This paper appears in *Proceedings of the Conference on Computer Vistion and Pattern Recognition*, pp. 656-662, Santa Barbera, June 1998

A radial cumulative similarity transform for robust image correspondence

T. Darrell
Interval Research Corp.
1801C Page Mill Road
Palo Alto CA 94304
trevor@interval.com
http://www.interval.com/papers/1997-090

Abstract

We develop a local image-correspondence algorithm which performs well near occluding boundaries. Unlike traditional robust methods, our method can find correspondences when the only contrast present is the occluding boundary itself and when the sign of contrast along the boundary is possibly reversed. We define a new image transform which characterizes local image homogeneity, defined as an attribute value in a central region and a function describing the surrounding local similarity structure. In this paper we use radial similarity functions and color attributes; within each window we compute the central color and an image with the cumulative probability that color is unchanged along a ray from the center to a given point in the window. This representation is insensitive to structure outside an occluding boundary, but can model the boundary itself. We show comparative results tracking finger, mouth, and eye features.

1 Introduction

Finding corresponding points in image pairs or image sequences is a central problem in computer vision. Most classical methods assume brightness constancy, and perform best when tracking high-contrast regions that lie on a single surface. However, many images have visually important features that violate this assumption. Developing methods to track corresponding points which lie on occluding boundaries is necessary if one is to track complicated objects with multiple articulated surfaces, such as the human face.

In recent years, robust estimation methods have been applied to image correspondence, and have been shown to considerably improve performance in cases of occlusion. Black and Anandan pioneered robust optic flow using redescending error norms that substantially discount the effect of outliers [1]. Shizawa and Mase derived methods for transparent local flow estimation [2]. Bhat and Nayar have advocated the use of rank statistics for robust correspondence [4]; Zabih and Woodfill use ordering statistics combined with spatial structure in the CENSUS transform [5]. Several authors have explored methods of finding image "layers" to pool motion information over arbitrarily shaped regions of support and to iteratively refine parameter estimates [6, 8, 7], but these methods generally assume models of global object motion to define coherence.

However, these methods make a critical assumption: that there will be sufficient contrast in the foreground ("inlier") portion of an analysis window to localize the corre-

Figure 1. Correspondence is difficult when a uniform surface moves across different background patterns. Consider the correspondence of window A with windows B or C; traditional robust methods equate the match between A:B and A:C, since the "outlier" regions in each is equally different.

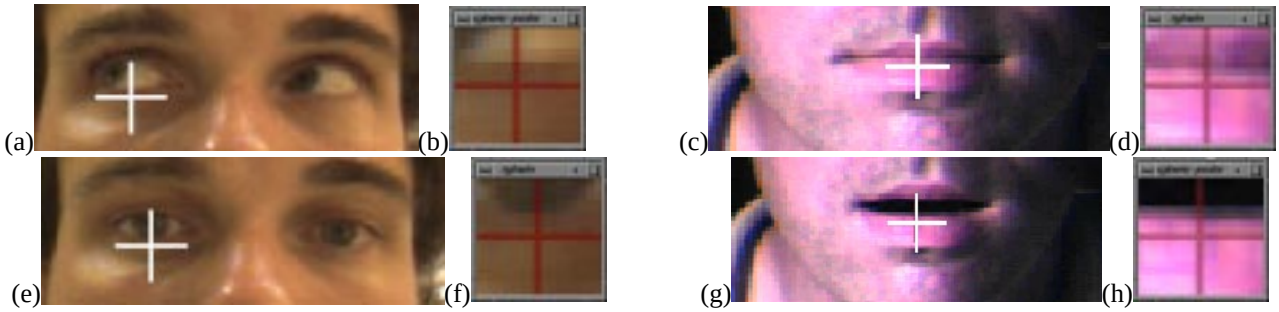


Figure 2. Finding local correspondences in regions with occlusion is a difficult challenge. (a,e) and (c,g) are images taken before and after user's expression changes; (b,f) and (d,h) are enlarged views of corresponding points, with a cross drawn to indicate the center point of the window. Traditional correspondence methods have difficulty at points such as these, where there is little foreground texture, substantial occlusion, and variable sign of contrast at the occlusion boundary.

spondence match. This is often not true, due either to a uniform foreground surface or low-resolution video sampling. This problem is illustrated in Figure 1, which shows a foreground region with zero contrast in front of two different background regions; note that the sign of contrast changes at the occlusion boundary between the two frames. A example in real imagery is shown in Figure 2; the marked locations pose a considerable challenge for existing robust correspondence methods, since any window large enough to include substantial foreground contrast will include a very large percentage of outliers.

