Semi-supervised learning of facial attributes in video

Neva Cherniavsky[1], Ivan Laptev[1], Josef Sivic[1], Andrew Zisserman[1,2]

[1]INRIA, WILLOW, Laboratoire d'Informatique de l'Ecole Normale Supérieure,
ENS/INRIA/CNRS UMR 8548
[2]Dept. of Engineering Science, University of Oxford

Abstract. In this work we investigate a weakly-supervised approach to learning facial attributes of humans in video. Given a small set of images labeled with attributes and a much larger unlabeled set of video tracks, we train a classifier to recognize these attributes in video data. We make two contributions. First, we show that training on video data improves classification performance over training on images alone. Second, and more significantly, we show that tracks in video provide a natural mechanism for generalizing training data – in this case to new poses, lighting conditions and expressions. The advantage of our method is demonstrated on the classification of gender and age attributes in the movie "Love, Actually". We show that the semi-supervised approach adds a significant performance boost, for example for gender increasing average precision from 0.75 on static images alone to 0.85.

1 Introduction

Classification of people according to their attributes is an area of active research, both as a first step in the larger problem of image search and classification on identity [1, 2], and as a goal in and of itself. For example, cultural sociologists are interested in measuring the evolution over time of the characterization of gender [3] in TV and movies. Video analysis for these purposes currently requires hours of tedious manual labeling, rendering large-scale experiments infeasible. Automating the detection and classification of human traits in video will potentially increase the quantity and diversity of experimental data.

Our goal in this paper is to learn and classify human attributes in video. The idea we explore is that video tracks provide a virtually free and limitless source of training data, since many human attributes, e.g. gender, race, age, hair colour, are unchanged over the course of a track. For example, if we can correctly determine the gender of a face in a video face-track, we can then apply that label to the rest of frames within the track, including faces that would normally be difficult to classify. We can thus take advantage of the full variation in poses and viewpoints.

It might be thought that videos could be classified by training a classifier on photos of faces, for example from flickr. However, as we show quantitatively, training on still image data does not generalize well to video data. Although image data sets contain a wide variety of distinct faces, the style of photo is often similar: the subject is usually facing the camera and smiling, and there are no strong shadows or unusual lighting conditions. Different poses or viewpoints are rare. In contrast, video, such as a feature length movie, often contains only a few distinct subjects, but their faces contain a wide variety of expressions, poses, and viewpoint (see Figure 1).

