appearance based methods also discussed in Liu[3], H. Schneiderman [4].

Turk and Pentland[6] and Belhumeur et al.[7] used distance thresholding in projected (face) space for face detection. Approach is discussed in section 3.2 in detail. Parallel to feature-based and template-matching methods by Wiskottet al.[9], image and appearance based methods are suggested. Under this methods, Rowley et al.[12] experimented neural network based face detection. Training data consists of face objects at different scales, with different poses, orientations and under various illuminating conditions and showed through most extensive training, computers can be quite good at detecting faces.

![Figure 1: overview of system](image)

H. Schneiderman and T. Kanade[4] used statistical method for face/object detection. Decomposition is performed on appearance in space, frequency and orientation using wavelet transform. Subset of quantized wavelet coefficients is used as feature for detection. Three levels of wavelet are calculated and statistical classifier like naive bayes is used and shown good result.

Boosting of haar features have been proposed by Viola and Jones[11]. It showed good performance in accuracy even in cluttered background as well in time. Implementation and results are discussed in section 3.

According to Wang[5], facial feature based methods have difficulties when that the face component, such as eye and mouth, and facial expression are blurred. Under such condition, the methods based on the whole face appearance are more robust to noise and face variance. Face poses was divided in seven view and each pose-view is again divided in three position like vertical, up and down. Separate SVM is trained for each category and train over set for iterations. In each iteration false positives are again feeded as negative image. On this system, good detection rate was shown.

Meanwhile, significant advances have been made in the design of classifiers for successful face recognition. Although 3-D recognition techniques are also available in literature, Here we will limit ourselves to 2-D recognition methods.

2-D face recognition methods mainly can be classified as shown in figure 2. Categorization based feature set are • Holistic matching methods/Appearance based methods • Local/Feature-based matching methods • Hybrid methods

In following subsection, we discussed all three in brief.

**Holistic matching methods/ Appearance based methods**

These methods use the whole face region as the raw input to a recognition system. Eigenfaces [Turk and Pentland, 1991] and Fisherfaces [Belhumeur et al. 1997; Zhao et al. 1998[8]] have proved to be effective in experiments. Both methods are discussed in section[3] with implementation. With these other methods like eigenpictures by Kirby and Sirovich[13], eigenedges, eigenhill by Gokemen[14], Probabilistic eigenfaces, Two-class problem with probabilistic measure, by Moghaddam and Pentland[17] are proposed which are based on principal component analysis. Eigenhill approach is showed to perform well under change in expression, lighting conditions
over eigenfaces, eigenedges by removing illumination
effects.
B. Park et al.[15] proposed Face-ARG, Attribute
Relation Graph over edges, and showed good results
in presence of occlusion, variation in expression.
Example of Face-ARG is shown in figure 3.
In SVM-based holistic approach by Phillips [18], face
images are directly given to support vector machine
and trained SVM for multiclass classification problem
with different kernels.

Local/Feature-based matching methods
In these methods, local features such as the eyes, nose,
and mouth are first extracted and their locations and
local statistics (geometric and/or appearance) are fed
into a structural classifier. Statistical method pro-
posed by T. Kanade[4], is part of feature based match-
ing.
One of the most successful of these systems is the
Elastic Bunch Graph Matching (EBGM) system by
Okada et al. [19] and Wiskott et al. 1997[9], which is
based on dynamic link architecture(DLA) proposed by
Buhmann et al. [21] and further discussed by Lades
et al.[20]. Wavelets, especially Gabor wavelets, play
a building block role for facial representation in these
graph matching methods. A typical local feature rep-
resentation consists of wavelet coefficients for different
scales and rotations based on fixed wavelet bases
called jets in Okada et al.[19]). These locally esti-
mated wavelet coefficients are robust to illumination
change, translation, distortion,rotation and scaling.

\[
g(x, y : u_0, v_0) = \exp\left(-\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2}\right) + 2\pi i(u_0x + v_0y)
\]
\[
G(u, v) = \exp\left(-2\pi^2(\sigma_x^2(u - u_0)^2 + \sigma_y^2(v - v_0)^2)\right)
\]

The basic 2D Gabor function and its Fourier trans-
form are shown in equation(1). Where \(\sigma_x\) and \(\sigma_y\) re-
represent the spatial widths of the Gaussian and \((u_0, v_0)\)
is the frequency of the complex sinusoid.
Hybrid methods

