Clothing Cosegmentation for Recognizing People

Andrew C. Gallagher
Carnegie Mellon University
Pittsburgh, Pennsylvania
agallagh@cmu.edu

Tsuhan Chen
Carnegie Mellon University
Pittsburgh, Pennsylvania
tsuhan@cmu.edu

Abstract

Reseachers have verified that clothing provides information about the identity of the individual. To extract features from the clothing, the clothing region first must be localized or segmented in the image. At the same time, given multiple images of the same person wearing the same clothing, we expect to improve the effectiveness of clothing segmentation. Therefore, the identity recognition and clothing segmentation problems are inter-twined; a good solution for one aides in the solution for the other.

We build on this idea by analyzing the mutual information between pixel locations near the face and the identity of the person to learn a global clothing mask. We segment the clothing region in each image using graph cuts based on a clothing model learned from one or multiple images believed to be the same person wearing the same clothing. We use facial features and clothing features to recognize individuals in other images. The results show that clothing segmentation provides a significant improvement in recognition accuracy for large image collections, and useful clothing masks are simultaneously produced.

A further significant contribution is that we introduce a publicly available consumer image collection where each individual is identified. We hope this dataset allows the vision community to more easily compare results for tasks related to recognizing people in consumer image collections.

1. Introduction

Figure 1 illustrates the limitations of using only facial features for recognizing people. When only six faces (cropped and scaled in the same fashion as images from the PIE [24] database often are) from an image collection are shown, it is difficult to determine how many different individuals are present. Even if it is known that there are only three different individuals, the problem is not much easier. In fact, the three are sisters of similar age. When the faces are shown in context with their clothing, it becomes almost trivial to recognize which images are of the same person.

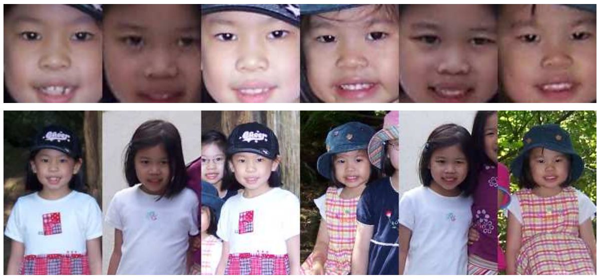

Figure 1. It is extremely difficult even for humans to determine how many different individuals are shown and which images are of the same individuals from only the faces (top). However, when the faces are embedded in the context of clothing, it is much easier to distinguish the three individuals (bottom).

To quantify the role clothing plays when humans recognize people, the following experiment was performed: 7 subjects were given a page showing 54 labeled faces of 10 individuals from the image collection and asked to identify a set of faces from the same collection. The experiment was repeated using images that included a portion of the clothing (as shown in Figure 1). The average correct recognition rate (on this admittedly difficult family album) jumped from 58% when only faces were used, to 88% when faces and clothing were visible. This demonstrates the potential of person recognition using features in addition to the face for distinguishing individuals in family albums.

When extracting clothing features from the image, it is important to know where the clothing is located. In this paper, we describe the use of graph cuts for segmenting clothing in a person image. We show that using multiple images of the same person from the same event allows a better model of the clothing to be constructed, resulting in superior clothing segmentation. We also describe the benefits of accurate clothing segmentation for recognizing people in a consumer image collection.

2. Related Work

Clothing for identification has received much recent research attention. When attempting to identify a person from the same day as the training data for applications such as teleconferencing and surveillance, clothing is an important cue [8, 11, 18]. In these video-based applications, good figure segmentation is achieved from the static environment.

In applications related to consumer image collections [1, 27, 29, 31, 32], clothing color features have been characterized by the correlogram of the colors in a rectangular region surrounding a detected face. For assisted tagging of all faces in the collection, combining face with body features provides a 3-5% improvement over using just body features. However, segmenting the clothing region continues to be a challenge; all of the methods above simply extract clothing features from a box located beneath the face, although Song and Leung [27] adjust the box position based on other recognized faces and attempt to exclude flesh.

Some reseachers have trained models to essentially learn the characteristics of the human form [7, 16, 19, 28]. Broadly speaking, these methods search for body parts (e.g. legs, arms, or trunk), and use a pre-defined model to find the most sensible human body amongst the detected parts. While a model-based approach is certainly justified for the problem, we wonder what can be learned from the data itself. Given many images of people, is it possible for the computer to learn the shape of a human without imposing a physical human model on its interpretation of the images?

