Contextual Identity Recognition in Personal Photo Albums

Dragomir Anguelov, Kuang-chih Lee[1], Salih Burak Göktürk, Baris Sumengen
Riya Inc.
3 Waters Park Drive, Suite 120, San Mateo CA 94305
{drago, kclee, burak, baris}@riya.com

Abstract

We present an efficient probabilistic method for identity recognition in personal photo albums. Personal photos are usually taken under uncontrolled conditions – the captured faces exhibit significant variations in pose, expression and illumination that limit the success of traditional face recognition algorithms. We show how to improve recognition rates by incorporating additional cues present in personal photo collections, such as clothing appearance and information about when the photo was taken. This is done by constructing a Markov Random Field (MRF) that effectively combines all available contextual cues in a principled recognition framework. Performing inference in the MRF produces markedly improved recognition results in a challenging dataset consisting of the personal photo collections of multiple people. At the same time, the computational cost of our approach remains comparable to that of standard face recognition approaches.

1. Introduction

Face recognition can help greatly with the organization, searching and sharing of personal photos. This has been demonstrated in popular software by the company Riya [20], which uses face recognition to annotate images with the names of people that appear in them. However, the recognition problem in personal photo collections is very challenging due to the fact that the photos are obtained in a completely uncontrolled fashion. Typically, faces are captured in a wide range of expressions and orientations, and at different scales. The lighting conditions can be quite challenging as well, especially in photos containing shadows and directional sunlight. Traditional face recognition systems attack these problems one by one [2, 16, 5, 1] and their performance deteriorates as each of these conditions varies.

In this paper, we show how to improve the recognition performance by exploiting additional cues present in sets of digital photographs. These cues include the time a photo was taken, the clothing people wear, as well as additional commonsense knowledge, such as the fact that the same person does not appear twice in the same photo. We show how to construct a Markov Random Field (MRF) that combines these additional cues with any standard face recognition algorithm in a principled way. Inference in the MRF searches in the combinatorial space of identity assignments and produces the desired recognition results. Our main contribution is a compact MRF representation, which allows us to obtain improved recognition results with little performance overhead. We show that in the absolute worst case, the number of MRF edges introduced by our method grows in $O(M^2)$, where M is the maximum number of faces in any event (an event is defined as a period of time with a duration of 6 hours). We show how even this bound can be drastically reduced in practice without a noticeable decrease in recognition rates.

The main details of the algorithm are introduced as follows. In Sec. 3 we describe the probabilistic framework that combines face, clothing and other contextual cues for identity recognition. In Sec. 4 we describe in detail our clothing similarity model, which is part of the framework. Experimental results are presented in Sec. 5 and in Sec. 6 we conclude by a discussion of our approach and possible future work.

2. Related Work

Cues such as hair appearance [21], global positioning information [9] and clothing appearance [23, 22, 15] have been investigated as a way of boosting recognition performance in images. We describe in detail how to incorporate clothing appearance into our recognition framework; the other cues can be introduced in a similar way.

Similar to our work, clothing models have been used to aid the task of identity annotation in personal photo collections [23, 22, 15]. The early work of Zhang *et al*. [23] proposes a specific similarity function between two people that takes face and clothing into account. Such a similarity function (although differing in the details) is one of the

[1]Now at DigitalPersona Inc.

components of our recognition framework. In our application, the goal is to to recognize some specific people of interest to the user, which is different from a clustering-based approach. An example of the latter is the work of Song and Leung [22], whose goal is to obtain clusters of faces in photo collections that correspond to different individuals. Song and Leung use face and clothing similarity functions to construct an affinity matrix over the identities of the detected people, and use a normalized-cut method to obtain this desired clustering. This is a rather expensive approach, where the size of the affinity matrix scales quadratically with the number of detected faces. We demonstrate that our recognition application introduces a considerably smaller number of MRF edges, and as a result can be performed much more efficiently.