Most robust and non-robust correspondence methods fail when there is no coherent foreground contrast. Transparent-motion analysis [2, 3, 9, 10] can potentially detect motion in these difficult cases, but has not, to date, been able to provide precise spatial localization of corresponding points. Smoothing methods such as regularization or parametric motion constraints (affine [11, 12, 13] or learned from examples [14]) can provide approximate localization when good motion estimates are available in nearby image regions, but this is not always the case. If a corpus of training images is available, techniques for feature or appearance modeling can solve these problems, c.f. [18, 19].

For many detailed image analysis/synthesis tasks, finding precise correspondences such as shown in these figures is extremely important. Image compositing [15], automatic morphing [16], and video resynthesis [17], all require accurate correspondence and slight flaws can yield perceptually significant errors. To obtain good results, authors of these methods have relied on either extreme redundancy of measurement, human-assisted tracking, substantial smoothing, or domain-specific feature-appearance models.

In this paper, we describe a new method that can solve the correspondence tasks illustrated in Figures 1 and 2 using purely local image analysis, without prior training, and without smoothing or pooling of motion estimates. Our approach defines an image transform; this transform characterizes the local structure of an image in a manner insensitive to points in an occluded region (e.g., outliers), but which *is* sensitive to the shape of the occlusion boundary itself. In essence, our method is to perform matching on a redundant, local representation of image homogeneity. In this paper we show examples where color is the attribute analyzed for homogeneity, but our method is applicable to other local image characteristics (such as texture, range data, or simply image intensity). While we only show sparse tracking results, our method can readily yield dense correspondences, assuming sufficient image contrast.

2 A robust image transform

Since contrast determines the ability to find unique correspondences, we motivate our approach by considering the sources of contrast within a local image window that contains an occlusion boundary. We define the "foreground" to be the scene layer on which the central point of the window resides; points on all other layers are considered "background". We desire a transform which ignores background contrast but is sensitive to contrast energy from the occluding boundaries of the foreground layer.

In general one does not know *a priori* whether con-

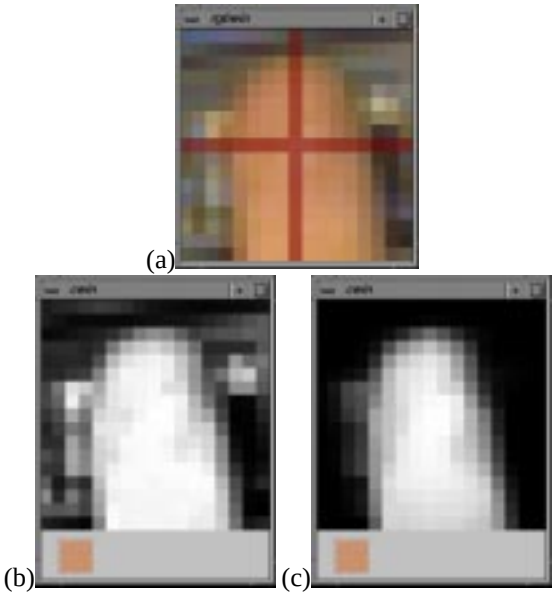


Figure 3. Construction of the Radial Cumulative Similarity (RCS) transform. (a) Color window, (b) central color C (in box at lower-left) and map of local similarity S. Bright pixels indicate similar value as central color. (c) neighborhood of cumulative similarity, N, where each pixel reflects the likelihood the ray from the center point has uniform color.

trast within a particular window is entirely within the foreground layer, is due to the occlusion boundary between foreground and background, or is entirely within the background layer. When contrast is in the foreground layer, an ideal template would model it fully, both in magnitude and sign. When the contrast is due to an occlusion edge, it is reasonable only to define a template based on the contrast energy, since the sign of contrast is arbitrary with changing background. When contrast is in the background layer, it should be ignored in an ideal template.

We define a robust local image representation that approximates this ideal, without any prior knowledge of the occlusion location. Our representation is comprised of a central image-attribute value (typically color) and of a local contrast neighborhood of this attribute, attenuated to discount background influence. Many different diffusion functions could be used to attenuate background influence; in this paper we explore radial cumulative probability functions. The local neighborhood is defined by estimating the contrast energy of the attribute relative to the center value, interpreting this energy probabilistically, and computing the cumulative likelihood that the attribute is unchanged along the ray from the template center to a particular neighborhood point.