Regarding segmenting objects of interest, researchers have attemped to combine the recognition of component object parts with segmentation [30], and to recognize objects among many images by first computing multiple segmentations for each image [22]. Further, Rother *et al*. extend their GrabCut [20] graph-cutting object extraction algorithm to operate on simultaneously on pairs of images [21], and along the same lines, Liu and Chen [15] use PLSA to initialize the GrabCut, replacing the manual interface. We extend this problem into the domain of recognizing people from clothing and faces. We apply graph cuts simultaneously to a group of images of the same person to produce improved clothing segmentation.

Our contributions are the following: We analyze the information content in pixels surrounding the face to discover a global clothing mask (Section 4). Then, on each image, we use graph-cutting techniques to refine the clothing mask, where our clothing model is developed from one or multiple images believed to contain the same individual (Section 5). In contrast to some previous work, we do not use any model of the human body. We build a texture and color visual word library from features extracted in putative clothing regions of people images and use both facial and clothing features to recognize people. We show these improved clothing masks lead to better recognition (Section 7).

	Set 1	Set 2	Set 3	Set 4
Total images	401	1065	2099	227
Images with faces	180	589	962	161
No. faces	278	931	1364	436
Detected faces	152	709	969	294
Images with multiple people	77	220	282	110
Time span (days)	28	233	385	10
No. days images captured	21	50	82	9
Unique individuals	12	32	40	10

Table 1. A summary of the four image collections.

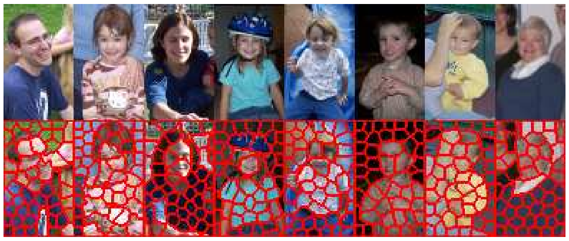

Figure 2. Person images at resolution 81×49 and the corresponding superpixel segmentations.

3. Images and Features for Clothing Analysis

Four consumer image collections are used in this work. Each collection owner labeled the detected faces in each image, and could add faces missed by the face detector [10]. The four collections, summarized in Table 1, contain a total of 3009 person images of 94 unique individuals. We experiment on each collection separately (rather than merging the collections), to simulate working with a single person's image collection.

Features are extracted from the faces and clothing of people. Our implementation of a face detection algorithm [10] detects faces, and also estimates the eye positions. Each face is normalized in scale (49×61 pixels) and projected onto a set of Fisherfaces [3], representing each face as a 37-dimensional vector. These features are not the state-of-the-art features for recognizing faces, but are sufficient to demonstrate our approach.

For extracting features to represent the clothing region, the body of the person is resampled to 81×49 pixels, such that the distance between the eyes (from the face detector) is 8 pixels. The crop window is always axis-aligned with the image. Clothing comes in many patterns and a vast pallette of colors, so both texture and color features are extracted. A 5-dimensional feature vector of low-level features is found at each pixel location in the resized person image. This dense description of the clothing region is used based on the work of [13, 14] as it is necessary to capture the information present even in uniform color areas of clothing. The three color features are a linear transformation of RGB color values of each pixel to a luminance-chrominance space (LCC).

The two texture features are the responses to a horizontal and vertical edge detector.

To provide some robustness to translation and movement of the person, the feature values are accumulated across regions in one of two ways. In the first (superpixel) representation, the person image is segmented into superpixels using normalized cuts [23], shown for example in Figure 2. For each superpixel, the histogram over each of the five features is computed. In turn, each pixel's features are the five histograms associated with its corresponding superpixel. This representation provides localization (over each superpixel) and maintains some robustness to translation and scaling. The notation $\mathbf{s}_p$ refers to the feature histograms associated with the p^{th} superpixel. Likewise, the notation $\mathbf{s}_{(x,y)}$ refers to the feature histograms associated with the superpixel that corresponds to position (x, y).

In the second (visual word) representation, the low-level feature vector at each pixel is quantized to the index of the closest visual word [25], where there is a separate visual word dictionary for color features and for texture features (each with 350 visual words). The clothing region is represented by the histogram of the color visual words and the histogram of the texture visual words within the clothing mask region (described in Section 4). Of course, this clothing mask is the putative region of clothing for the face; the actual clothing in a particular person image may be occluded by another object. The visual word clothing features are represented as $\mathbf{v}$.