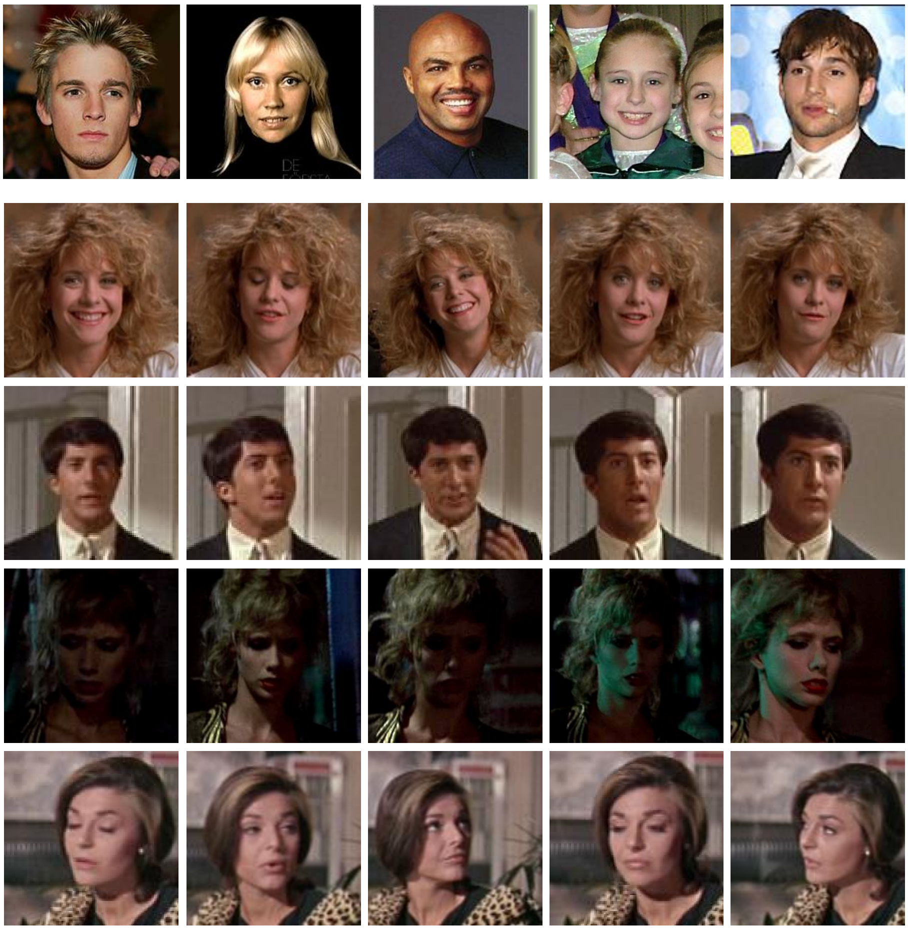

Fig. 1. Labeled still images from the FaceTracer database (top row) versus faces from video (four bottom rows). Faces in video contain more variety of expression, lighting, and viewpoint.

This is a shame as labeled still image training data for human attributes is readily available from several public databases [1, 2], and also can be obtained automatically. For example, many attributes can be obtained by crawling the descriptions and photos available on dating web sites, or, in the case of gender, by using web image search engines for common male and female names [4].

Previous work on classifying facial attributes [1, 2] has used strong supervision in the form of fully labeled still image datasets. Here we also train from fully labeled still images but show that results can be significantly improved by incorporating a large pool of additional *unlabeled* videos. To achieve that we turn to semi-supervised learning [5, 6] and in particular self-training, where labels are hypothesized for the most confidently classified unlabeled examples. These predicted labels are then considered as additional labeled training data. Whilst very simple, this strategy may suffer from limited generalization [5] as only confidently classified unlabeled data lying far from the decision boundary is considered for labeling. To overcome this problem, we employ the video tracks to provide additional generalization over pose, lighting and expression. In particular, we select tracks which contain faces that are very confidently classified, and then use other faces in the track with a low classification score in order to provide training examples close to the decision boundary. We illustrate this approach here by learning a classifier for the attributes gender and age from faces.

This strategy contrasts with others who have used tracks for providing training data. The closest work is that of Yan *et al.* [7] where tracks in video are considered as constraints in support vector machine classifier training, forcing the same classification output for all detections along the track. Others have used video for person/object recognition and retrieval [4, 8–10], but in these works detections within a track are used only as additional labels or query examples and no semi-supervised learning is performed.

The paper is organized as follows: section 2 describes the train and test datasets and attribute annotation; section 3 reviews the video processing pipeline used to obtain tracks and a face descriptor vector for each face in the track; section 4 then compares classification performance for training on still images alone to also training on faces from video tracks; the semi-supervised approach is developed in section 5 and the performance investigated fully in section 6.

2 Databases and attributes

We train the attribute classifier from labeled face images and unlabeled video tracks from Hollywood movies. It is then tested on tracks from a disjoint movie. Here we describe the data, ground truth attribute annotation and performance measure.

2.1 Labeled image database

The labeled still images are obtained from the FaceTracer database, available on the web [1]. The FaceTracer database consists of 15,000 images downloaded from the Internet. The images were collected and labeled by researchers using a standard search engine, and contain both amateur photos and professional photos of celebrities. Some subjects are represented multiple times. A subset of these images are labeled by at least one of several attributes. The choices for each attribute were determined by the FaceTracer creators. The attributes that we use are gender (male/female) and age (baby/child/youth/middle-aged/senior). We group age into two supersets, *young and* old; the first containing baby/child/youth and the second containing middle-aged and senior. There are 303 labeled images for gender and 208 for age. The top row of figure 1 shows some typical images from the FaceTracer database.

2.2 Unlabeled track database

The unlabeled movie set consists of tracks from five distinct Hollywood movies: *Roman Holiday*, a black and white feature made in 1953; *The Graduate*, a coming-of-age drama made in 1967; *Desperately Seeking Susan*, a thriller/romantic comedy made in 1985; *When Harry Met Sally*, a romantic comedy with a small cast made in 1989; and *Insomnia*, a thriller made in 2002. The five movies have no overlapping actors with each other or with the test set. We chose a wide variety of eras and genres so that our work would be applicable across different time periods for sociological research. After face detection, tracking, and filtering (described in section 3), there are a total of 3,661 tracks. Of these, 43.6% are female and 57.4% are male, and 35.3% are young and 65.7% are old. However, the number of distinct people is only around 200.

