Automatic Surveillance Video Matting Using a Shape Prior

Ting Yu Xiaoming Liu Sernam Lim Nils O. Krahnstoever Peter H. Tu
{yut,liux,limser,krahnsto,tu}@research.ge.com

Computer Vision Lab, GE Global Research, Niskayuna, NY 12309

Abstract

This paper presents a novel algorithm for performing video matting, which is built upon a proposed image matting algorithm that is fully automatic. The proposed algorithm utilizes a PCA-based shape model as a prior for guiding the matting process, so that manual interactions required by most existing image matting methods are unnecessary. We specifically consider a surveillance environment in which foreground windows are identified via a person detector. By applying the image matting algorithm to these foreground windows, on a per frame basis, we aim to fully automate the video matting process. Due to the inherent inaccuracy of any person detector, it is critical that the shape model be aligned with the object. We achieve this in a framework where the estimation of the alpha matte guided by the shape prior model, and the alignment process are simultaneously optimized based on a quadratic cost function. We report very promising results on a people data set collected from surveillance environments.

1. Introduction

Image matting is an important operation in photo editing [2, 11, 12, 7] that allows the user to extract and composite a foreground region onto a background of choice, and is accomplished by estimating the foreground opacity or "alpha matte" at every pixel and extracting those pixels that have a high foreground opacity. The biggest challenge here lies in extracting, with high confidence, initial foreground and background regions that would then guide the matting process in fully determining the foreground opacity at every pixel. To accomplish this, most existing methods, such as [14, 12], rely on manual input that indicates foreground and background regions. Most notably, the work by Levin et al. [9, 10] are excellent examples of recent advances along this direction, where the alpha matte can be estimated efficiently in close form through an elegant formulation of a quadratic cost function [9].

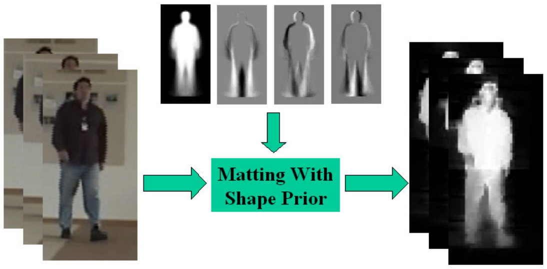

Figure 1. Guided by a shape prior, and given image windows from video, we present a fully automatic matting algorithm.

The use of manual interactions is, however, unsuitable for performing video matting, a process in which one wishes to estimate the matte of a foreground object from a video sequence. This is a much more challenging problem when compared to image matting, since it is obvious that manually marking foreground and background regions for every frame of a video sequence comprising a large number of image frames is impractical. Several attempts to automate the matting process includes the work in [2], where foreground and background regions in keyframes are marked, followed by interpolation based on background and optical flow estimation. Along the same direction, a recent paper by Levin et al. in [10] has proposed an extension of their matting work [9] towards unsupervised matting by utilizing cues from spectral segmentation. Since it is well known that the image segmentation problem itself is an ill-posed problem, manual interactions are inevitable if one wishes to achieve a reasonable level of accuracy. In this paper, we sought a matting algorithm that is fully automatic, but yet capable of producing good results. We particularly consider a surveillance environment, whereby a person detector would provide us with foreground windows, per frame, in which we can perform image matting automatically (Figure 1). In this context, it is important to point out that a typical person detector (e.g., [15, 17]) is inherently inaccurate, and an automatic matting algorithm that is capable of determining foreground opacity accurately would be extremely valuable.

Towards this end, this paper proposes a fully automatic algorithm built upon the work by Levin et al. [9], who proposed a quadratic cost function that could efficiently compute the matte in closed form. We incorporate into this quadratic cost function a shape prior that has been built from a set of training data, which essentially replaces the manual inputs required to "kickstart" the matting process. The application of shape prior in this paper has been largely motivated by its successful application to several domains, most notably object segmentation, where learned shape priors are used for guiding the segmentation process [3, 4, 8, 13], as opposed to segmentation algorithms such as [1, 11, 12] that require manual interactions. Additionally, we adopt a PCA-based approach towards learning the shape priors. Such approach can also be found in other work that include the PCA-based shape representation used in the level-set segmentation algorithm described in [13]. More sophisticated shape model can also be found in work such as [3].