Formally, given a discrete color image intensity function $\mathbf{I}(x, y)$ we compute a local robust representation:

$$\mathcal{R}_{\mathbf{I},x,y} \;=\; \{\mathbf{C}_{\mathbf{I},x,y}, N_{\mathbf{I},x,y}(i,j)\}$$

where $-M_n \leq i, j \leq M_n$. Our representation is comprised of two terms, a central value and a neighborhood function; the central value is simply the image attribute averaged over the center point or a small central area:

$$\mathbf{C}_{\mathbf{I},x,y} \;=\; \frac{1}{(2M_c+1)^2} \sum_{i,j=-M_c}^{i,j \leq M_c} \mathbf{A}(\mathbf{I}, x+i, j-i).$$

where $\mathbf{A}(\mathbf{I}, x, y)$ is an image attribute function and can be defined to be any local image property. In this paper we explore attribute functions which return the color or hue vector corresponding to the pixel at the given location. We typically keep the central region small, with $M_c = 0$ or 1. The neighborhood is defined over window coordinates $-M_n \leq i, j \leq M_n$ using the similarity of other image attribute values to the central value:

$$S_{\mathbf{I},x,y}(i,j) \;=\; e^{-\mathbf{E}_{\mathbf{I},x,y}(i,j)^T \mathbf{E}_{\mathbf{I},x,y}(i,j)}.$$

$$\mathbf{E}_{\mathbf{I},x,y}(i,j) \;=\; (\mathbf{C}_{\mathbf{I},x,y} - \mathbf{A}(\mathbf{I}, x+i, y+j))$$

Note that $-\log S$ is a local contrast energy function, and is thus independent of contrast sign.

When tracking a single feature of known size, we could simply use $S_{\mathbf{I},x,y}(i,j)$ over a fixed (possibly non-rectangular) window cropped to resolve the entire feature and the occlusion boundary. This would yield a template which captures both the foreground and occlusion contrast, and was insensitive to contrast sign. However, when automatically tracking features for image analysis/synthesis, or when computing dense correspondence for stereo or motion, we rarely have the luxury of knowledge of appropriate window size.

For fully automatic processing, we define a function which substantially attenuates the influence of exterior pixels. We define our neighborhood function by propagating the attribute similarity function S outward along a ray from the center of the window, so that once we encounter a dissimilarity (i.e., contrast energy) we attenuate the influence of any contrast found farther out along that ray. We are essentially making the assumption that the most proximate contrast is due either to surface contrast or occlusion contrast; background contrast must lie beyond an occurrence of occlusion contrast. Our algorithm

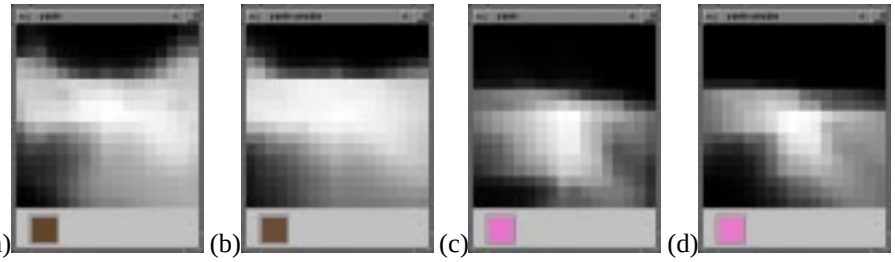


Figure 4. The RCS transform is stable despite occlusion boundaries of different contrast sign. (a,b) show the RCS transform of the marked locations in Figure 2(b,f), while (c,d) show the RCS transform of Figure 2(d,h).

reflects the conservative assumption that, in the absence of any prior knowledge of occlusion location, correspondence judgments are best made on the most proximate contrast.

Our neighborhood function is the cumulative product of S, computed radially from the center point:

$$N_{\mathbf{I},x,y}(i,j) \;=\; \prod_{(k,l)\in r_{i,j}} S_{\mathbf{I},x,y}(k,l)$$

where $r_{i,j}$ is the set of points that lie along the ray from $(0,0)$ to (i,j), inclusive. Other possible neighborhood functions include pixel-fill or diffusion operators; these would also capture non-convex local similarity structure.