Clothing appearance has also been used for person detection in sets of photos taken over a short period of time. A popular set of methods uses pictorial structures [4, 15, 11] to find instances of the person that were missed by the face detector. This set of methods can detect additional instances of a person in a scene, assuming the person's clothing appearance is known. Our method does not make such an assumption, and as a result can be used as a natural preprocessing step, which determines the identities of some of the detected faces. The results can then be used to initialize the methods above, in order to to detect additional instances of the identified people. However, this process is outside the scope of this paper.

Clothing models have also been used by tracking methods to improve the annotation of video data [11, 12, 6, 3]. For example, Ramanan *et al*. [12] detect the human body in some canonical poses, estimate a color model for the body parts, and uses the model to track backwards and forwards in time, which allows additional frames to be annotated. Everingham *et al*. [3] combine face and clothing information with text transcriptions to annotate each movie frame with the people that are present. This set of methods use the small motion assumption to disambiguate the identity correspondence problem between successive frames. Our approach applies to images, where this assumption does not hold. Our MRF models the correlations between the identities of the detected faces, based on their clothing similarity. In principle, our model can be generalized to the case of video, where instead of recognition of faces we model the relationships and perform recognition for entire face tracks.

3. Joint Probabilistic Framework

In this section we describe a principled way of combining face similarity scores, clothing similarity scores, and additional constraints in the recognition process. We show how to encode all the relevant information into a Markov Random Field (MRF); performing inference in the MRF yields the desired identity recognition results.

We have a set of photos $\mathcal{P}$; each photo in this set has an associated *timestamp* that indicates when the photo was taken. Let $\mathcal{F} = \{f_1, \ldots, f_N\}$ be the set of all the detected faces and $\mathcal{C} = \{c_1, \ldots, c_N\}$ be the set of associated clothing features (to be described later in Sec. 4). Let $\mathcal{X} = \{X_1, \ldots, X_N\}$ denote the set of unknown identities associated with these faces. The domain of each variable is $\mathrm{dom}(X_i) = \{1, \ldots, D, u\}$, where D is the total number of identities we are trying to recognize, and u corresponds to the *unknown* identity label.

We assume that the photos and the faces in them can be partitioned into a set of *events* $\mathcal{E}$. The event length corresponds to the amount of time that we expect a person to wear the same clothing. We use the simplest possible greedy algorithm: sort the photos by timestamp and go through them in increasing order, each photo that doesn't fall inside an existing event starts a new event. We let e_i denote the event to which face f_i is assigned. We found in practice that 4 or 6 hours is a suitable event length.

Our MRF is a representation of a joint probability distribution over the identity variables $\mathcal{X}$, which encodes the necessary knowledge for identity recognition. Below we first describe a naïve design of the MRF, and then we show how to obtain a more efficient construction.

3.1. Naïve MRF Model

The MRF is made up of three kinds of potentials: single *face similarity* potentials $\phi_i^f(X_i)$, pairwise *clothing similarity* potentials $\psi_{i,j}^c(X_i, X_j)$ and pairwise *uniqueness potentials* $\psi_{i,j}^u(X_i, X_j)$.

The model contains N *face similarity* potentials $\phi_i^f(X_i)$. There are $D+1$ probabilities for each face potential, corresponding to the likelihood of face f_i matching each of the people we want to recognize (or none of those). Any face recognition model can be used to elicit probabilities for $\phi_i^f(X_i)$ [24, 14]. A comparison of the different possible face models is outside the scope of this work.

We also introduce *clothing similarity* potentials $\psi_{i,j}^c(X_i, X_j)$ between all pairs of faces in the same event, but not in the same photo. These potentials express the idea that two similar pieces of clothing are likely to belong to the same person in that event. These are potentials of the following form:

$$\psi_{i,j}^c(x_i, x_j) = \begin{cases} 1 & \text{if } x_i \neq x_j; \\ 1 & \text{if } x_i = x_j = u; \\ w_c S(c_i, c_j) & \text{otherwise.} \end{cases} \quad (1)$$

where $S(c_i, c_j)$ is the clothing similarity score described in Sec. 4.3 and w_c is a scaling weight.