4. Finding the Global Clothing Mask

In previous recognition work using clothing, either a rectangular region below the face is assumed to be clothing, or the clothing region is modeled using operator-labeled clothing from many images [26]. We take the approach of learning the clothing region automatically, using only the identity of faces (from labeled ground-truth) and no other input from the user. Intuitively, the region associated with clothing carries information about the identity of the face. For example, in a sporting event, athletes wear numbers on their uniforms so the referees can easily distinguish them. Similarly, in a consumer image collection, when two people in different images wear the same clothing, the probability increases that they might be the same individual. We discover the clothing region by finding pixel locations that carry information about facial identity. Let $p_i = p_j$ be the event S_{ij} that the pair of person images p_i and p_j share an identity, and $\langle \mathbf{s}_{i(x,y)}, \mathbf{s}_{j(x,y)} \rangle_s$ be the distance between corresponding superpixel features $\mathbf{s}_{i(x,y)}$ and $\mathbf{s}_{j(x,y)}$ at pixel position (x, y). The distance is the sum of χ^2 distances between the five feature histograms:

$$\langle \mathbf{s}_{i(x,y)}, \mathbf{s}_{j(x,y)} \rangle_s = \sum_u \chi^2(\mathbf{s}^u_{i(x,y)}, \mathbf{s}^u_{j(x,y)}) \qquad (1)$$

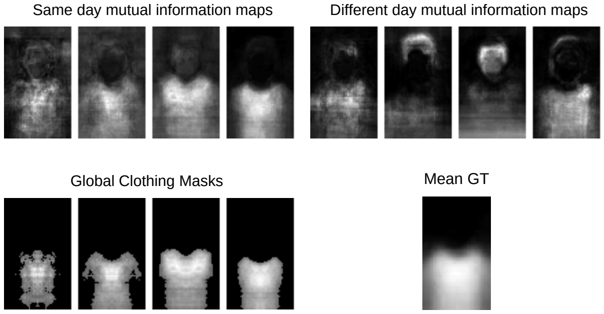

Figure 3. **Top Left:** The clothing region carries information about identity. Maps of mutual information between S_{ij} and $\langle s_{i(x,y)}, s_{j(x,y)} \rangle_s$ for four image sets all yield a map with the same qualitative appearance. In Set 3, the mutual information reaches 0.17, while the entropy of S_{ij} is only 0.19. **Top Right:** The mutual information maps for person images captured on different days. The overall magnitude is only about 7% the same-day mutual information maps, but the clothing region (and the hair region) still carry information about the identity of the person. **Bottom Left:** The clothing masks created from the mutual information masks all have the same general appearance, though Set 1's mask is noisy probably due to the relatively small number of people in this set. **Bottom Right:** The average of 714 hand-labeled clothing masks appears similar to the mutual information masks.

where u is an index over each of the five feature types (three for color and two for texture).

In the region surrounding the face, we compute the mutual information $I(S_{ij}, \langle \mathbf{s}_{i(x,y)}, \mathbf{s}_{j(x,y)} \rangle_s)$ between the distance between corresponding superpixels, and S_{ij} at each (x, y) position in the person image. Maps of the mutual information are shown in Figure 3. For each image collection, two mutual information maps are found, one where p_i and p_j are captured on the same day, and one otherwise.

Areas of the image associated with clothing contain a great deal of information regarding whether two people are the same, given the images are captured on the same day. Even for images captured on different days, the clothing region carries some information about identity similarity, due to the fact that clothes are re-worn, or that a particular individual prefers a specific clothing style or color.

In three image Sets (1, 2, and 4), the features of the face region itself carry little information about identity. (Remember, these features are local histograms of color and texture features not meant for recognizing faces). These collections have little ethnic diversity so the tone of the facial skin is not an indicator of identity. However, Set 3 is ethnically more diverse, and the skin tone of the facial region carries some information related to identity.

This mutual information analysis allows us to create a mask of the most informative pixels associated with a face that we call the *global clothing mask*. The same-day mutual information maps are reflected (symmetry is assumed), summed, and thresholded (by a value constant across the image collections) to yield clothing masks that appear remarkably similar across collections. We emphasize again that our global clothing mask is learned without using any manually labeled clothing regions; simply examining the image data and the person labels reveals that the region corresponding roughly to the torso contains information relevant to identity.

5. Graph Cuts for Clothing Segmentation

Single Image: The global clothing mask shows the location of clothing on average, but on any given image, the pose of the body or occlusion can make the clothing in that image difficult to localize. We use graph cuts to extract an image-specific clothing mask. Using the idea of GrabCut [20], we define a graph over the superpixels that comprise the image, where each edge in the graph corresponds to the cost of cutting the edge. We seek the binary labeling f over the superpixels that minimizes the energy of the cut. We use the standard graph cutting algorithms [2, 5, 6, 12] for solving for the minimum energy cut. Using the notation in [12], the energy is:

$$E(f) = \sum_{p \in \mathcal{P}} D_p(f_p) + \sum_{p,q \in \mathcal{N}} V_{p,q}(f_p, f_q) \tag{2}$$

where $E(f)$ is the energy of a particular labeling f, p and q are indexes over the superpixels, $D_p(f_p)$ is the data cost of assigning the p^{th} superpixel to label f_p, and $V_{p,q}(f_p, fq)$ represents the smoothness cost of assigning superpixels p and q in a neighborhood $\mathcal{N}$ to respective labels f_p and f_q.