2.3 Test set

The test set consists of 1,708 tracks from the movie *Love, Actually*, produced in 2003. This movie, made up of several interweaving story lines, contains a wide variety of characters. The gender distribution in the test set is 35.8% female and the age distribution is 33.1% young.

2.4 Ground truth annotation and performance measure

Each track is represented by the face with the highest facial feature score (see below) and this is annotated to provide the ground truth. The annotation is positive (female), negative (male) or ambiguous. Tracks labeled ambiguous are not reported for the training or test sets. Of course, in the semi-supervised learning we do not use the ground truth track labels on the 'unlabeled' Hollywood tracks, but these are required so that the performance of the learning can be assessed.

Average precision (AP) is used as the performance measure. So, for example, in gender classification the AP is unity if all the females are returned first before the males.

3 Video processing

We review here the video processing pipeline of Everingham *et al.* [8] and Sivic *et al.* [11], which we adopt without change. Section 3.3 describes the kernel and learning framework and here we differ from the approach of [11].

3.1 Face extraction and tracking

Faces in the video are detected using the OpenCV [12] implementation of the Viola and Jones face detector [13]. A color histogram-based shot detector is run, and the faces are then grouped into tracks on a shot-by-shot basis using the Kanade-Lucas-Tomasi tracker [14]. The tracker is seeded with feature points from every face detection, and it is run in both the forward and backward directions to ensure that good feature points from late in the track help form connections between faces. The output point tracks are aggregated into face tracks by counting, for each pair of faces, the number of point tracks in common and normalizing by the total number of point tracks. Intra-frame detected face tracks are then merged using agglomerative clustering based on the overlap of the faces. Aggregating point tracks to create face tracks in this manner is robust in that it can handle missing detections and it does not make any false connections between people of different identity (no drift). For example, successful tracking of nearly 45,000 face detections with no mistakes is reported in [11]. This property is important here as we rely on the tracker to provide additional noise-free training data during the semi-supervised learning.

3.2 Facial descriptors

The output of the face detector exhibits some noise over location and scale. Facial feature (eye, mouth, etc.) localization is therefore useful as a means to better align pairs of faces, and subsequently extract descriptors based on the facial features after a viewpoint normalization. We follow the approach of [8], which combines a discriminative model of feature appearance in the form of boosted classifiers using Haar-like features [13] with a generative model of feature locations. The location model uses a mixture of Gaussians, where each mixture component has a tree-structured covariance such that efficient inference for the MAP locations can be performed using the generalized distance transform [15].

Following [8] we also use the score (log-probability) of the joint facial feature model to remove false positive faces. Face tracks are then considered unreliable and removed if they are short (less than 5 frames) or if their minimal facial feature score is low (less than -15). We experimentally determined that these thresholds provided a good balance between eliminating false positives and obtaining enough true positives. We additionally use a higher threshold on facial feature score for the unlabeled track data. We only include a face from a true positive track in the classification if its facial feature confidence score is greater than zero.

A pixel-wise descriptor of the local appearance around a facial feature is extracted by taking the vector of pixels in the patch and normalizing (so that the intensity has zero mean and unit variance) to obtain local photometric invariance. The descriptor for each face in the track is then formed by concatenating descriptors extracted around 13 frontal facial feature locations, e.g. the corners and centers of the eyes, nose and mouth. This results in a 1,937 dimensional descriptor for each face in the track.

3.3 Kernel and classifier

To measure similarity between faces, we use the intersection kernel, as originally described by [16]. Given two d-dimensional face descriptor vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ the intersection kernel is given by

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} \min(x_i, y_i). \tag{1}$$

We use the explicit feature map of Maji and Berg [17] to approximate the intersection kernel. This represents each dimension of the descriptor vector by 20 components. Thus the 1,937 original descriptor vector is represented as a 38,740 dimensional vector. A linear support vector machine is then used as a classifier using the efficient LIBLINEAR package [18].

We have compared this method with the min-min distance SVM classifier used by Sivic *et al.* [11]. The performances are very similar, but the intersection kernel and linear SVM is orders of magnitude faster at a cost of storing larger feature vectors.

4 Images vs tracks

As previously illustrated in Figure 1, faces in video typically have greater variability in pose, lighting and expression than those in still images. Here we investigate the effect of this on face attribute classification applied to video data. Gender classifiers are trained from different data sources, namely, (i) faces from the still image set, (ii) faces with high facial feature scores from the tracks of the training videos and, (iii) all faces from the tracks of the training videos. The performance is then measured on the test set. To compare the classifiers against each other for the same number of annotations, we count using all the faces in the track (case (iii)) as a single annotation. For each labeled set size, three subsets of the training data are randomly chosen in order to obtain the mean and standard deviation of the AP. Results for different numbers of annotations over the three different training methods are shown in Fig. 2.