We also want to enforce the constraint that no person appears in the same photo twice. To do this we introduce pairwise *uniqueness* potentials between all faces in the same

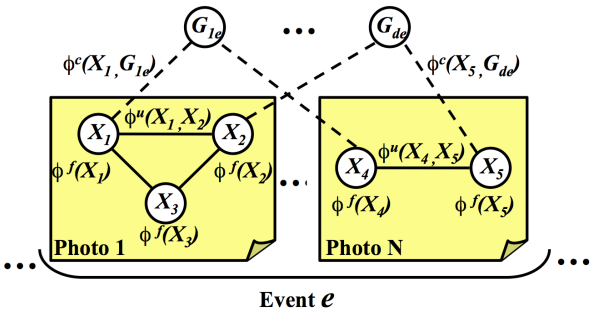

Figure 1. A sketch of the MRF decribed in Sec. 3.2, illustrating the potentials between the identity variables X and the clothing variables G. Each identity variable X_i is associated with a face similarity potential $\phi_i^f(X_i)$. Uniqueness potentials $\phi_{i,j}^u(X_i, X_j)$ are shown as solid lines between variables X_i in the same photo. Clothing potentials $\phi_{i,de}^c(X_i, G_{de})$ are shown as dashed lines in the same event.

photo:

$$\psi_{i,j}^u(x_i, x_j) = \begin{cases} 1 & \text{if } x_i \neq x_j; \\ \tau & \text{if } x_i = x_j. \end{cases} \quad (2)$$

This type of MRF constraints has been used before in different applications, including tracking [7] and the face clustering work of Song and Leung [22]. Setting $\tau = 0$ enforces the constraint strictly, however in our experiments $\tau = 0.25$ performed slightly better.

While the the model above allows us to combine information about face similarity, clothing similarity and uniqueness constraints, it results in MRF graphs that are fairly large. Each face potential in the MRF requires the estimation of $D+1$ probability values, which typically include expensive comparisons of face features. Even more importantly, this model requires the introduction of at least $O(M^2)$ clothing similarity potentials, where M is the maximum number of faces in an event. Since in practice M can go up to a few hundreds and even thousands, this approach can quickly become computationally impractical, especially if we rely on obtaining the recognition results quickly. Fortunately, there are effective ways in which this computational overhead can be reduced.

3.2. Efficient MRF Representation

The face potentials as described above require the computation of $D+1$ identity probabilities for N faces, where D can be in the hundreds and N can be in the thousands. Depending on the complexity of the underlying face representation, this can take a prohibitively long amount of time and calls for a pruning approach. It has been observed that clothing is a less informative feature than face, and simply using clothing for recognition produces many false positives [22]. As a result, the face potential values are weighted more, and tend to dominate the values of the clothing potentials. This justifies a strategy, which prunes all values in $\text{dom}(X_i)$ with low face probability, because even a good clothing match would not be able to push those hypotheses over the recognition threshold. This strategy is a natural fit with efficient nearest-neighbor search structures for face feature comparison that efficiently avoid the need to compare the faces to prototypes that look very different. We base the face similarity probability only on the K-nearest neighbor prototypes, which naturally limits the maximum domain size to $K+1$ and produces efficiency improvements.

We also transform the MRF in order to decrease the number of clothing similarity potentials. We introduce additional variables G_{de} into the MRF, which correspond to the clothing person d wears in event e (see Fig. 1). These variables are connected to the variables X_i in that event with potentials that match the clothing c_i to the clothing estimate in G_{de}. In theory, we could perform inference in such a MRF that will not only figure out the face identities, but also what kind of clothing each person is wearing in each event.

In practice, the idea as stated is problematic, because G_{de} would be continuous variables in the high-dimensional space of clothing parameters. Inference methods in continuous MRFs are much more difficult, less accurate and less efficient. However, this issue can be addressed by discretizing the domains of G_{de} using actual observed clothing examples in event e. Consider all clothing examples $\mathcal{C}_{de} \equiv \{c_i : (e_i = e, \phi_i^f(d) \geq \sigma)\}$ – these are the clothes of people in event e whose face is a good match for identity d.