As the human perception system uses both local features and the whole face region to recognize a face, a machine recognition system should use both. These methods could potentially offer the best of the two types of methods. For example, the modular eigenfaces approach [Pentland et al. 1994] uses both global eigenfaces and local eigenfeatures. In mugshot applications, usually a frontal and a side view of a person are available; in some other applications, more than two views may be appropriate. One can take two approaches to handling images from multiple views. The first approach pools all the images and constructs a set of eigenfaces that represent all the images from all the views. The other approach uses separate eigenspaces for different views, so that the collection of images taken from each view has its own eigenspace. The second approach, known as view-based eigenspaces, performs better. The concept of eigenfaces can be extended to eigenfeatures, such as eigennose, eigenmouth, etc. Using a limited set of images (45 persons, two views per person, with different facial expressions such as neutral vs. smiling), recognition performance as a function of the number of eigenvectors was measured for eigenfaces only and for the combined representation. For lower-order spaces, the eigenfeatures performed better than the eigenfaces [Pentland et al. 1994]; when the combined set was used, only marginal improvement was obtained. These experiments support the claim that feature-based mechanisms may be useful when gross variations are present in the input images.

A active appearance model based method for face recognition was presented by T.Coots et al.,[16] Where appearance model is an integrated statistical model which combines a model of shape variation with a model of the appearance variations in a shape-normalized frame. An AAM contains a statistical model if the shape and gray-level appearance of the object of interest which can generalize to almost any valid example. Matching to an image involves finding model parameters which minimize the difference between the image and a synthesized model example projected into the image. Example of fitting AAM is shown in figure[4].

Other categorization is based on processing of face or structure data. (1) Linear model: Processing of face data and structures assumed of containing linear relations. Typically used for dimensionality reduction. As shown figure 2, all methods like fisher faces, eigenfaces, modular eigenfaces etc. are linear model.

(2) Nonlinear model: Processing of face data and structures which have been projected into some non-linear feature space. It is assumed that nonlinear relations in the data become linear relations in the feature space. Methods like Non-linear Neural Networks, Kernel LDA, Kernel PCA, Kernel Features etc are non linear/ kernalized model.

3 Approaches Implemented

3.1 Face detection

Boosting haar features using LDA as weak hypothesis: Viola and jones[24] showed importance of rectangular haar like features for face and object fention. But
haar feature varies with different pattern, size, position and it can only give block difference. Discriminant capabilities is limited of these features is limited, boosting helps to get good classifier using these features. Linear FDA is used to select weak hypothesis at each boosting iteration. As training real time dataset is taken of all frontal face images. Training is performed in two stages.

1. First stage classifier is used to learn skin color model for human faces using boosted stump.
2. Second stage classifier is trained over haar features. To detect face from images, image was given to stage 1. It will give possible rectangular candidate region for face using skin color. Then subcandidate generation is performed and one by one this subcandidate is given to stage 2 which classify it as face or not face. Conclusion: Results are good. But this detection can not give alignment of face so difficulty for recognition increases. One result is shown in figure. Where big boxes shows output of stage1 and small boxes are detected face by stage2.

3.2 Face Recognition Technique

1. Eigenfaces: Idea here is to extract the relevant information in face, encode it as efficiently and compare one face encoding with a databases of models encoded similarity. In mathematical terms, principal components of the distribution of faces or eigenvectors of covariance matrix of set of images. Eigenvectors, shown as images in figure 4, were called as eigenfaces. In order to recognize a face, it is projected into face-space, producing weights that are then used as features in a nearest neighbor classifier that simply finds the minimum euclidean distance between the recognition weights and the training weights.

Tested on UMIST dataset, contains images of 20 peoples.
1. For testing one image per person is selected (20 images) other 545 images are used for testing. Experiment is repeated for 10 iteration. Result is 99.57% average accuracy.
2. Two image per person is selected, other 525 are used for testing. Result is 98.78% average accuracy on 10 experiment.
3. Number of testing images is same as experiment 2, but from training 15 images per person are selected for training. Result is 91% accuracy.

Conclusion: Eigenfaces works well with densely and uniformly sampled inputs.

2. Fisherface: Belheumeur[7] solved problem in Eigenfaces that Scatter is maximized not due only to Between-class Scatter but Within-class Scatter too using fisher discriminant analysis, analyze eigenvector of $S_b^{-1}S_w$ and called fisherfaces. Where $S_b$ is between-class scatter matrix and $S_w$ is within-class scatter matrix.