Possible labels for each superpixel are $f_p \in \{0, 1\}$ where the index 0 corresponds to foreground (i.e. the clothing region that is useful for recognition) and 1 corresponds to background. The clothing model M_0 is formed by computing the histogram over each of the five features over the region of the person image corresponding to clothing in the global clothing mask. In a similar manner, the background model M_1 is formed using the feature values of pixels from regions corresponding to the inverse of the clothing mask. Then, the data cost term in Eq. (2) is defined:

$$D_p(f_p) = \exp(-\alpha \langle s_p, M_{f_p} \rangle) \tag{3}$$

where again the distance is the sum of the χ^2 distances for each of the corresponding five feature histograms. The smoothness cost term is defined as:

$$V_{p,q}(f_p, f_q) = (f_p - f_q)^2 \exp(-\beta \langle s_p, s_q \rangle) \tag{4}$$

Experimentally, we found parameter values of $\alpha = 1$ and $\beta = 0.01$ work well, though the results are not particularly sensitive to the chosen parameter values. The lower value of β is explained by considering that clothing is often occluded by other image objects, and is often not contiguous in the image. Figure 4 illustrates the graph cutting process for segmenting the clothing region. Except for the selection of a few constants, the algorithm essentially learned to segment clothing first by finding a global clothing mask describing regions of the image with high mutual information with identity, then performing a segmentation to refine the clothing mask on any particular image.

Multiple Images: When multiple images of the same person with the same clothing are available, there is an opportunity to learn a better model for the clothing. We use the idea from [21] that the background model for each image is independent, but the foreground model is constant across the multiple images. Then, the clothing model is computed with contribution from each of the images:

$$M_0 = \sum_i M_{0i} \tag{5}$$

This global clothing model M_0 is the sum for each feature type of the corresponding feature histograms for each image's individual clothing model. However, each image i has its own individual background model M_{1i}, formed from the feature values of the inverse global clothing mask. Conceptually, the clothing is expected to remain the same across many images, but the background can change drastically.

When applying graph cuts, a graph is created for each person image. The smoothness cost is defined as before in Eq. (4), but the data cost for person image i becomes:

$$D_{pi}(f_{pi}) = \begin{cases} \exp(-\alpha \langle s_{pi}, M_0 \rangle) & \text{if } f_{pi} = 0 \\ \exp(-\alpha \langle s_{pi}, M_{1i} \rangle) & \text{if } f_{pi} = 1 \end{cases} \tag{6}$$

Figure 5 shows several examples of graph cuts for clothing segmentation by either treating each image independently, or exploiting the consistency of the clothing appearance across multiple images for segmenting each image in the group.

6. Recognizing people

For searching and browsing images in a consumer image collection, we describe the following scenario. At first, none of the people in the image collection are labeled, though we do make the simplifying assumption that the number of individuals is known. A user provides the labels for a randomly selected subset of the people images in the collection. The task is to recognize all the remaining people, and the performance measure is the number of correctly recognized people. This measure corresponds to the usefulness of the algorithm in allowing a user to search and browse the image collection after investing the time to label a portion of the people. We use an example-based nearest neighbor classifier for recognizing people in this scenario.