Using the set $\mathcal{C}_{de}$ we can discretize the domain of G_{de}, such that each value g_{de} corresponds to a particular clothing candidate c_{de}, or *unknown*. The clothing potentials $\psi_{i,de}^c(X_i, G_{de})$ then are added between each G_{de} and all the variables $\{X_i : d \in \text{dom}(X_i), e_i = e\}$:

$$\psi_{i,de}^c(x_i, g_{de}) = \begin{cases} 1 & \text{if } x_i \neq d; \\ 1 & \text{if } x_i = d, g_{de} = u; \\ 1 & \text{if } x_i = d, c_i = c_{de}; \\ w_c S(c_i, c_{de}) & \text{otherwise.} \end{cases}$$

It is important to note the potential value in the case $c_i = c_{de}$, corresponding to the match of a clothing example c_{de} to the item from which it was originally borrowed. A potential value of $w_c S(c_i, g_{de})$ here would be rewarding the item for matching to itself, which is incorrect.

So what did we achieve with this model? First, it can be shown that if $\sigma = 0$ (all clothing items are considered as candidates for all the clothing variables G_{de}), we end up with roughly the same number of clothing potential parameters as the MRF from Sec. 3.1. One important reduction in parameters occurs as we increase σ. More importantly, another decrease results from a pruning procedure that merges all clothing examples in C_{de} that look very similar according to the score $S(c_i, c_{de})$. The result of this procedure is

a substantial reduction in the domain sizes of G_{de}. Such a pruning procedure *cannot* be carried out in the MRF representation from Sec. 3.1.

3.3. MRF Inference

The MRF described in Sec. 3.2 defines a joint probability distribution over the detected face identities $\mathcal{X}$ and the clothing variables $\mathcal{G}$. Our task is to find the posterior marginal probabilities of each variable X_i. Among the multiple possible algorithms for MRF inference, we chose to use *loopy belief propagation (LBP)* [10], which has been shown to work effectively in a broad range of applications. We run LBP with parallel *sum-product* updates. Since we do not have pairwise potentials between variables assigned to different events, we run separate inference for each subset of the MRF graph that contains only variables from a single event.

4. Clothing Model

In this section we describe a model, which can be used to identify the same (or very similar) pieces of clothing in photographs. Similar to face recognition, clothing recognition is a process dependent on several stages: detection and segmentation of clothing, extraction of features, and computing the similarity between those features.

4.1. Clothing Detection and Segmentation

In the ideal case, we would want to perform accurate segmentation of the different clothing pieces a person is wearing. However, this is a very difficult open problem due to several reasons. People are viewed from different points of view and appear in different poses. Clothes can exhibit different folds and wrinkles, and there can be significant occlusion and clutter in the photos.

We adopt the approach used in the majority of the methods for clothing feature extraction [3, 23, 22, 6]. The assumption underlying all these methods is that modeling the clothing covering the person's torso is sufficient to improve the recognition quality significantly. We use the position and relative scale of each detected face in order to predict a bounding box which is expected to contain the person's torso. After some experimentation, we got the best results with fairly narrow clothing boxes (shown in Fig. 2). Such boxes are likely to contain a part of the torso, even in the presence of some face orientation and body pose variation. We do further postprocessing to detect occlusions — when another detected face in the scene occludes a clothing box, the clothing descriptor in that image is set to *unknown*. Unlike Song and Leung [22], who use a skin color model to ignore parts of the clothing box that correspond to human skin, we keep these parts. Keeping them was beneficial in our experiments, because they provide valuable information about whether the clothing exposes the upper chest and the neck.