For experiment, 20 recognition faces (one for each person) were randomly picked from the database, then 60 more images were used as training faces. Three training faces were picked for each person: a frontal, side, and 45-degree view. As result is 80% classification accuracy can be achieved on average. While with eigenface could give only 66% in same setup.

Conclusion: Fisherface is performed well compare to eigenfaces under varying light and pose condition even less training samples.
3.3 Tagging using local descriptors/local features

Previously we showed global approaches need large and densely sampled training set. As well global approaches are more sensitive to change in expression and illumination condition. We tried local feature based approach. We discussed local descriptors and its results in this section. There are wide range of local descriptors are available in image domain. We tried SIFT and evaluated it for face.

Scale Invariant Feature Transform

David Lowe, 2004 [23] presented a method to extract distinctive invariant features from images. He named them Scale Invariant Feature Transform (SIFT). This particular type of features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise. Following are the major stages of computation used to generate the set of image features:

- Scale-space extrema detection

The first stage of computation searches over all scales and image locations. It is implemented efficiently by using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation. Given a Gaussian-blurred image,

\[ L(x, y, \sigma) = G(x, y, \sigma) \ast I(x, y) \]  

(2)

where \( I(x, y) \) is given image and \( G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \). To efficiently detect stable keypoint locations in scale space, the method proposed in [23] could be used. It makes use of the scale-space extrema in the difference-of-Gaussian function convolved with the image, \( D(x, y, \sigma) \), which can be computed from the difference of two nearby scales separated by a constant multiplicative factor \( k \):

\[ D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \ast I(x, y) \]

\[ = L(x, y, k\sigma) - L(x, y, \sigma) \]

Interest points (called keypoints in the SIFT framework) are identified as local maxima or minima of the DoG images across scales. Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales. If the pixel is a local maximum or minimum, it is selected as a candidate keypoint.

- Accurate keypoint localization: At each candidate location, a detailed model is fit to determine location and scale. Keypoints are selected based on measures of their stability. Once a keypoint candidate has been found by comparing a pixel to its neighbors, the next step is to perform a detailed fit to the nearby data for location, scale, and ratio of principal curvatures. This information allows points to be rejected when having low contrast (and therefore be sensitive to noise) or are poorly localized along an edge.

- Orientation assignment: One or more orientations are assigned to each keypoint location based on local image gradient directions. To determine the keypoint orientation, a gradient orientation histogram is computed in the neighborhood of the keypoint (using the Gaussian image at the closest scale to the keypoints scale). The contribution of each neighboring pixel is
weighted by the gradient magnitude and a Gaussian window with a \( \sigma \) that is 1.5 times the scale of the keypoint. Peaks in the histogram correspond to dominant orientations. A separate keypoint is created for the direction corresponding to the histogram maximum, and any other direction within 80% of the maximum value. All the properties of the keypoint are measured relative to the keypoint orientation, this provides invariance to rotation.

- Keypoint descriptor: The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination. Once a keypoint orientation has been selected, the feature descriptor is computed as a set of orientation histograms on 4 x 4 pixel neighborhoods. The orientation histograms are relative to the keypoint orientation, the orientation data comes from the Gaussian image closest in scale to the keypoints scale. Just like before, the contribution of each pixel is weighted by the gradient magnitude, and by a Gaussian with \( \sigma \) 1.5 times the scale of the keypoint. Histograms contain 8 bins each, and each descriptor contains an array of 4 histograms around the keypoint. This leads to a SIFT feature vector with \( 4 \times 4 \times 8 = 128 \) elements. This vector is normalized to enhance invariance to changes in illumination. In this way the descriptor is invariant to affine changes in illumination.
- This descriptor was used object detection. Here we tried to use it for face tagging. Results are quite good. Main advantage of this approach is no need of localization of face. So it can speedup whole operation. Figure 9 is face with descriptor. Figure 10 is shown as example using nearest neighbour approach.

4 Conclusions and Future Work

Standard face recognition methods can be used to solve the face tagging problem. But the given problem is simpler than general face recognition problem. So we use simpler approach of finding similarities. For that landmarks of some person’s face are enough to help define similarity search. Good local features are sufficient which are transformation invariants. If such local point can be given, easy to search then tagging process can be fast. Using local features as points in point set we formulate as machine learning problem of learning over point sets. Alignments over pointset are used to define similarity measure. We have presented nearest neighbour as proof of concept. In future, we would explore different similarity measure including alignment methods and different learning methods available in machine learning literature.

References


[5] PENG WANG AND QIANG JI. Multi-View Face Detection under Complex Scene based on Combined SVMs. ICPR, 2004