We observe that the use of video training data results in improved gender classification compared to the classifier learned from the same amounts of labeled

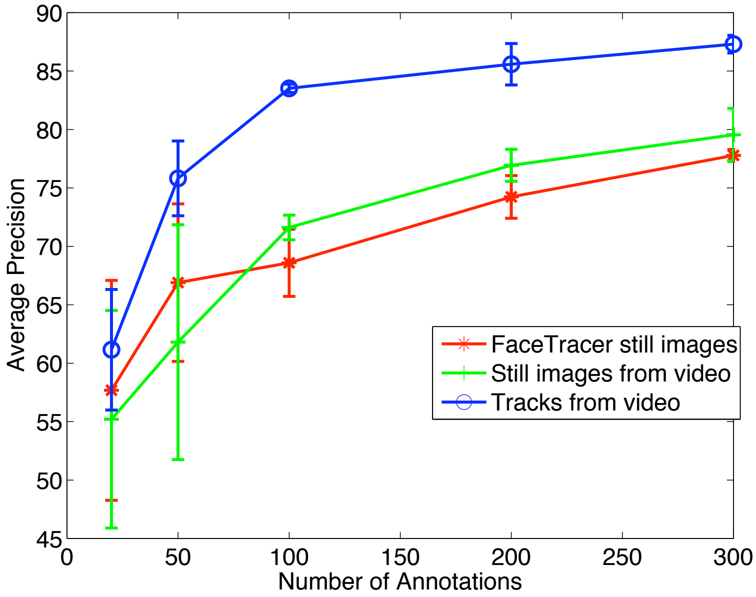

Fig. 2. Comparison of training on still images or video tracks for gender classification. Average precision of the classifier is measured on the test set.

faces in still images. For low numbers of labeled images, the variance is quite high (0.18 in the case of 20 annotations), but as the variance decreases, using still images from video (case (ii)) provides a consistent advantage. This confirms our hypothesized importance of video data for training. Moreover, by automatically propagating manual labels among the faces in the track and, in this way, generating additional training samples (case (iii)), we are able to significantly increase classification performance further. We conclude that faces within the same track contain non-redundant and highly useful information which helps to improve the classifier.

5 Semi-supervised learning with unlabeled tracks

As we have seen above, training on labeled video track data improves the classification results on video compared to training on still images. We now describe how this can be achieved in a semi-supervised setting where the tracks are unlabeled. We will use the attribute gender as our running example.

Suppose we have trained an initial classifier on the fully labeled still images. One natural way to try to include unlabeled video training data is to hypothesize the track labels on the basis of the initial classifier, e.g. if all the faces in the track are classified as positive. However, there are two problems with this approach. The first is that the initial classifier can be quite inaccurate on video, leading us to hypothesize the wrong labels and thus add noise to the training data. The

```
Input: Labeled face still images. Unlabeled video face tracks.

1. Train initial classifier from fully labeled still images.

2. Iteratively add training data from the unlabeled videos:
   (a) Classify all faces in all unlabeled face tracks.
   (b) In each track, $i$, select the closest face, $minF_i$, and the furthest face,
       $maxF_i$, from the current decision boundary.
   (c) Consider only tracks whose maximum classification score is in the top
       10% of tracks.
   (d) Use $minF_i$ scores to choose a track from each movie. Choose the track
       with the smallest $minF_i$ for each of the $m$ movies.
   (e) From the $m$ selected face tracks, add the $maxF_i$ face and the $minF_i$
       face to the pool of labeled faces if it is not too closely correlated
       (normalized correlation between face descriptors below 0.9) to faces
       that have been previously added.
   (f) Re-train the classifier from the extended labeled set.

Output: Classifier trained from labeled still image faces and video face tracks
with hypothesized labels.
```

Fig. 3. The algorithm for semi-supervised learning from labeled face still images and unlabeled video face tracks.

second, more pernicious problem, is that the initial classifier does well on the same types of faces, and that adding this data in as labeled examples might overtrain, rather than generalize, the classifier.