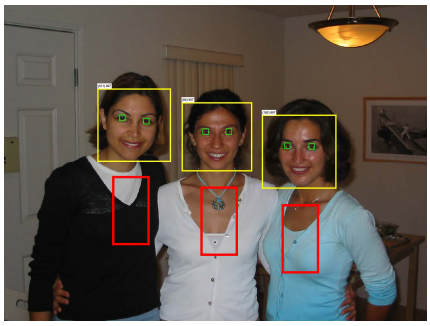

Figure 2. Image showing detected faces and clothing boxes used.

4.2. Clothing Features

Our choice of clothing representation was motivated by the need to devise efficient features that can handle the variability in clothing appearance. We use two features that capture the color and the texture of clothing, respectively. The color of clothing is typically modeled using histogram features [3, 15, 6]. However, we obtained slightly better results with an adaptive binning technique which uses the standard K-means algorithm to cluster the RGB colors in each clothing box. The result is a set of clusters $L \equiv \{(l_1, m_1), \ldots, (l_K, m_K)\}$, where l_k denotes the RGB color of the cluster k, and m_k is the relative amount of color in the cluster ($\sum_k m_k = 1$). The number of clusters we obtain is data dependent, based on a penalty discouraging very small clusters. The penalty is set such that we typically obtain between 5 and 10 color clusters.

The color distance between two clothing boxes then is computed using Earth mover's distance (EMD) between the respective cluster centers:

$$EMD(L^1, L^2) = \sum_{i=1}^{K^1} \sum_{j=1}^{K^2} d(l_i^1, l_j^2) \cdot f_{i,j} \qquad (3)$$

$$\text{s.t.} \quad f_{i,j} \geq 0,\ f_{i,j} \leq m_i^1,\ f_{i,j} \leq m_j^2,\ \sum_{i,j} f_{i,j} = \gamma$$

Here $f_{i,j}$ are the set of "flows" that match color amounts from clusters i and j. The EMD computation reduces to finding the set of flows that minimizes the objective, and can be performed efficiently [13]. In our setting, we require that only a subset of the total color mass is matched. To this effect, we set $\gamma = 0.85$ in the constraint $\sum_{i,j} f_{i,j} = \gamma$ above. This setting of γ was found to perform better in practice, because it provides some tolerance for misalignment of clothing boxes, and for some lighting variations.

The choice of distance function $d(l_i, l_j)$ is important for obtaining good recognition performance. Because people

are captured in varying lighting conditions in photographs, we want a color distance that is less sensitive to illumination. If $\alpha \in [-\pi/2, \pi/2]$ is the angle between l_i and l_j, and $r \in (0,1]$ is the ratio of their lengths, our distance is:

$$d(l_i, l_j) = \alpha^2 + b \cdot \delta_{(r<s)}(r-s)^2, \quad (4)$$

where $\delta_{(r<s)}$ is 1 if $r < s$ and 0 otherwise. The intuition behind this distance is that it mostly penalizes change in color, and somewhat less the change in illumination. Furthermore, there is zero penalty for small changes in illumination $(r > s)$. We found that good results are obtained with the setting $s = 0.8$.

To capture the texture of clothing, we apply Gabor filters [8] in 4 orientations and 3 scales, and quantize the texture space using the k-means algorithm into 75 clusters. This allows us to associate each pixel in the clothing box with a particular cluster, obtaining a texture histogram.

4.3. Combining the Clothing Features

We learn the relative importance of the different clothing features using ground truth dataset containing people whose identities are known. We only compare clothes in photos which were taken within the period during which people are expected to wear the same set of clothes. In defining the score, we use the idea in [23]. We apply regression on this ground truth dataset to predict the log probability $P(X_i = X_j \mid c_i, c_j)$, and use that probability as the clothing similarity score $S(c_i, c_j)$. In the regression, we learn a separate weight for different texture clusters, to capture the idea that some textures are more unique and stable than others.

To summarize, the features for clothing comparison proposed here are designed to possess several desirable properties. 1) The color and texture features are robust to imperfect alignment of clothing boxes. 2) The color features are also largely robust to lighting variation. 3) Both the color and texture features, due to the use of clustering, tend to ignore small amounts of noise and clutter in the picture.