Given our interest in the surveillance domain, it is important to consider the spatial alignment of the shape prior to the object, since, as mentioned, it is common knowledge that existing person detectors are incapable of providing "perfect" foreground windows. In fact, it is typical to see an offset between the true location of the person and the center of the window due to localization uncertianties in the training data. Even under the assumption of static background so that typical background subtraction algorithm can be employed, the presence of shadows or changes in lighting would still cause such a mis-alignment problem. To address this issue, we will show that the unknown transformation parameters of aligning the shape prior to the detection window can be recovered using Gauss-Newton method simultaneously during the optimization process.

The contributions of this paper can thus be summarized as follow: (1) a fully automatic image matting algorithm, guided by a shape model, is proposed towards achieving automatic video matting in surveillance environments, (2) in contrast to previous work that uses shape priors, which still need small amount of manual interactions to help deal with mis-alignment between the applied shape prior and the object region, our approach is capable of recovering the transformation parameters automatically, and (3) our approach elegantly unifies the estimation of the matte guided by the shape model and the alignment of the shape prior with the object in a single objective function.

The rest of the paper is organized as follow. In the next section, we will look at the objective function, to which we will incorporate the shape model in Section 3 and the alignment parameters in Section 4. Optimization of the unified objective function is then described in Section 5. Finally, experimental results are given in Section 6, and conclusions in Section 7.

2. Laplacian Matting

To compute the alpha matte given an image I, one can consider the color of the i^{th} pixel, I_i, as a linear combination of the foreground and background colors

$$I_i = \alpha_i F_i + (1-\alpha_i)B_i, \tag{1}$$

where α_i, referred to as the foreground opacity, controls the contribution of the foreground, F_i, and background, B_i, terms. Estimating these unknown quantities is, however, an underconstrained problem, since each pixel of a 3-channel color image would be associated with three equations and seven unknowns.

Consequently, it is impossible to solve for Eq. 1 without introducing additional constraints. Such constraints have indeed been proposed previously by Levin et al. [9]. They proved that if an assumption of *color linear model* could be made, then in a small window, w, around each pixel, F and B could be represented with a linear mixture of two colors. As a result the compositing equation in Eq. 1 can be transformed and approximated by a 4D linear model:

$$\alpha_i \approx \sum_c a^c I_i^c + b, \qquad \forall i \in w \tag{2}$$

where I_i^c is the c_{th} channel color value of i_{th} pixel, and a^c and b are unknown variables related to the foreground and background colors of pixels in the local window w.

By manipulating Eq. 2, Levin et al. [9] derived a cost function that is quadratic in α and in which the a and b terms can be eliminated

$$J(\alpha) = \alpha^T L \alpha. \tag{3}$$

L, referred to as the *matting Laplacian*, is a square matrix of size $M \times M$, that captures the local color properties of the input image containing M pixels. Its $(i,j)^{th}$ element is given as

$$\sum_{k|(i,j)\in w_k} (\delta_{ij} - \frac{1}{|w_k|}(1+(I_i-\mu_k)^T(\Sigma_k+\frac{\epsilon}{|w_k|}I_3)^{-1}(I_j-\mu_k))), \tag{4}$$

where δ_{ij} is the Kronecker delta function. Within w_k, the color distribution is described by a 3×3 covariance matrix, Σ_k, and a 3×1 vector, μ_k, representing the mean pixel colors. I_3 is a 3×3 identity matrix.

If there are no other constraints, it is obvious that any α vector that lies in the null space of L constitutes a valid solution. On the other hand, any meaningful solution would have to be consistent with a well-defined notion of the foreground and background regions. To obtain such information, we can rely on manual interactions for explicitly marking initial foreground and background regions so that a valid

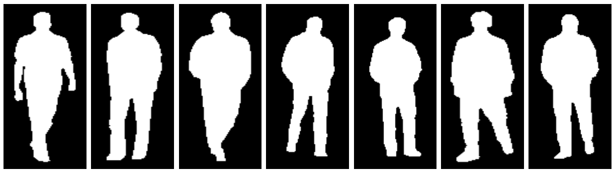

Figure 2. Sample training images used to learn the PCA-based shape prior model.

solution can subsequently be obtained by minimizing

$$\arg\min_{\alpha} J(\alpha) = \arg\min_{\alpha} \alpha^T L\alpha + \lambda(\alpha - b_s)^T D_s(\alpha - b_s), \quad (5)$$

where λ is a weighting factor, D_s is a diagonal matrix whose diagonal elements contain 1 for marked pixels and 0 otherwise, and b_s is a vector containing the user-specified alpha values for the marked pixels and 0 for all other pixels. The optimal solution can then be obtained by computing the derivative of Eq. 5 over α, setting it to 0, and then solving a sparse linear system equation as follows

$$(L + \lambda D_s)\alpha = \lambda b_s. \quad (6)$$

3. Adding Shape Prior

While the closed form solution in Eq. 6 is appealing, its dependency on manual interactions makes it unsuitable for video matting. The task of marking foreground and background regions in every frame of a video sequence is prohibitive. To overcome such a problem, we propose in this paper the utilization of a shape model that would essentially be used to replace manual interactions.