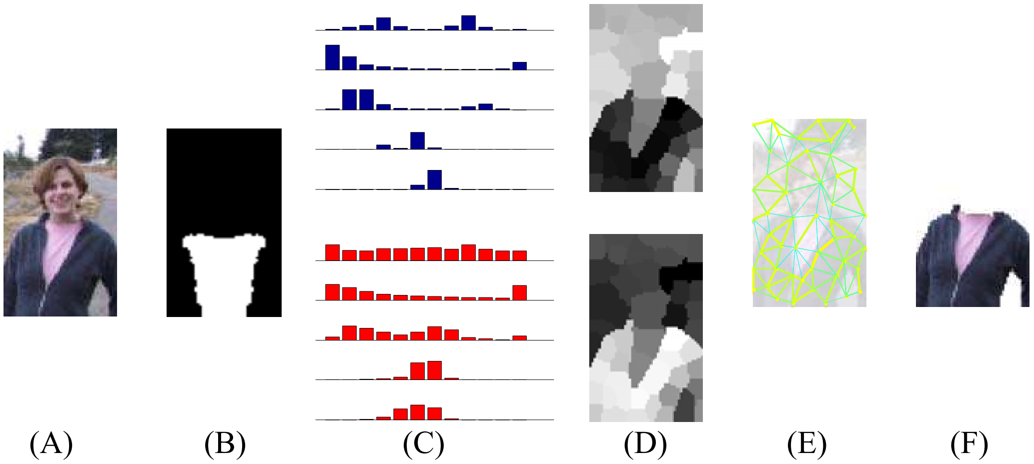

Figure 4. Using graph cuts to segment the clothing from a person image. The automatically learned global clothing mask (B) is used to create a clothing model (C, top) and a background model (C, bottom) that each describe the five feature types from the person image (A). Each superpixel is a node in a graph, and the data cost of assigning each superpixel to the clothing and background are shown (D, top) and (D, bottom), respectively, with light shades indicating high cost. The smoothness cost is shown in (E), with thicker, yellower edges indicating higher cost. The graph cut solution for the clothing is shown in (F).

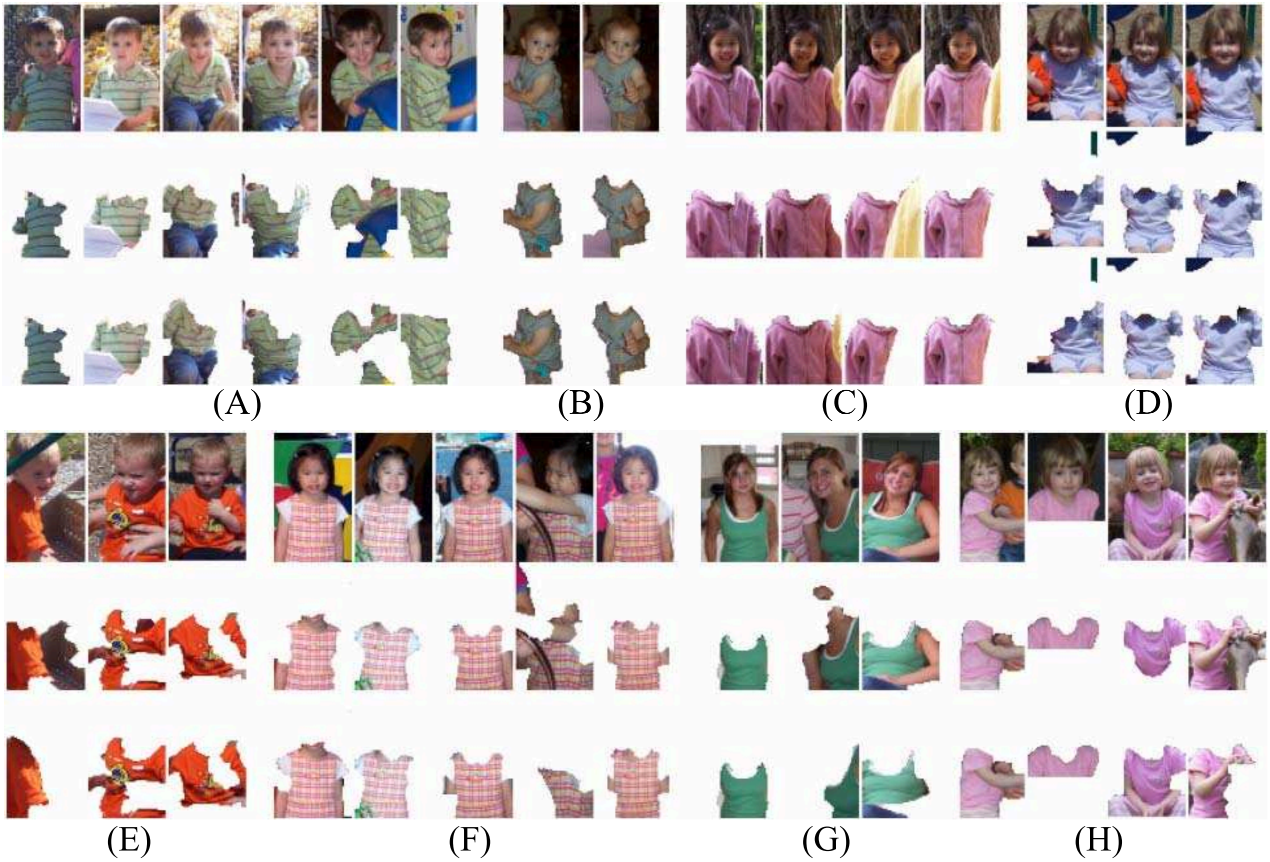

Figure 5. See Section 5. For each group of person images, the top row shows the resized person images, the middle row shows the result of applying graph cuts to segment clothing on each person image individually, and the bottom row shows the result of segmenting the clothing using the entire group of images. Often times, the group graph cut learns a better model for the clothing, and is able to segment out occlusions (A, C, F, H) and adapt to difficult poses (E, G). We do not explicitly exclude flesh, so some flesh remains in the clothing masks (B, G, H).