For example, often the initial classifier can determine with high accuracy that Audrey Hepburn is a female; as there are many tracks of Audrey Hepburn in the movie *Roman Holiday*, the most confident tracks according to the classifier contain a disproportionate number of different instances of Audrey Hepburn; hypothesizing that Audrey Hepburn is a female thus does not actually aide in improving the classifier at all, but rather skews the hyperplane towards more Hepburn-like female faces. The classifier then has a harder time recognizing that Cate Blanchett, e.g., is female.

Solving these two problems ((i) avoiding noisy training data; and (ii) avoiding correlation in the training data) is crucial to achieving success in the semi-supervised approach. To solve the first problem we only consider *confident tracks*, i.e. tracks with the most confident classified face in the top 10% of tracks. This threshold was selected manually, but could be chosen automatically on a small separate validation set. In addition we require that the facial feature score for that face is above a conservative threshold – this ensures that for the face chosen in the track the facial features are well detected. The key idea then is to obtain generalization by adding the face in a positive track *minimizing* the classifier score, as well as the face maximizing the classifier score. The minimum, i.e. the face from the track closest to the decision boundary, is typically a face that

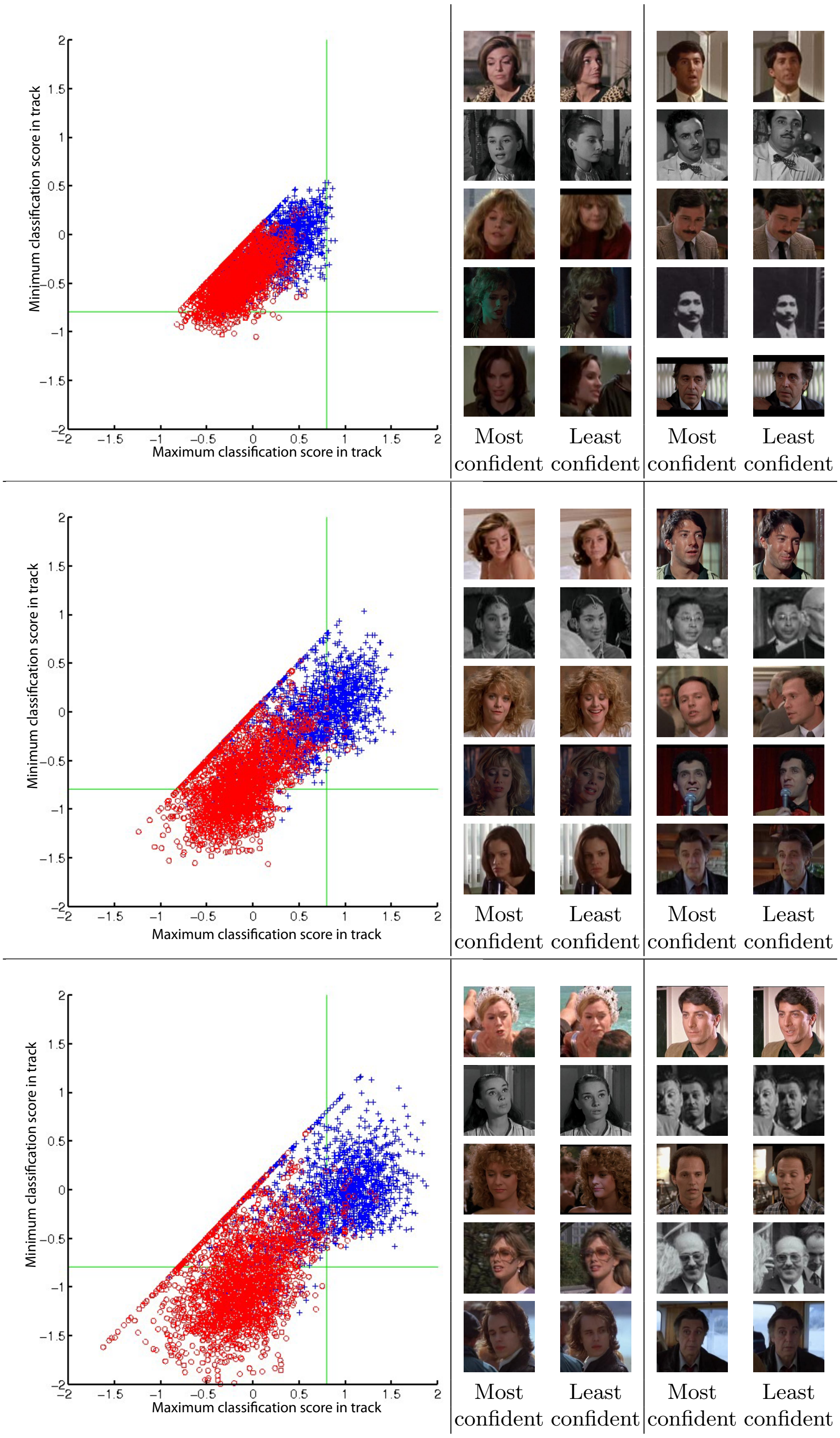

Fig. 4. Maximally confident faces versus minimally confident faces for each track. Blue crosses are positive (female) tracks and red circles are negative (male) tracks. The green lines indicate thresholds beyond which all the tracks have the same label. Top row shows an early iteration, middle row a midway iteration, and bottom row the final iteration. Note the increase in separation and spread of the classes as the iterations proceed.

differs significantly from the type of faces in the FaceTracer training still images (e.g. different pose, expression or lighting as in Figure 4) but, as the attribute does not change on a track, it is guaranteed to be positive (provided the max is a positive).

The semi-supervised learning then proceeds in an iterative manner, with at each iteration a number of max/min pairs from the track above a classification threshold being added (for the positives) as well as a number of max/min pairs below a classification threshold (for the negatives). We address the second, correlation, problem by choosing only a few tracks per training movie at each iteration and adding only faces not too closely correlated (normalized correlation between face descriptors below 0.9) to faces that have been previously added. The algorithm is detailed in Figure 3.