Given a shape database for an object category of interest, $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$, where S_i is the i^{th} shape training data represented as a binary map and all shape images are spatially registered, we train a PCA-based shape prior model through eigen-analysis. The trained model can then be used to represent a shape as

$$S(u) = Vu + \Delta = \sum_{i=1}^{N} V_i u_i + \Delta, \quad (7)$$

where Δ is the mean shape, $V = [V_1, V_2, \ldots, V_N]$ are the shape bases, and $u = [u_1, u_2, \ldots, u_N]$ are the basis coefficients. Figure 2 shows some training examples we have used to learn the shape model for walking people. Some learned PCA-shape bases are shown in Figure 1. Incorporating such a shape prior model would then modify the cost function to

$$\begin{aligned} &\arg\min_{\alpha,u} J(\alpha, u) \\ &= \arg\min_{\alpha,u} \alpha^T L\alpha + \lambda(\alpha - (Vu + \Delta))^T(\alpha - (Vu + \Delta)), \end{aligned} \quad (8)$$

which can be easily solved with the following sparse linear system

$$\begin{pmatrix} (L + \lambda I) & -\lambda V \\ -V^T & V^T V \end{pmatrix} \begin{pmatrix} \alpha \\ u \end{pmatrix} = \begin{pmatrix} \lambda\Delta \\ -V^T\Delta \end{pmatrix}. \quad (9)$$

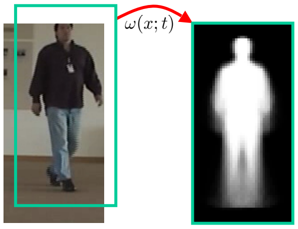

Figure 3. Alignment through a spatial transformation.

4. Shape-Object Alignment

So far, we have implicitly assumed that the shape model is properly aligned with the object. Such an assumption is frequently violated, particularly due to our interest in the surveillance domain, where it is impractical to assume that the foreground window provided by a person detector is well-aligned with the object. The spatial transformation that would re-align the shape model with the object (see Figure 3 for an example) is, however, an unknown property until we can correctly solve for the foreground matte. To overcome such a dilemma, we propose solving both estimation problems simultaneously through the following iterative optimization process.

Let $\omega(x; t)$ be the spatial transformation that maps a pixel from an image location, x, to a location $\omega(x; t)$ in the shape model coordinate system. Here, $t = [t_1, t_2, \ldots, t_q]$ denotes the unknown parameter vector of ω. It is also important to point out that the spatial transformation from image to shape model, as opposed to the reverse, is computationally desirable since we only need to compute the Laplacian matrix, L, once for each input image (note that such an "input image", in our case, would come from the foreground window provided by a person detector).

After applying the transformation to obtain $V(\omega(x; t)) = [V_1(\omega(x; t)), V_2(\omega(x; t)), \ldots, V_N(\omega(x; t))]$, and mean shape $\Delta(\omega(x; t))$, our task is then to find an optimal (α, u, t) that minimizes the quadratic cost defined over L, i.e., we have

$$\begin{aligned} &\arg\min_{\alpha,u,t} J(\alpha, u, t) \\ &= \arg\min_{\alpha,u,t} \alpha^T L\alpha + \lambda\|\alpha - (V(\omega(x; t))u + \Delta(\omega(x; t)))\|. \end{aligned} \quad (10)$$

With this formulation, there are three unknowns to be estimated simultaneously, namely α the unknown matte, u the shape basis coefficients, and t the transformation parameters. Such a cost function is quadratic over α and u, but nonconvex over t, since $V_i(\omega(x; t))$ is essentially nonlinear over $\omega(x; t)$, and solving it may require some type of costly global optimization procedure. For this reason, we make a concession and assume that the unknown center of the object is near the center, t_0, of the input image (see Figure 3), which is a valid assumption in most cases (e.g., [6]). Starting from t_0, we can now iteratively solve for a transformation update δt through the Gauss-Newton method. Specifically, we propose to solve these three unknowns in two iterative steps, which update (α, u) and t respectively.