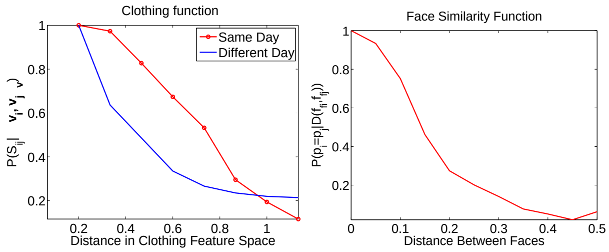

Figure 6. **Left:** The probability that two person images share a common identity given the distance between the clothing features and the time interval between the images. **Right:** In a similar fashion, the probability of two person images sharing a common identity given the distance between faces $\mathbf{f}_i^f$ and $\mathbf{f}_j^f$.

Given an unlabeled person p, $P(p = n|\mathbf{f})$ where $\mathbf{f} = \{\mathbf{f}^f, \mathbf{v}\}$ includes the facial features $\mathbf{f}^f$ and the clothing features $\mathbf{v}$, the probability that the name assigned to person p is n is estimated using nearest neighbors. In our notation, name set $\mathbf{N}$ comprises the names of the U unique individuals in the image collection. An element $n^k \in \mathbf{N}$ is a particular name in the set. The K nearest labeled neighbors of a person p_i are selected from the collection using facial similarity and clothing similarity. When finding the nearest neighbors to a query person with features $\mathbf{f}$, both the facial and clothing features are considered using the measure P_{ij}, the posterior probability that two person images p_i and p_j are the same individual. We propose the measure of similarity P_{ij} between two person images, where:

$$P_{ij} = P(S_{ij}|\mathbf{f}_i, \mathbf{f}_j, t_i, t_j) \quad (7)$$

$$\approx \max\,[P_{ij}^v, P_{ij}^f] \quad (8)$$

The posterior probability $P_{ij}^v = P(S_{ij}|\langle\mathbf{v}_i, \mathbf{v}_j\rangle_v, |t_i - t_j|)$ that two person images p_i and p_j are the same individual is dependent both on the distance between the clothing features $\langle\mathbf{v}_i, \mathbf{v}_j\rangle_v$ using the visual word representation, and also on the time difference $|t_i - t_j|$ between the image captures. The distance between the clothing features $\langle\mathbf{v}_i, \mathbf{v}_j\rangle_v$ for two person images p_i and p_j is simply the sum of the χ^2 distances between the texture and the color visual word histograms, similar to the superpixel distance in Eq. (1). The probability P_{ij}^v is approximated as a function of the distance $\langle\mathbf{v}_i, \mathbf{v}_j\rangle_v$, learned from a non-test image collection for same-day and different-day pairs of person images with the same identity, and pairs with different identities. Figure 6 shows the maximum likelihood estimate of P_{ij}^v. The posterior is fit with a decaying exponential, one model for person images captured on the same day, and one model for person images captured on different days. Similarly, the probability P_{ij}^f, the probability that faces i and j are the same person, is modeled using a decaying exponential.

We justify the similarity metric P_{ij} based on our observations of how humans perform recognition by combining multi-modal features to judge the similarity between faces. If we see two person images with identical clothing from the same day, we think they are likely the same person, even if the images have such different facial expression facial expressions that a judgement on the faces is difficult. Likewise, if we have high confidence that the faces are similar, we are not dissuaded by seeing that the clothing is different (the person may have put on a sweater, we reason).

Using the metric P_{ij}, a nearest neighbor is one that is similar in either facial appearance or in clothing appearance. These K nearest neighbors are used to estimate $P(p = n|\mathbf{f})$ using a weighted density estimate, which can in turn be used to recognize the face according to:

$$p_{\text{MAP}} = \arg\max_{n \in \mathbf{N}} P(p = n|\mathbf{f}) \quad (9)$$

When multiple people are in an image, there is an additional constraint, called the *unique object constraint*, that no person can appear more than once in an image [4, 26]. We seek the assignment of names to people that maximizes $P(\mathbf{p} = \mathbf{n}|\mathbf{F})$, the posterior of the names for all people in the image, assuming that any group of persons is equally likely. The set of M people in the image is denoted $\mathbf{p}$, $\mathbf{F}$ is the set of all the features $\mathbf{f}$ for all people in the image, and $\mathbf{n}$ is a subset of $\mathbf{N}$ with M elements and is a particular assignment of a name to each person in $\mathbf{p}$. Although there are $\binom{U}{M}$ combinations of names to people, this problem is solved in $O(M^3)$ time using Munkres algorithm [17].