5. Experimental Results

Experiments are performed on real consumer photos. We use 8 personal photo collections obtained from different people, and containing a variety of settings including indoors, outdoors, pool scenes, birthday parties and class reunions. Some pictures contain more than 30 faces. An Adaboost-based face detection algorithm [19] is run, and the identities of the people in the photos are manually labeled to provide the ground truth for evaluation. We also used an automatic Adaboost-based registration algorithm for the eyes, the nose and the mouth corners. No postprocessing was made to correct the registration errors. No simplifications that benefit the algorithm were made. Almost all faces captured in the photos larger than a certain minimal size (to avoid people very far in the background) were labeled.

	# Faces	# Labeled	# Identities	Ms/face	Total(s)
D1	913	260	49	0.46	0.42
D2	1387	244	13	0.67	0.93
D3	2123	456	49	0.59	1.25
D4	3079	716	117	0.79	2.43
D5	4493	1237	35	1.11	4.99
D6	4238	345	12	0.69	2.92

Table 1. Running times of the algorithm. A subset of the faces is labeled and provided to the algorithm, when then is run on the remaining faces. The running times remain relatively small even for large datasets.

Our face similarity potentials are based on K-nearest neighbors classification in two distinct feature spaces (we set K=5, but also tried K=7 and K=10 without noticeable improvements in the recognition rate). One of those spaces is based on the standard Eigenfaces method [18]. This method is applied separately to the entire face, as well as the two regions around the eyes, and the scores from the three features are combined in the final score. The other space is obtained by first registering the face image to a 3D model. After using the 3D model to correct for pose variations, a separate transformation based on discriminative space learning is applied [17]. The two models are combined to obtain the final face similarity score. The precise details of the face model are outside the scope of this paper. The framework presented here is suitable for use with any good face similarity measure.

We used cross-validation on several datasets to obtain a good tradeoff between the probabilities obtained by the face model, and those of the clothing model. The algorithm was then evaluated on six of the original eight datasets. We randomly label 30% (or 10%) of the faces and average 5 runs to get our curves. When some identities are unlabeled by the process, the theoretical recognition rate can be much less than 100%. Also, a small fraction of the photos were obtained using a digital scanner and therefore missing timestamps. For those, no clothing potentials were introduced.

We evaluated our algorithm against two simpler alternatives, and plot the recognition performance at 5% false positive rate (Fig. 3). This number is motivated by the expectation that users of the algorithm would not be willing to tolerate a much higher error rate. We compare against a baseline method that uses only the face similarity potentials, and post-processes the results to identify cases when the same identity is found twice in the same image. Only the more likely instance is picked in such cases. The second alternative combines the baseline algorithm with the uniqueness potentials, and shows that even though they offer a small improvement over the baseline strategy, it is fairly minimal.