5. Unified Optimization

5.1. Solving α and u

Given an updated transformation parameter $t^{'} = t + \delta t$, we warp the shape model as $V^{'} = [V_1(\omega(x;t^{'})), V_2(\omega(x;t^{'})), \ldots, V_N(\omega(x;t^{'}))]$ and mean shape $\Delta^{'} = \Delta(\omega(x;t^{'}))$, and solve α and u using Eq. 9.

Recall that the left hand side matrix (LHSM) of the linear equation in Eq. 9 is a block matrix comprising four blocks. $(L + \lambda I)$ is the largest sub-matrix in this LHSM with dimension $M \times M$, where M is the number of pixels in the input image. $V^{'}$, the shape prior space, is a $M \times N$ matrix, where N is the number of learned shape bases and is typically much smaller than M. In addition, the Laplacian matrix L does not change during iteration due to the spatial transformation from image space to shape model space. As a result, the inverse of $(L + \lambda I)$ need only be computed once. We then compute the inverse of LHSM using the efficient method in [5]. Though this block matrix inverse involves an inverse computation of a $M \times M$ sub-matrix, defined by $[(L+\lambda I) - \lambda V^{'}(V^{'T}V^{'})^{-1}V^{'T}]$, we can, due to its symmetric form, simplify this inverse operation by using matrix inversion lemma [5]. This means that in every iteration, only the inverse of the $V^{'T}V^{'}$ needs to be computed, which is only a $N \times N$ matrix, and thus much cheaper to compute.

5.2. Solving t

Once we obtained an updated $(\alpha^{'}, u^{'})$ and current estimate t, we look for an update δt that minimizes the cost function in Eq. 10. Since we fix $(\alpha^{'}, u^{'})$ at this stage, it is equivalent to minimizing the quantity

$$\begin{aligned}
&\arg\min_{\delta t} J(\delta t) \\
&= \arg\min_{\delta t} \|\alpha^{'} - (V(\omega(x;t+\delta t))u^{'} + \Delta(\omega(x;t+\delta t)))\| \\
&= \arg\min_{\delta t} \|\alpha^{'} - (\sum_{i=1}^{N} V_i(\omega(x;t+\delta t))u_i^{'} + \Delta(\omega(x;t+\delta t)))\|.
\end{aligned} \tag{11}$$

The term $V_i(\omega(x;t+\delta t))$ could then be expanded using a first-taylor expansion around its current t, i.e.,

$$V_i(\omega(x;t+\delta t)) = V_i(\omega(x;t)) + J_i(t)\delta t, \tag{12}$$

where J_i is the *Jacobian matrix* of V_i with respect to t. This is a $M \times q$ matrix that could be written in column form as

$$J_i(t) = [V_{i,t_1}\omega(x;t)|V_{i,t_2}\omega(x;t)|\ldots|V_{i,t_q}\omega(x;t)]. \tag{13}$$

Similar linear expansion can also be applied to the mean shape vector $\Delta(\omega(x;t+\delta t))$ to obtain its Jacobian $J_{\Delta}(t)$.

After expansion, the cost function defined in Eq. 11 becomes quadratic with respect to δt, so that the solution can now be obtained in closed form by solving a linear equation. The biggest problem, however, is that we are faced with computing the Jacobian matrices $J_i(t)$ of all shape bases $V_i, i \in [i, N]$ during each iteration, which is expensive.

This computational burden can, fortunately, be reduced by realizing that we do not have to compute the Jacobian terms for the shape basis and mean shape separately due to the linear relationship between them. Rather, we can define a new term as

$$\beta(\omega(x;t+\delta t)) = \sum_{i=1}^{N} V_i(\omega(x;t+\delta t))u_i^{'} + \Delta(\omega(x;t+\delta t)), \tag{14}$$

where $\beta(\omega(x;t+\delta t))$ is essentially the reconstructed matte from the updated shape prior, and conduct a taylor-expansion around the new term instead. The transformation update δt can now be derived as

$$\delta t = (J_{\beta}(t)^T J_{\beta}(t))^{-1} J_{\beta}(t)^T (\alpha^{'} - \beta(\omega(x;t))), \tag{15}$$

which solves a $q \times q$ matrix inverse problem, and thus can be computed very efficiently.