7. Experiments

Better Recognition Improves Clothing Segmentation: The following experiment was performed to evaluate the performance of the graph-cut clothing segmentation. In our Sets 1 and 4, every superpixel of every person image was manually labeled as either clothing or not clothing. This task was difficult, not just due to the sheer number of superpixels (35700 superpixels), but because of the inherent ambiguity of the problem. For our person images, we labeled as clothing any covering of the torso and legs. Uncovered arms were not considered to be clothing, and head coverings such as hats and glasses were also excluded.

We apply our clothing segmentation to each person image in both collections. Table 2 reports the accuracy of the clothing segmentation. We compare the graph cut segmentation against the prior (roughly 70% of the superpixels are *not* clothing). A naïve segmentation is to find the mean value of the clothing mask corresponding to the region covered by each superpixel, then classify as clothing if this value surpasses a threshold. The threshold was selected by minimizing the equal error rate. This method considers only the position of each superpixel and not its feature values. In both collections, using the graph cut clothing segmentation provides a substantial improvement over the naïve approach.

	Set 1	Set 4
Prior	70.7%	68.2%
Naïve	77.2%	84.2%
GC Individual	87.6%	88.5%
GC Group	**88.5%**	**90.3%**

Table 2. Graph cuts provides effective clothing recognition. For each of two image collections, the accuracy of classifying superpixels as either clothing or non-clothing with four different algorithms is shown. Using Graph Cuts for groups of images proves to be the most effective method.

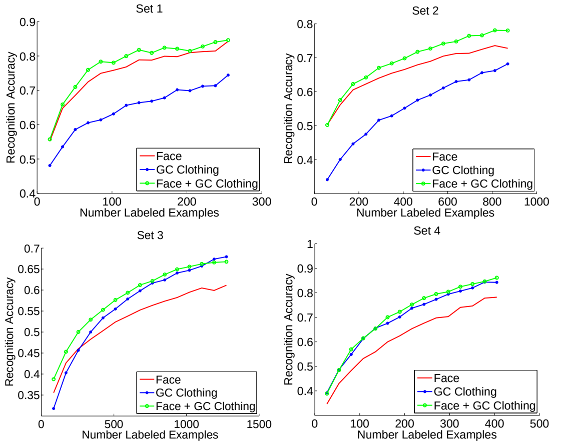

Figure 7. Combining facial and clothing features results in better recognition accuracy than using either feature independently.

Further improvement is achieved when the person images are considered in groups. For this experiment, we assume the ground truth for identity is known, and a group includes all instances of an individual appearance within a 20 minutes time window, nearly ensuring the clothing has not been changed for each individual.

Better Clothing Recognition Improves Recognition: The following experiment is performed to simulate the effect on recognition of labeling faces in an image collection. People images are labeled according to a random order and the identity of all remaining unlabeled faces is inferred by the nearest-neighbor classifier from Section 6. Each classification is compared against the true label to determine the recognition accuracy. We use nine nearest neighbors and repeat the random labeling procedure 50 times to find the average performance. The goal of these experiments is to show the influence of clothing segmentation on recognition.

Figure 7 shows the results of the person recognition experiments. The combination of face and clothing features improves recognition in all of our test sets. If only a single feature type is to be used, the preferred feature depends on the image collection. For this experiment, the clothing features are extracted from the clothing mask determined by graph cuts on each image individually.

Figure 8 compares the performance of recognizing people using only clothing features. For all of our collections, the graph cut clothing masks outperform using only a box (shown in Figure 9). Also, for each collection, the clothing masks are generated by segmenting using group segmentation, and these segmentations unanimously lead to better recognition performance. Finally, we show in collection Sets 1 and 4, where ground-truth labeled clothing masks exist, that the best performance is achieved using the ground truth clothing masks. This represents the maximum possible recognition accuracy that our system could achieve if the clothing segmentation is perfect.

To summarize, these experiments show that:

- Multiple images of the same person improve clothing segmentation.
- Person recognition improves with improvements to the clothing segmentation.

Ongoing work includes merging the recognition and clothing segmentation into a single framework where each assists the other in the following fashion: based on a labeled subset of people, the other people in the collection are recognized. Then, based on these putative identities, new clothing masks are found using multiple images of the same person within a given time window.

8. Publically Available Dataset

One persistant problem for researchers dealing with personal image collections is that there is a lack of standard datasets. As a result, each research group uses their own datasets, and results are difficult to compare. We have made our image Set 2 of 931 labeled people available to the research community [9]. The dataset is described in Table 1, and contains original JPEG captures with all associated EXIF information, as well as text files containing the identity of all labeled individuals. We hope this dataset provides a valuable common ground for the research community.