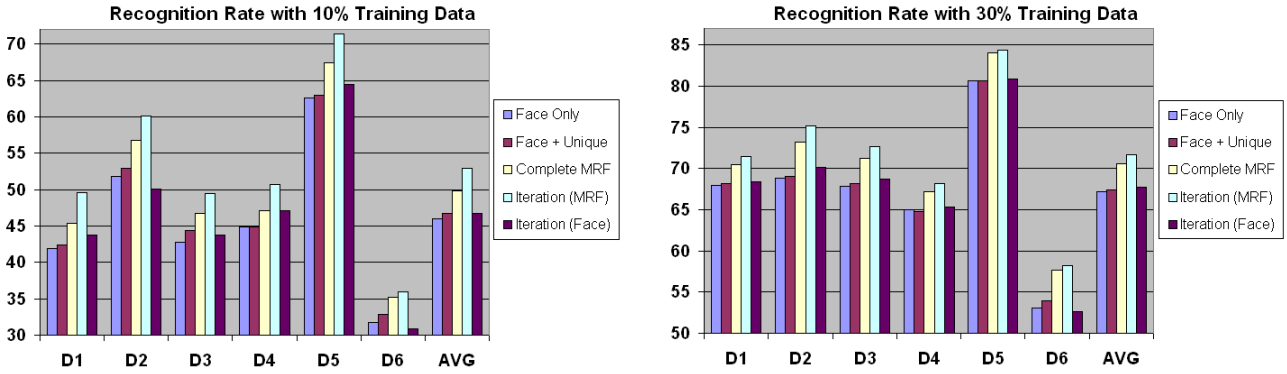

Figure 3. Comparison of the recognition accuracy in six different datasets. The algorithm takes as input the labels for 10% and 30% of the faces in each dataset, respectively. The plot shows the correct recognitions obtained at 5% false positive rate. The improvement of our method (Complete MRF) over the baseline (Face Only) is 3.8% and 3.6% on average. Adding likely recognitions to the ground truth and iterating the algorithm (Iteration MRF) yields additional increases of 3.1% and 1.1% on average. Using the same iteration technique with the baseline face recognition method (Iteration Face) yields almost no improvement.

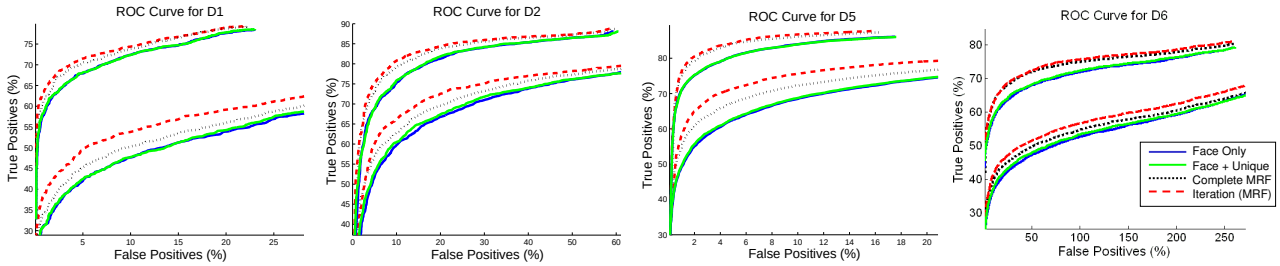

Figure 4. ROC comparison for four datasets. The horizontal axis displays the false positive rate (FP rate) relative to the number of labeled faces. As some detected faces are not labeled in the ground truth, the FP rate can exceed 100%. Each graph holds two sets of curves, when 10% and 30% of the face labels are provided to the algorithm, respectively. The datasets are of varying level of difficulty, in particular D6 has a lot of low-resolution photographs, and in some scenes different people wear similar clothing. In all cases our algorithm provides reasonable improvement along the entire curve. Our iteration strategy is less effective when more identities are labeled. In this case, the ground truth is more likely to contain labeled instances for each person in each event, which reduces the uncertainty about the kinds of clothing people are wearing.

Our complete algorithm is shown to outperform the alternatives in all cases. The average improvement of 3.42% or 3.81% of the recognition rate for runs with 30% and 10% training data is not very large, but is reasonable given the difficulty of the datasets and the testing methodology. In some of the datasets, the improvement approaches 5%.

We found that additional improvement can be achieved by using the following iteration strategy. We run our algorithm, and then take the faces which are recognized with high certainty (over 0.825), add them to the ground truth and re-run the algorithm. We observe an average additional improvement of 1.1% and 3.1% for the runs with 30% and 10% training data (see Fig. 3). The same iteration strategy applied to the baseline face recognition algorithm yields marginal improvement, and even deterioration in some cases. The reason the iteration strategy was more beneficial for our approach, is that through our use of clothing we end up adding examples further away on the face similarity manifold to the ground truth. In the second iteration, these examples contribute towards the recognition of faces which are close to these new examples but relatively far from the original ones. The iteration approach naturally works better when there are fewer labeled examples. Otherwise, the face similarity manifold will be populated with a sufficient number of examples from the start.