The above two-step optimization is then conducted iteratively until either the maximum number of iterations allowed is reached or little improvement is observed, noting that we have found that a good solution can typically be found within 20 iterations.

6. Experiments

The proposed matting algorithm was evaluated with real-world images of people walking in typical indoor and outdoor surveillance environments. In all experiments, we set $\lambda = 0.01$.

We first evaluated our algorithm on still images and compare the results quantitatively with the method due to Levin et al. [9]. It is important to point out that the latter method was run with extensive manual interactions, whereby foreground and background regions were repeatedly marked as required to get the best possible matting results. Comparatively, we run our method in a fully unsupervised manner. That is, the goal here is to demonstrate quantitatively the “closeness” of the performance of our method to the user-guided method, with the expectation that the user-guided method would logically produce better results.

Upon establishing from the still image experiments the efficacy of our method, we proceeded to conduct experiments for evaluating the utility of our method when applied to video sequence. The quality of the video sequences used in these experiments, being captured from typical CCTV cameras, is naturally much poorer than those datasets used

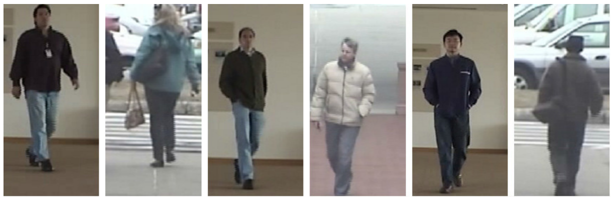

Figure 4. A subset of the test samples used in our experiments, captured from both indoor and outdoor surveillance scenes.

in most previous work. Despite that, the video results demonstrate the capability of our method in consistently producing good matte maps, unsupervised, for a video sequence. Testing on these real-world video sequences thus reinforce the efficacy of our method for practical usage.

6.1. Shape Database

To learn the PCA-based shape prior, we manually labeled 215 binary images of the foreground of people. Each training image was resized to a standard 80×40 pixels, i.e., $M = 3200$, and spatially aligned. Some training samples are shown in Figure 2. We kept 99% of the sample covariance, leading to a total of $N = 182$ shape bases. A subset of learned PCA shape bases is shown in Figure 1. The set of shape bases was then used for both the still image and video experiments as shape priors.

6.2. Still Image Results

A set of 375 images were then collected, and the image patch containing the walking people was cropped and used as input, as shown in Figure 4. Note that while the windows containing subjects were manually selected here, so that we have the ground truths, in the real applications, these windows are presumably provided by a person detector. To simulate the scenario that the foreground window may not be well aligned with the center of the image window, we then randomly perturb the bounding box location around the true location. Based on the ground truths, we proceeded to measure the accuracy of the spatial aligning capability of our algorithm. For simplicity, we applied only translation to the test images, i.e., $\omega(x;t)$ does not contain any rotational component. In practice, specifically under a surveillance context where people are expected to walk upright, this is generally a valid simplification. However, as presented in Section 4, complex transformation should also be recoverable under our framework. We also selected a subset of test samples to perform a pixel-wise quantitative comparison with Levin's algorithm [9]. As mentioned, manual interactions were provided as required to achieve the best results from the algorithm. We then compared these results with those of our algorithm by computing pixel-wise differences in the matte values.

In Figure 5, we first show some of the matting results using our algorithm. The top row displays the original test images, and the bottom row shows the computed mattes. As seen in the figure, the matte maps are qualitatively able to match the body shapes, with the poses correctly estimated. In this case, even though there were significant distractions from background clutters (notice the cars in Figure 5), the algorithm manages to return good matte maps that are spatial aligned automatically.

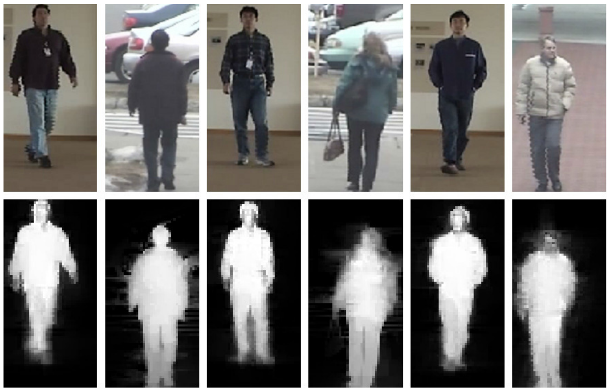

Figure 5. Sample matting results using the proposed approach. Top Row: original test images, Bottom Row: matting results. For more results, please refer to the supplemental video submission.