We call the representation $\mathcal{R}$ the *Radial Cumulative Similarity* (RCS) transform, since it reflects the radial homogeneity of a given attribute value. Figure 3 illustrates the computation of color RCS for a image window containing a fingertip. The substantial benefit of the RCS transform is invariance to sign of contrast at an occluding boundary, as well as invariance to background contrast. As an example Figure 4 shows the RCS transform for the marked locations in Figure 2; despite dissimilar background structure and occlusion contrast sign reversal, the transformed pairs are substantially similar.

3 Finding correspondences

We define a distance metric using the RCS transform as the weighted L_2 error in central attribute and neighborhood function value:

$$D_\lambda(\mathcal{R}_{\mathbf{I},x,y},\mathcal{R}_{\mathbf{I}',x',y'}) \;=\; (1-\lambda)\Delta N \;+\; \lambda\Delta C$$

where the neighborhood difference is

$$\Delta N \;=\; \frac{1}{(2M_n+1)^2}\sum_{i,j}(N_{\mathbf{I},x,y}(i,j)-N_{\mathbf{I}',x',y'}(i,j))^2.$$

The central attribute difference is similarly,

$$\Delta C \;=\; \frac{1}{a}((\mathbf{C}_{\mathbf{I},x,y}-\mathbf{C}_{\mathbf{I}',x',y'}))^T(\mathbf{C}_{\mathbf{I},x,y}-\mathbf{C}_{\mathbf{I}',x',y'}).$$

where a is the dimension of $\mathbf{A}$ (and $\mathbf{C}$).

The bias term λ expresses a trade-off between the contribution of the central attribute error and the neighborhood function error. Generally the neighborhood error is the most important, since it captures the spatial structure at the given point. However, in certain cases of spatial ambiguity the central attribute value is critical for making the correct match unambiguous. For example in the image shown in Figure 2(c), the neighborhood component of the RCS transform would be roughly equal for the marked point and a point located just below the top lip (centered in the dark region of the open mouth). A modest value of λ disambiguates this case.

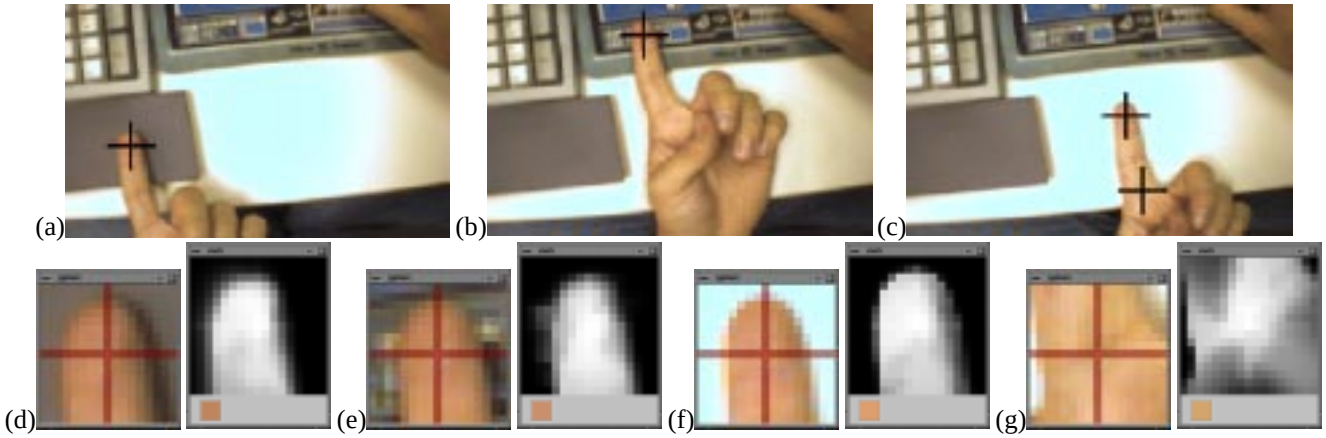
To perform a correspondence search given a point (x,y) in an image $\mathbf{I}$, we compute the RCS transform $\mathcal{R}_* = \mathcal{R}_{\mathbf{I},x,y}$ and search for the point $(\hat{x},\hat{y})$ in a second image $\mathbf{I}'$ such that

$$(\hat{x},\hat{y}) = \arg\min_{x',y'\in W_{x,y}} D_\lambda(\mathcal{R}_*,\mathcal{R}_{\mathbf{I}',x',y'})$$

where $W_{x,y}$ is a search window of radius M_w centered at (x,y).