9. Conclusion

In this paper, we describe the advantages of performing clothing segmentation with graph cuts in a consumer image collection. We showed a data-driven (rather than driven by a human model) approach for finding a global clothing mask that shows the typical location of clothing in person images. Using this global clothing mask, a clothing mask for each person image is found using graph cuts. Further clothing segmentation improvement is attained using multiple images of the same person which allows us to construct a better clothing model.

This work can be viewed as a case study for the merits of combining segmentation and recognition. Improvements in clothing segmentation improve person recognition in consumer image collections. Likewise, using multiple images

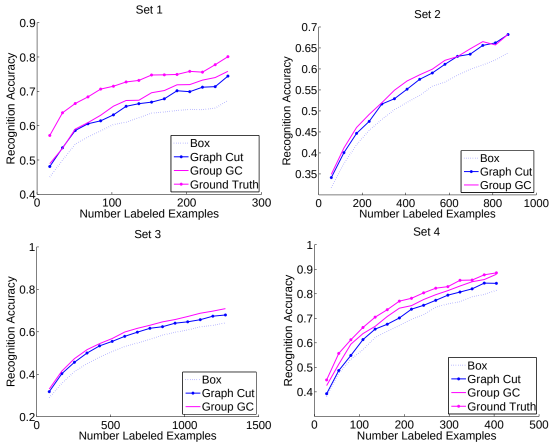

Figure 8. Using graph cuts for the extraction of clothing features improves the accuracy of recognizing people over using a simple box region. Further improvement is attained by using multiple person images when performing clothing segmentation. Sets 1 and 4 demonstrate even more room for improvement when ground-truth clothing segmention is used for feature extraction.

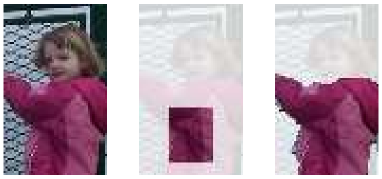

Figure 9. Given an image (left), using the clothing features from a graph cut clothing mask (right) results in superior recognition to using a box (middle).

of the same person improves the results of clothing segmentation. We are working on the next steps by merging the clothing segmentation and person recognition into a framework where each assists the other, and anticipate applying these concepts to other computer vision problems as well.

Acknowledgements. We would like to thank Alexei Efros and Jean-François Lalonde for enlightening and entertaining discussions on the topics addressed here, Lei Zhang for providing access to his image data, and Shai Bagon for sharing his code.

References

[1] D. Anguelov, K.-C. Lee, S. Burak, Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *Proc. CVPR*, 2007.

[2] S. Bagon. Matlab wrapper for graph cut. Downloaded July 2007 from the Weizmann Institute. http://www.wisdom.weizmann.ac.il/~bagon.

[3] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI*, 1997.

[4] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proc. CVPR*, 2004.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 2004.

[6] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *PAMI*, 2001.

[7] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *Proc. CVPR*, 2006.

[8] I. Cohen, A. Garg, and T. Huang. Vision-based overhead view person recognition. In *ICPR*, page 5119, 2000.

[9] A. Gallagher. Consumer image person recognition database. http://amp.ece.cmu.edu/downloads.htm.

[10] M. Jones and P. Viola. Fast multiview face detector. In *Proc. CVPR*, 2003.

[11] D. Klünder, M. Hähnel, and K.-F. Kraiss. Color and texture features for person recognition. In *IJCNN*, 2004.

[12] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 2004.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[14] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[15] D. Liu and T. Chen. Background cutout with automatic object discovery. In *Proc. ICIP*, 2007.

[16] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proc. CVPR*, 2004.

[17] J. Munkres. Algorithms for the assignment and transportation problems. *SIAM*, 1957.

[18] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 2003.

[19] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, 2005.

[20] C. Rother, V. Kolomogorov, and A. Blake. Grabcut- interactive foreground extraction using iterated graph cuts. In *Proc. ACM Siggraph*, 2004.

[21] C. Rother, T. Minka, A. Blake, and V. Kolomogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proc. CVPR*, 2004.

[22] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. CVPR*, 2006.

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 2000.

[24] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *Proc. ICAFGR*, May 2002.

[25] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[26] J. Sivic, C. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proc. BMVC*, 2006.

[27] Y. Song and T. Leung. Context-aided human recognition- clustering. In *Proc. ECCV*, 2006.

[28] N. Sprague and J. Luo. Clothed people detection in still images. In *Proc. ICPR*, 2002.

[29] R. X. Y. Tian, W. Liu, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In *Proc. CVPR*, 2007.

[30] S. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. In *Proc. NIPS*, 2002.

[31] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *Proc. MM*, 2003.

[32] L. Zhang, Y. Hu, M. Li, and H. Zhang. Efficient propagation for face annotation in family albums. In *Proc. MM*, 2004.