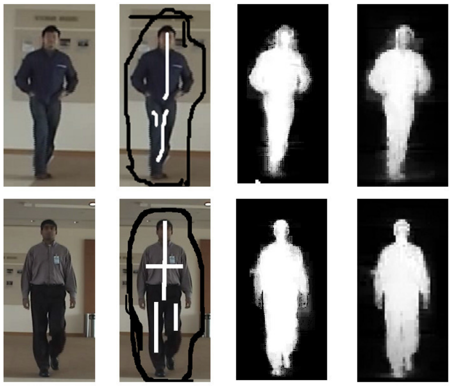

Figure 6. Comparative study between the results obtained using the proposed approach and Levin's method [9]. Left Col: original test images, Second Col: manual stroke inputs, Third Col: results using Levin's method [9], Right Col: results from the proposed approach.

The results returned by our automatic approach and Levin's method [9] are shown in Figure 6. The left column shows the original images, the second column displays manual inputs required by Levin's method [9], the third column shows the results using Levin's method [9], and the right column shows our results. Qualitatively, the difference between these two approaches are relatively minor. The main difference is that the matte map obtained through our method is relatively darker, due to the PCA-based shape prior, which provides a probabilistic score instead of a bi-

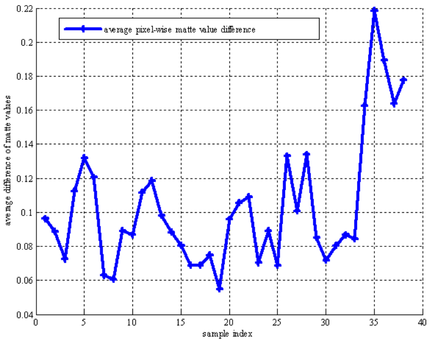

Figure 7. The average pixel-wise matte difference between our results and the results obtained by Levn's method [9].

narized score as in Levin's method [9].

Figure 7 plots the average pixel-wise matte difference between the results returned by our approach and those by Levin's method [9] for a subset of the test images. The average difference here is less than 0.14, which is a good indication that our unsupervised method achieves matting accuracy close to the interactive matting method.

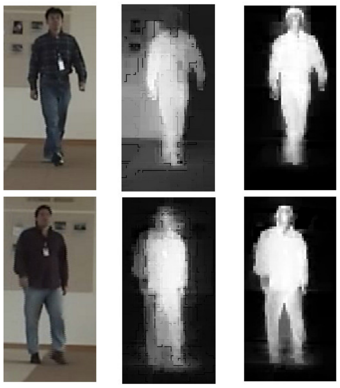

Figure 8. Comparative study between the results obtained using the proposed approach and those obtained using shape prior without spatial alignment. Left Col: original test images, Middle Col: results using shape prior without spatial alignment, Right Col: results with the full approach. Without spatial alignment, the matting results are much worse.

We also look at the qualitative effect of not conducting spatial alignment. Figure 8 demonstrates the necessity of spatial alignment. The left column shows the original images, the middle column displays the results obtained by directly applying the shape prior without dealing with alignment, and the right column shows the results of the full approach. It is obvious that without spatially aligning the shape priors, the results are much worse.

Finally, to quantify the accuracy of conducting spatial alignment simultaneously in the optimization process, we compared the spatial alignment estimated by our algorithm and the actual amount of alignment required based on the ground truths, as shown in Figure 9, using the same image subset in Figure 7. The shifting is computed along x and y direction respectively. In most cases, the estimated amount is close to the actual amount. The average alignment error between the estimation and ground truth over this subset is 1.25 pixels and 2.18 pixels along x and y direction respectively. Overall, with respect to the size of the test images (80×40), the performance of the proposed alignment algorithm is very promising.

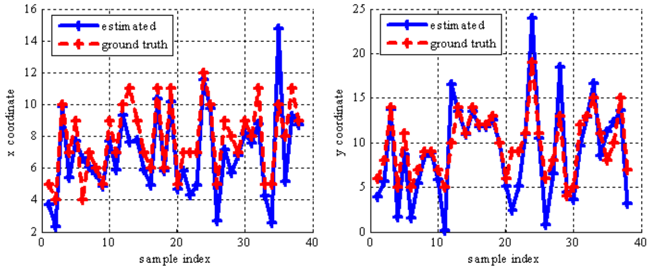

Figure 9. The estimated shift and ground truth shift on a subset along X direction (Left) and Y direction (Right).