4 Results

In the present implementation we recompute all $\mathcal{R}_{\mathbf{I}',x',y'}$ each time D is evaluated. We have not optimized for speed; using $M_n = 8$, $M_w = 50$, $M_c = 0$ a substantial fraction of a second is consumed to find a correspondence minima per feature. However since RCS is a transform, we could easily precompute $\mathcal{R}$ over the entire image and then be faced with the run-time cost of a standard least-squares template search with template radius M_n and search radius M_w.



(h)

	A:B	A:C	B:C	A:D	B:D	C:D
L_2	0.05	0.35	0.33	0.27	0.22	0.15
ROBUST	0.04	0.62	0.65	0.38	0.37	0.20
RCS	0.05	0.11	0.07	0.35	0.27	0.30
(ideal)	0	0	0	1	1	1

Figure 5. Fingertip feature locations (a) feature A, (b) feature B, (c) features C (top) and D (bottom). Feature D is a distractor. (d-g) Raw color values and RCS transform for features A-D. (h) Correspondence values between features for three different metrics: L_2 on intensity, robust norm (Lorentzian $\sigma = 0.4$) on intensity, and L_2 on the RCS transform. The ideal correspondence values would be near 0 for pairs of fingertips, and near 1 for pairs with the distractor. Values are not normalized and so can only be compared within-method.

We compared our method to correspondence search using classic L_2 norms, using normalized correlation, using a robust redescending norm (from [1], a Lorentzian ρ with $\sigma = 0.1$), and using our RCS transform with $\lambda = 0.1$. The L_2 norm and normalized correlation yielded substantially similar results, and so for brevity we only show L_2 results here.

First we note that in the majority of image locations, all three methods yield accurate results. It is only at points near discontinuities, and further at points where the discontinuity changes contrast sign between images, that there is a dramatic difference between RCS and the comparison methods. We will thus demonstrate performance in a disproportionate number of these cases (these are often critical locations for image analysis/synthesis tasks).

Figure 5 shows a comparison of correspondence values for a fingertip at various background locations (A,B,C), and a distractor region (D) of the hand. The table in Figure 5(h) shows that only the RCS method has correct performance: low distance measures for all the cases of correspondence between actual fingertips (A:B, A:C, B:C) and high distance for cases with the distractor (A:D, B:D, C:D).

Figures 6 and 7 show results from tracking 16 features simultaneously on image pairs of an eye, mouth, and fingers, and from comparing to hand-labeled ground truth. The mean coordinate error across the three images was 5.6 pixels for the L_2 norm, 5.2 pixels for the redescending robust norm, and 0.97 pixels for the RCS method. The images were processed at 320x240 resolution. As expected, the L_2 norm had difficulty at regions where substantial occlusion was present, and the redescending robust norm had problems where the designated correspondence was at a region of occlusion contrast sign reversal. At points where no occlusion was present the L_2 and redescending norm had no coordinate error, but the RCS did return erroneous correspondences in approximately 5% of points.

This lower performance of RCS away from occlusion boundaries is not surprising: When analyzing an image window of a single surface where brightness constancy holds (e.g., there is no occlusion) suboptimal performance results from downweighting portions of the window that are actually foreground. Informally, regions of high contrast that are prone to aliasing in the RCS representation