Figure 4 illustrates the progress of the algorithm starting from the initial classifier trained on FaceTracer images (the details of the experiment are given in Section 6). Each point on this graph is a track. The minimally confident score in the track is plotted against the maximally confident score. (Some tracks only have a few faces when filtered by facial feature score, which is why there are points along the line x=y.)

The tracks we're most interested in are those whose maximum confidence score is high enough to ensure that the track is correct, but whose minimum confidence score is low. This is the area in the lower right hand corner of the graph. For females, these are faces with maximum score (x axis) high and minimum score low; for males, it's faces with minimum score (y axis) low and maximum score high.

The figure shows three iterations of the algorithm and the corresponding min-max graph and faces chosen. In the beginning, the data is not as well separated, with both the blue and red points largely centered around zero. As the algorithm progresses, the classification scores become more spread out. Note that the points in the plots are color-coded according to the gender for visualization purposes, but the algorithm does not have access to the labels.

The faces beside each graph are the tracks chosen in this iteration; the maximum and minimum face by classification score are added to the classifier. The least confident face is often not frontal – the eyes may be closed, the viewpoint shifted, or the expression changed. The final iteration shows that the classifier does make some mistakes, labeling a male as female and including that (incorrect) data in the next round. But as we show in the next section, this does not seem to too adversely affect the classification. Note also the increase in separation and spread of the classes as the iterations proceed.

6 Experimental results

We report results on the gender and age attributes. In all the experiments the training data is FaceTracer labeled images, the unlabeled tracks are from the Hollywood movies, and the test data is the film 'Love Actually'.

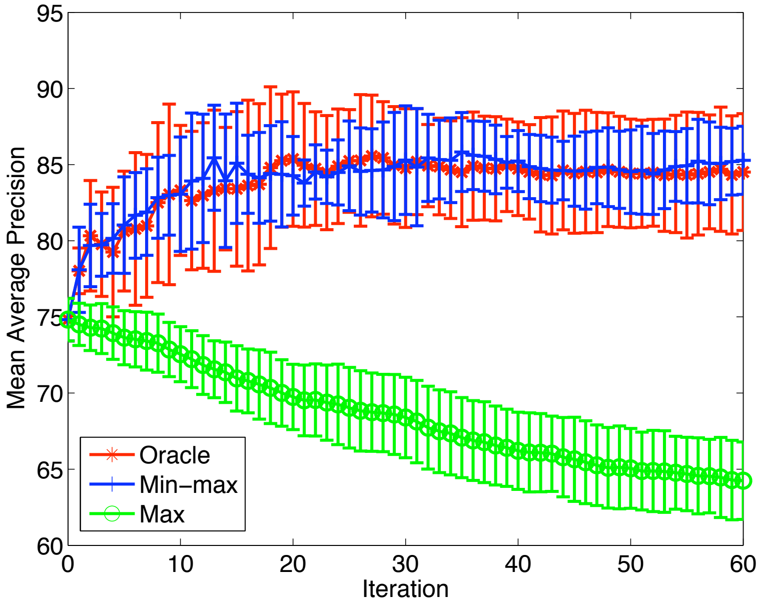

Fig. 5. Results for gender classification on three splits of 200 labelled still images. The leftmost point, at iteration 0, is the mean average precision when no video data is included in the classifier. Our approach, Min-max, performs similarly to the Oracle and outperforms the standard self-training (Max) approach.