6.3. Video Results

At this point, the results obtained from the still image experiments are very promising, and we proceed on to demonstrate the true value of our algorithm, which is its capability to perform unsupervised video matting. We applied our method, on a per frame basis, to an indoor and outdoor sequence. Figure 10 shows the results, where each row of video frames is followed by the corresponding matte maps. The results demonstrate the utility of our algorithm for performing unsupervised video matting.

7. Conclusions

7.1. Summary

We have presented a fully automatic matting algorithm, and shown that the algorithm is capable of consistently generating good matting results when applied to video sequences. Towards achieving a fully unsupervised matting algorithm, we conjecture that utilizing shape priors is more reliable than, for example, cues from spectral segmentation as proposed by [10], due to the lower ambiguities. Here, we reiterate two important points. Firstly, we were able to perform video matting in a fully unsupervised fashion while producing good matting results. This, in our view, has made a significant contribution towards automatic video matting; our experiences with most existing methods reveal that it is very hard to achieve good matting results without extensive manual interactions. Secondly, the difficulty of our experimental setup is clear. We had to work with video sequences

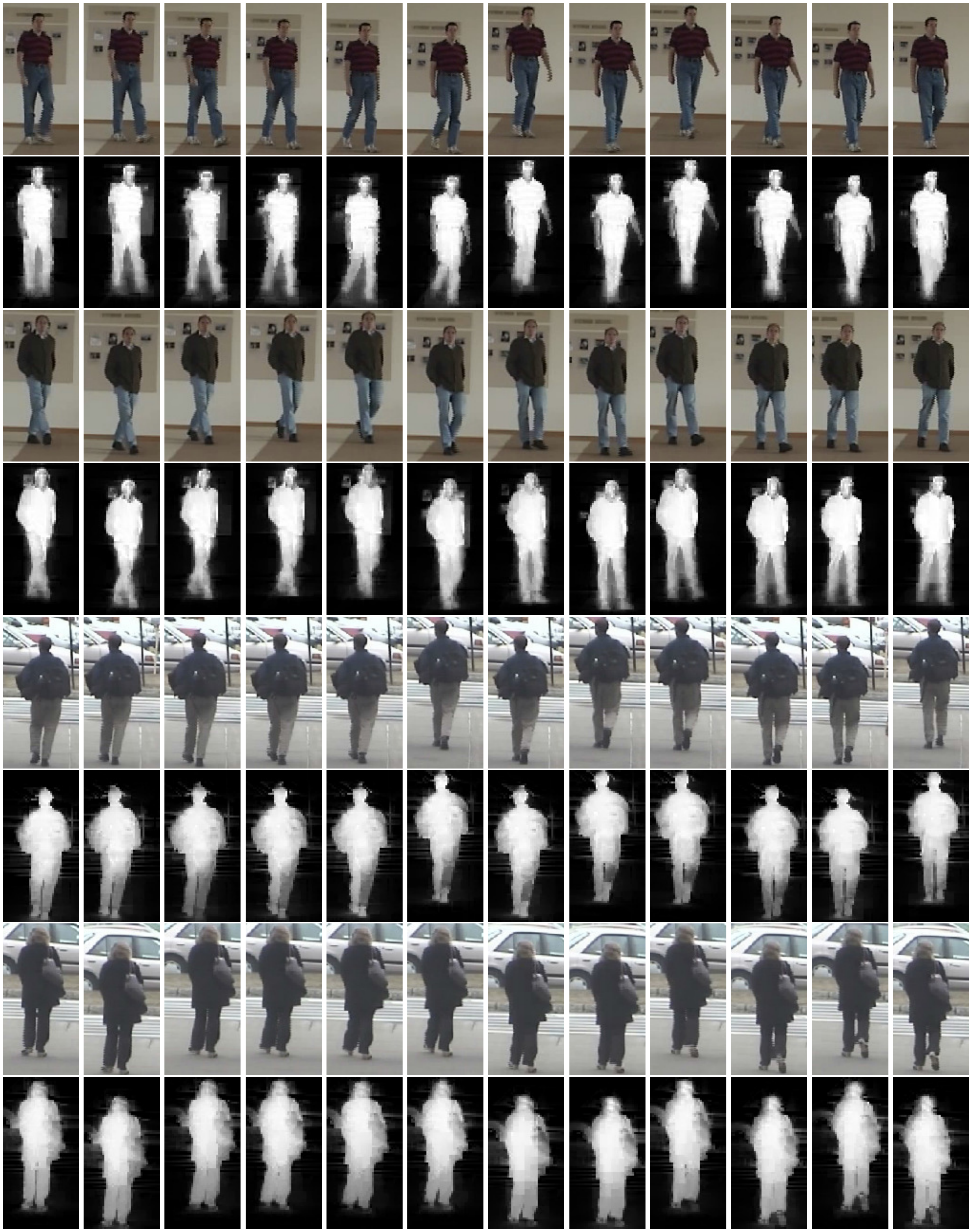

Figure 10. Video matting results.

acquired from typical CCTV cameras, of which the quality are generally poor.

7.2. Extensions and Applications

However, such an experimental setup is necessary as we are motivated by the successful application of our algorithm to the surveillance domain, which demands an algorithm that works in practical situations. In a surveillance context, several potential applications of our algorithm can be considered. One example where such an automatic matting algorithm can be extremely useful is the area of foreground detection. It is commonly understood that the computed matte can be used to guide the foreground detection process, but the supervised nature of the matting process has so far prohibited such synergy. In the surveillance community, researchers are also frequently faced with difficulties in conducting experiments, where privacy issues often prevent them from running experiments on unsuspecting subjects, or, where there is often a lack of subjects. The need for video synthesis is becoming increasingly evident due to such reasons, but synthesizing video is unfortunately a very hard problem. An automatic matting algorithm is however a big step forward in this direction where the user could conceivably extract participating subjects from video sequences for the purpose of synthesizing new video sequences. Additionally, visual appearance modeling and signature learning of humans, which are mainly used for person tracking and identity recognition in a typical surveillance scene, can also benefit from this automatic matting method [16], because a soft matte provides a more detailed confidence measure for these methods to choose the right pixels in order to learn their models.