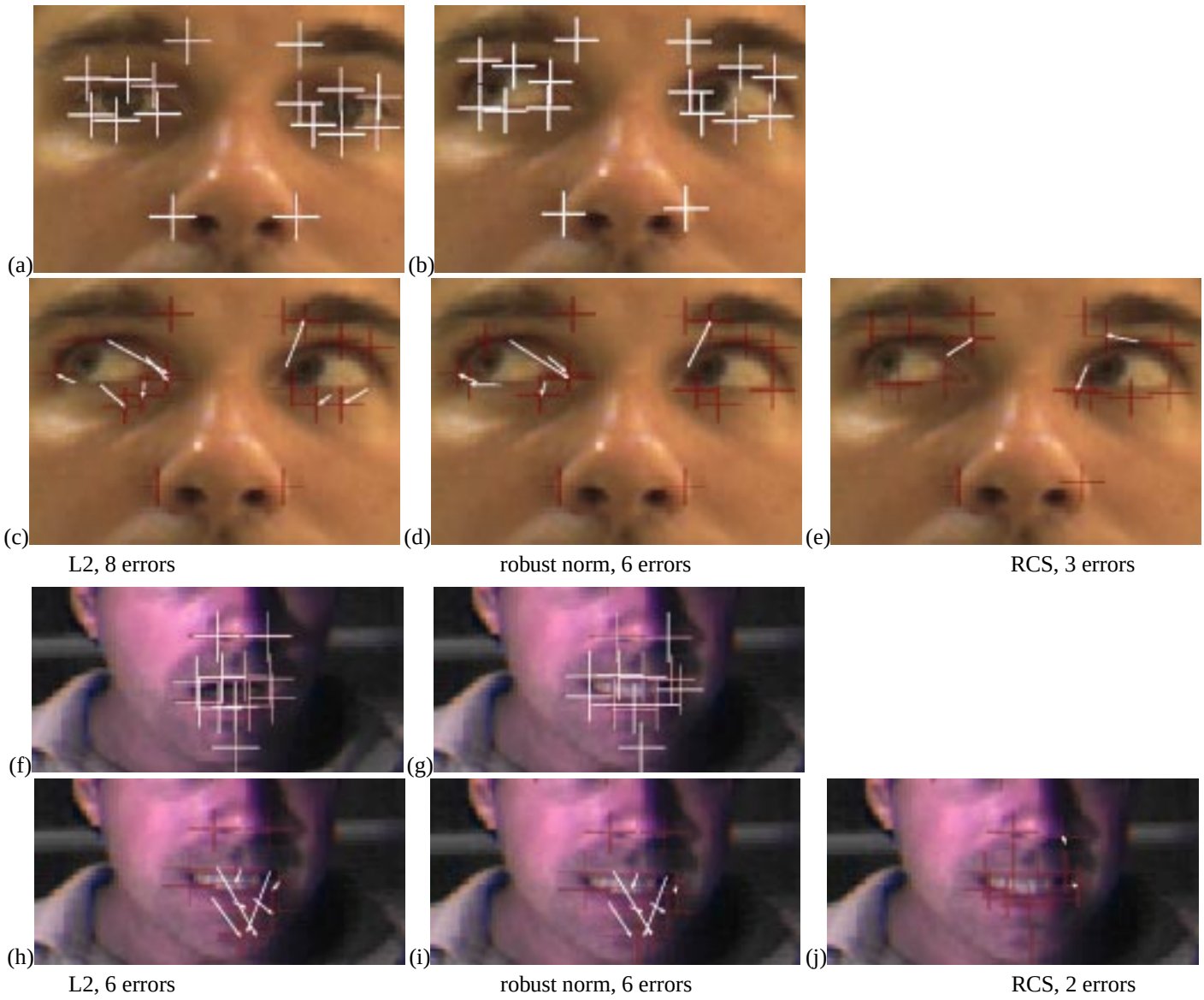


Figure 6. Results of exhaustive correspondence search for 16 different features in various image pairs. (a,b) hand-labeled feature locations an image pair with moving eyeballs, (f,g), an image pair with changing mouth expression. For each feature in the first image (a,f), we searched for the point in the second image (b,g) with minimum correspondence error using three different distance metrics: L_2, robust norm, and RCS. (c,h) Results using L_2 norm on intensity, showing arrows where incorrect correspondences were returned. There were 8 and 6 correspondence errors, with mean squared coordinate error of 6.1 and 3.4 pixels, respectively. (d,i) Results using robust norm on intensity: 6 and 6 correspondence errors, mean squared coordinate error of 5.0 and 3.3 pixels. (e,j) Results using L_2 norm on RCS transform: 3 and 2 correspondence errors, mean squared coordinate error of 2.3 and 0.4 pixels.