We compare our semi-supervised method to two other possible approaches. The first method we test against is the semi-supervised approach of Section 5, modified to select a threshold so that only face tracks with the correct labels are added. We refer to this as the *Oracle* approach. The only difference between the Oracle and actual approach is the possibility of incorrect labels, i.e. noisy supervision, which may have an adverse effect.

Second, we also compare our method to the natural approach of adding the most confident faces at each iteration. Given the classification, this method chooses the face from each movie that is most confidently classified (e.g. most positive for females and most negative for males). We then exclude that track from the pool of unlabeled training data; without exclusion, the classifier chooses the same track at every iteration. We call this method *Max*.

Figures 5 and 6 show the mean and standard deviations for the three methods over 60 iterations of gender classification and age classification. The size of the labeled set is 200 images. Figure 7 shows some examples of the most confidently classified faces after 60 iterations. Our method is able to deal with a wide variety of pose, lighting, and expression, including tracks taken from photos, blurry faces, and partially obscured faces.

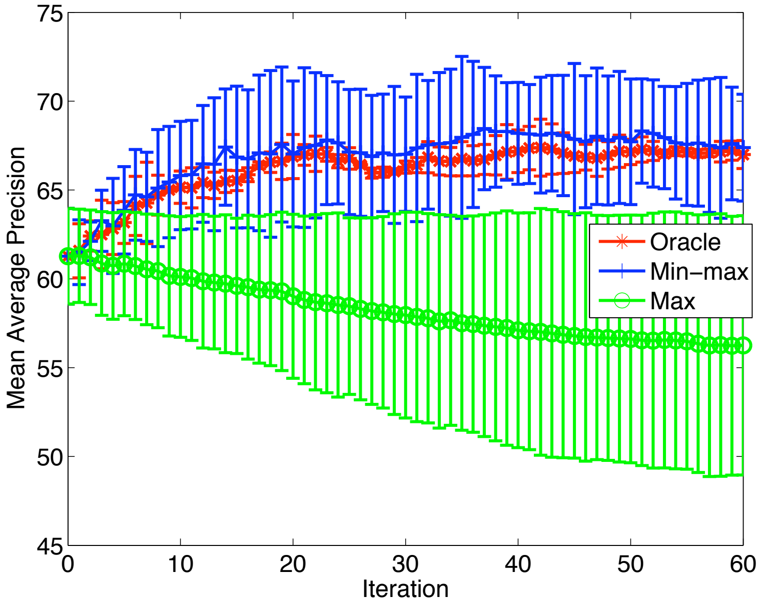

Fig. 6. Results for age classification on three splits of 200 labelled still images. The leftmost point, at iteration 0, is the mean average precision when no video data is included in the classifier. Our approach, Min-max, performs similarly to the Oracle and outperforms the standard self-training (Max) approach.

The Min-max method achieves comparable results to Oracle, improving the mean average precision by almost 10% for gender and 5% for age. Though we observe with Min-max that the classifier does sometimes mislabel data, adding incorrectly classified faces to the training set, this does not seem to worsen performance. Furthermore, in some cases Min-max may outperform Oracle, because the threshold chosen by Oracle could be quite high in some instances. Oracle chooses the threshold so that no faces are misclassified; if there is an outlier, the threshold could be so high as to remove the advantage of obtaining minimum faces from the track.

The intuitive Max method does not perform well. It appears that the classifier overtrains, only improving on faces that it already classifies well. In examining the blue-red min-max graphs (similar to figure 4) for Max, we do not see the gradual separation of the training data into two classes, indicating that the classifier is not improving its performance.

7 Discussion

We have shown that tracks, which are readily available in video data, can be harnessed to provide a natural means of generalization in semi-supervised learning.