7.3. Limitations

Despite the promising results and applicability of our method, the biggest challenge we faced comes from the lack of details in our results. For most existing work, the level of details in the reported results were excellent. This can be seen in results reported by [2], [9], etc, where the matting algorithm was able to extract contour of object as fine as, say, a strand of hair. Admittedly, our algorithm in its current form is not able to match these algorithms from this perspective. While this can be credited to the use of manual interactions in these algorithms, the level of details that can be captured by a shape prior is also limited. On the other hand, as mentioned, it is with the guidance of the shape prior that we are able to automate the video matting process. Possible future work could thus extend our algorithm to preserve desirable details in the extracted matte map.

References

[1] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *In Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, 2001. 2

[2] Y. Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski. Video matting of complex scenes. *ACM Transaction on Graphics*, 21(3):243–248, 2002. 1, 8

[3] D. Cremers, T. Kohlberger, and C. Schnörr. Nonlinear shape statistics in mumford-shah based segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, volume 2, pages 93–108, 2002. 2

[4] D. Freedman and T. Zhang. Interactive graph cut based segmentation with shape priors. In *IEEE conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 755–762, June 2005. 2

[5] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996. 4

[6] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 20(10):1025–1039, 1998. 3

[7] K. He, J. Sun, and X. Tang. Fast matting using large kernel matting laplacian matrices. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, June 2010. 1

[8] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18–25, San Diego, CA, June 2005. 2

[9] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York City, June 2006. 1, 2, 4, 5, 6, 8

[10] A. Levin, A. Rav-Acha, and D. Lischinski. Spectral matting. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, June 2007. 1, 6

[11] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum. Lazy snapping. *ACM Transaction on Graphics*, 23(3):303–308, 2004. 1, 2

[12] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Transaction on Graphics*, 23(3):309–314, 2004. 1, 2

[13] D. Shen and C. Davatzikos. An adaptive-focus deformable model using statistical and geometric information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):906–913, 2000. 2

[14] D. Singaraju, C. Rother, and C. Rhemann. New appearance models for natural image matting. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, June 2009. 1

[15] P. F. M. P. Tuzel, O. Human detection via classification on riemannian manifolds. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, June 2007. 1

[16] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *In Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, 2007. 8

[17] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *In Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, volume I, pages 90–97, 2005. 1