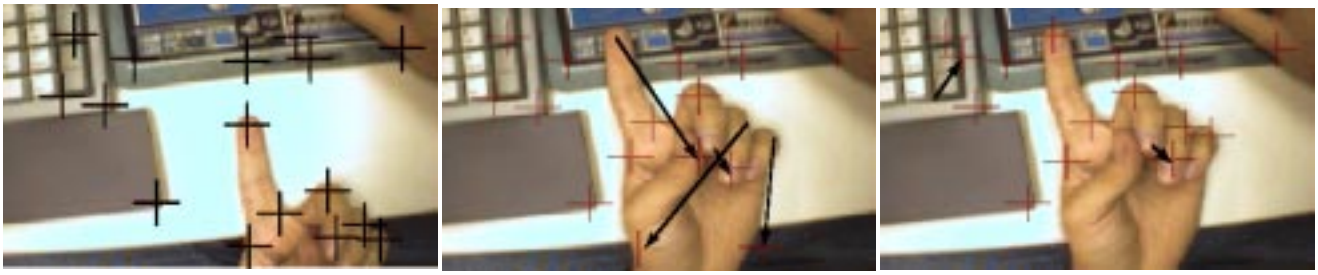


Figure 7. Results as in previous figure on an image pair with a hand moving over different backgrounds. (a) shows first image of pair, (b) results on second image using robust norm (L_2 norm yielded same result): 4 errors, mse=7.3 pixels, (c) RCS result: 2 errors, mse=0.21 pixels.

can be detected by computing the sum of the radial cumulative similarity function, N: if that sum is below a certain threshold the RCS transform should be considered degenerate. Fortunately, occlusion-free regions of high contrast are cases where the traditional methods perform exceedingly well. We are currently implementing a hybrid algorithm which reverts to a L_2 when that method yields good results. Alternatively a smoothing or regularization stage would also greatly alleviate this problem.

5 Conclusion

Radial Cumulative Similarity (RCS) is a new image transform that describes local image homogeneity, comprised of a central attribute value and a function of the surrounding radial similarity structure. We compute radial-similarity as the cumulative product of the probability the attribute value is constant along a given ray from the center. When applied to color attributes, this representation is insensitive to structure outside an occluding boundary, yet it can model the boundary itself. The RCS can therefore be used to track foreground surfaces near occlusion where there is no foreground contrast other than the from the occlusion boundary.

References

[1] M. Black and P. Anandan, A framework for robust estimation of optical flow, 4th Proc. ICCV, 1993.

[2] M. Shizawa and K. Mase, Simultaneous multiple optical flow estimation, Proc. CVPR, 1990.

[3] M. Irani, B. Rousso, and S. Peleg, Computing Occluding and Transparent Motions, IJCV, 12(1), 1994.

[4] D. Bhar and S. Nayar, Ordinal measures for visual correspondence, Proc. CVPR, 1994.

[5] R. Zabih and J. Woodfill, Non-parametric local transforms for computing visual correspondence, Proc 3rd ECCV, 1994.

[6] T. Darrell, and A. Pentland, Robust Estimation of a Multi-Layer Motion Representation, Proc. IEEE Workshop on Visual Motion, Princeton, NJ, 1991.

[7] J. Wang, and E. H. Adelson, Layered Representations for Image Sequence Coding, Proc. CVPR, 1993.

[8] S. Ayer and H. Sawhney, Layered representation of motion video using robust maximum likelihood estimation of mixture models and MDL encoding, Proc. ICCV, 1995.

[9] E. P. Simoncelli and E. H. Adelson, Probability distributions of optical flow, Proc. CVPR, 1991.

[10] T. Darrell, and E. Simoncelli, Nulling Filters and the Separation of Transparent Motions, Proc. CVPR, 1993.

[11] J. R. Bergen, P. J. Burt, K. Hanna, R. Hingorani, P. Jeanne, and S. Peleg, Dynamic multiple-motion computation. In Y. A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*, Elsevier Science Publishers B.V., 1991.

[12] M. Black and Y. Yacoob, Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions, Proc. ICCV, 1995.

[13] G. Hager and P. Belhumeur, Real-time tracking of image regions with changes in geometry and illumination, Proc. CVPR, 1996.

[14] M. Black, Y. Yacoob, A. Jepson, D. Fleet, Learning Parameterized Models of Image Motion, Proc. CVPR, 1997.

[15] Kanade, T., Yoshida, A., Oda, K., Kano, H., and Tanaka, M., A Video-Rate Stereo Machine and Its New Applications, Proc. CVPR, 1996.

[16] M. Covell, and M. Withgott, Automatic Morphing: spanning the gap between motion estimation and morphing, Proc ICASSP, 1994.

[17] C. Bregler, M. Covell, M. Slaney, Video Rewrite, Proc. SIGGRAPH, 1997.

[18] M. Covell and C. Bregler, Eigenpoints, Proc. ICIP, 1996.

[19] M. Black, Y. Yacoob, D. Fleet, Modeling Appearance Change in Image Sequences, Proc. 3rd Intl. Workshop on Visual Form, Capri, Italy, 1997.