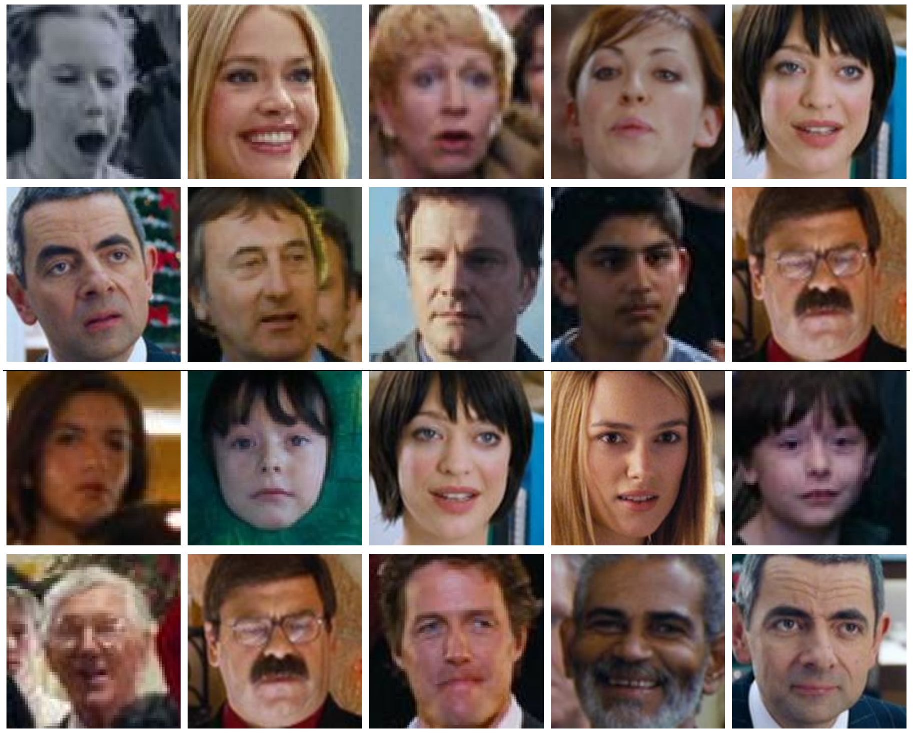

Fig. 7. Examples of most confidently labeled faces at final iteration of the test set. The top two rows are the most confidently classified gender faces (female and male) and the bottom two rows are the most confidently classified age faces (young and old). Our method is able to accurately classify these despite the wide variety of pose, lighting, and expression.

We are now applying this learning method to other track invariant attributes, such as race, age, eyewear (glasses), facial hair (beards, mustache), color of hair, etc. Of course, the method is not applicable to attributes that change within a track, such as expression or smoking. A similar method could be applied to learn attributes for other trackable objects, such as pedestrians and cars.

There are obvious links between our training method and finding the maximally violated constraints in cutting plane optimization algorithms, and we are currently investigating this.

Acknowledgements. We are grateful for financial support from the Royal Academy of Engineering, Microsoft, ERC grant VisRec no. 228180, the MSR-INRIA laboratory and the Quaero Programme, funded by OSEO.

References

1. Kumar, N., Belhumeur, P.N., Nayar, S.K.: Face Tracer: A search engine for large collections of images with faces. In: Proc. European Conference on Computer Vision. (2008)
2. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: Proc. International Conference on Computer Vision. (2009)
3. Kenneth, A., Coltrane, S.: Gender displaying television commercials: A comparative study of television commercials in the 1950s to 1980s. Sex roles **35** (1996)
4. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: Proc. Computer Vision and Pattern Recognition. (2009)
5. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
6. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison (2005)
7. Yan, R., Zhang, J., Yang, J., Hauptmann, A.G.: A discriminative learning framework with pairwise constraints for video object classification. PAMI **28** (2006)
8. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" - automatic naming of characters in tv video. In: Proc. British Machine Vision Conference. (2006)
9. Ramanan, D., Baker, S., Kakade, S.: Leveraging archival video for building face datasets. In: Proc. International Conference on Computer Vision. (2007)
10. Sivic, J., Schaffalitzky, F., Zisserman, A.: Object level grouping for video shots. International Journal of Computer Vision **67** (2006) 189–210
11. Sivic, J., Everingham, M., Zisserman, A.: "Who are you?" - learning person specific classifiers from video. In: Proc. Computer Vision and Pattern Recognition. (2009)
12. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly, Cambridge, MA (2008)
13. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. Computer Vision and Pattern Recognition. (2001)
14. Shi, J., Tomasi, C.: Good features to track. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on. (1994) 593–600
15. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision **61** (2005)
16. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7** (1991) 11–32
17. Maji, S., Berg, A.: Max-margin additive models for detection. In: Proc. International Conference on Computer Vision. (2009)
18. Fan, R., Chang, K., Hsieh, C., Wang, R., Lin, C.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874