In Fig. 4 we show the ROC curves for four of the datasets. The curves show that there is a significant difference in difficulty between the datasets. While in some datasets such as D1 and D5 the algorithm offers significant advantages, in D6 the improvement is quite minimal, which is due to the fact that it contains pictures of weddings and reunions at which many people wear the same clothing. Even in such a challenging circumstance the algorithm improves the results, albeit only slightly.

The algorithm's performance on a 1.8 GHz AMD Opteron processor is shown in Table 1. These running times

do not contain the time necessary to compute the original face and clothing representations. However, the time to compute similarities between pairs of faces and clothings is included, as well as the time to construct the Markov Net, to perform the inference and to return the results. The algorithm performed very efficiently even for very large datasets with many faces and labeled examples. We observed that about 80% of the time is spent in comparing face features, and only about 20% is used to compare clothing similarity and to run MRF inference. We feel that this overhead is a reasonable price to pay for the consistent improvement in the recognition results that the algorithm offers.

6. Conclusions and Future Directions

We have presented a flexible and principled framework for exploiting multiple cues for identity recognition in personal digital photo collections. We have shown that our algorithm provides consistent improvement in recognition results without significantly exceeding the running time of standard face recognition algorithms.

There are many directions in which this work can be extended. The clothing features can be perfected, and compared to those used by Song and Leung [22]. Additional cues can be incorporated, such GPS location information, photo captions, and learning that certain groups of people tend to appear together in the same events. A way for learning all parameters of the MRF potentials directly from data needs to be investigated, as opposed to our current approach, where a few key parameters were set by cross-validation. Finally, it would be interesting to extend this approach to video data, where for each appearance of a person we can pool information not from a single, but from a multiple adjacent frames. Our approach can be used to construct a MRF that models the clothing and face similarity, as well as uniqueness constraints, for entire groups of frames.

Acknowledgements

We thank Lorenzo Torresani for the useful discussions and Diem Vu for his data processing help.

References

[1] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible lighting conditions. In *Int'l. J. Computer Vision*, volume 28, pages 245–260, 1998.

[2] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang. Coding, analysis, interpretation, and recognition of facial expressions. *Computer Vision and Image Understanding*, 2003.

[3] M. R. Everingham, J. Sivic, and A. Zisserman. 'hello! my name is... buffy' - automatic naming of characters in tv video. In *Proc. of the 17th British Machine Vision Conference (BMVC2006)*, pages 889–908, September 2006.

[4] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 66–73, 2000.

[5] A. Georghiades, D. Kriegman, and P. Belhumeur. From few to many: Generative models for recognition under variable pose and illumination. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 40:643–660, 2001.

[6] G. Jaffre and P. Joly. Costume: A new feature or automatic video content indexing. In *Proc. RIAO*, 2004.

[7] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *Proc. European Conf. on Computer Vision*, pages 279–290, 2004.

[8] B. S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI - Special issue on Digital Libraries)*, 18(8):837–42, Aug 1996.

[9] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. *Proc. of the Fifth ACM/IEEE-CS Joint Conf. on Digital Libraries*, 2005.

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, 1988.

[11] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2003.

[12] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2005.

[13] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[14] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition*, 25(1):65–77, 1992.

[15] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proc. of the 16th British Machine Vision Conference*, pages 909–918, 2006.

[16] T.Coots, G.Edwards, and C.Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[17] Lorenzo Torresani and Kuang chih Lee. Large margin component analysis. In *Advances in Neural Information Processing Systems*, December 2006.

[18] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[19] Paul Viola and Michael Jones. Robust real-time object detection. *Int'l. J. Computer Vision*, 2001.

[20] Riya visual search. http://www.riya.com.

[21] Y. Yacoob and L. Davis. Detection, analysis and matching of hair. In *Proc. Int'l. Conf. on Computer Vision*, 2005.

[22] Y.Song and T.Leung. Context-aided human recognition - clustering. *In Proc. ECCV*, 3:382–395, 2006.

[23] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. *ACM Multimedia*, 2003.

[24] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.