Video Object Tracking Based On Extended Active Shape Models With Color Information

A. Koschan, S.K. Kang, J.K. Paik, B. Abidi, and M. Abidi
Imaging, Robotics, and Intelligent Systems Laboratory, University of Tennessee
Knoxville, Tennessee

Abstract

Tracking and recognizing non-rigid objects in video image sequences are complex tasks of increa sing importance to many applications. For the tracking of such objects in a video sequence e.g. "active shape models" can be applied. The existing active shape models are usually based on intensity information and they do not consider color information. However, active shape models are sensitive to outliers, especially in the case of partial object occlusions. In this paper, we present an extension of the active shape model for color images and we examine to what extent the use of color information can contribute to the solution of the outlier problem.

Introduction

The problem of tracking people and recognizing their actions in video sequences is of increasing importance to many applications.[1,2,3] Examples include video surveillance, human computer interaction, and motion capture for animation, to name a few. Special considerations for digital image processing are required when tracking objects whose forms change between two frames. For example, pedestrians in a road scene belong to this class of objects denoted as *non-rigid objects*. For the tracking of non-rigid objects in a video sequence, active shape models (ASMs) could be applied. The existing active shape models usually do not consider color information. In this paper, we present several extensions of the active shape model for color images using different color adapted objective functions.

Tracking and recognizing non-rigid objects in video image sequences are complex tasks. Using color information as a feature to describe a moving object or person can support these tasks. The use of four-dimensional templates for tracking objects in color image sequences was suggested in Ref. 4. However, if the observation is accomplished over a long period of time and with many single objects, then both the memory requirements for the templates in the database and the time requirements for the search of a template in the database increase. In contrast to this, ASMs represent a compact model for which the form variety and the color distribution of an object class are taught in a training phase.[5]

Several systems use skin color information for tracking faces and hands.[6,7,8] The basic idea is to limit the search complexity to one single color cluster (representing skin color) and to identify pixels based on their membership to this cluster. Several problems affect these approaches. First, skin colors cannot be uniquely defined and, in addition, a person cannot be identified when seen from behind. Here tracking clothes instead of skin is more appropriate.[9]

Second, color distributions are sensitive to shadows, occlusions, and changing illuminations. Addressing the problem occurring with shadows and occlusions, Lu and Tan[10] assume that the only moving objects in the scene are persons. This assumption does not hold for many applications. Most of the approaches mentioned above cannot be easily extended to multi-colored objects other than persons. In this paper, we present a general technique to track colored non-rigid objects (including persons).

A very efficient technique for the recognition of colored objects is color indexing[11] Based on comparisons between color distributions, an object in the image is assigned to an object stored in a database. This technique usually needs several views of the object to be recognized, which is not always ensured when tracking people in a road scene, for example. Furthermore, color indexing partly fails with partial occlusions of the object. Active shape models do not need several views of an object, since by using energy functions they can be adapted to the silhouette of an object represented in the image. However, the outlier problem, which can occur particularly with partial object occlusion, represents a difficulty for these models. In the following, an extension of the active shape models for color images is presented. We examine to what extent the use of color information can contribute to the solution of the outlier problem, especially in the case of occlusions.

Active Shape Models

For tracking a human target in video, detecting the shape and position of the target is the fundamental task. Since the shape of a human object is subject to deformation and random motion in the two-dimensional image space, ASM is one of the best-suited approaches in the sense of both accuracy and efficiency.

ASM falls into the category of deformable shape models with a priori information about the object. ASM-based object tracking models the contour of the silhouette of an object, and the set of model parameters is used to align different contours in each image frame.

More specifically, an ASM-based tracking algorithm consists of the following steps: (i) landmark points assignment, (ii) principal component analysis (PCA), (iii) model fitting, and (iv) local structure modeling.

Landmark Points

Given a frame of input video, suitable landmark points should be assigned on the contour of the object. Figure 1 shows manually selected, 42 landmark points on the contour of the human object. Good landmark points should be consistently located from one image to another. In a two-dimensional image, we represent n landmark points by the $2n$ vector as

$$\mathbf{x} = [x_1, \ldots, x_n, y_1, \ldots, y_n]^T . \qquad (1)$$

Various automatic, systematic ways of obtaining landmark points were discussed in Ref. 12.

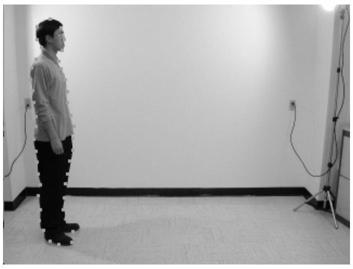

Figure 1. A human object with 42 landmark points (n=42).

Principal Component Analysis

A set of n landmark points represents the shape of the object. Figure 2 shows a set of 56 different shapes, called a training set.

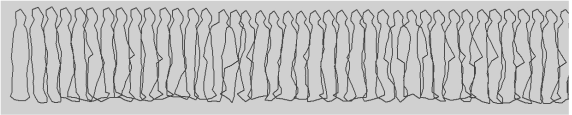

Figure 2. Training set of 56 shapes (m=56).

Although each shape in the training set is in the $2n$-dimensional space, we can model the shape with a reduced number of parameters using the principal component analysis (PCA) technique.

Suppose we have m shapes in the training set, such as $\mathbf{x}_i$, $i = 1,\ldots,m$. The PCA algorithm is as follows.

PCA Algorithm

1. Compute the mean of the m sample shapes in the training set.

$$\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i . \qquad (2)$$

2. Compute the covariance matrix of the training set.

$$\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T . \qquad (3)$$

3. Construct the matrix

$$\mathbf{\Phi} = [\phi_1 \mid \phi_2 \mid \cdots \mid \phi_t] , \qquad (4)$$

where ϕ_i,i=1,…,t represent eigenvectors of $\mathbf{S}$ corresponding to t largest eigenvectors.

4. Given $\mathbf{\Phi}$ and $\bar{\mathbf{x}}$, each shape can be approximated as

$$\mathbf{x}_i \approx \bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}_i , \qquad (5)$$

where

$$\mathbf{b}_i = \mathbf{\Phi}^T(\mathbf{x}_i - \bar{\mathbf{x}}) . \qquad (6)$$

In step 3 of the PCA algorithm, t is determined so that the sum of t largest eigenvalues is greater than 98% of the sum of all eigenvalues.

In order to generate plausible shapes, we need to evaluate the distribution of $\mathbf{b}$. To constrain $\mathbf{b}$ to plausible values we can either apply hard limits to each element b_i or constrain $\mathbf{b}$ to be in a hyper-ellipsoid. The nonlinear version of this constraint is discussed in Ref. 13.

Model Fitting

We can find the best pose and shape parameters to match a shape in the model coordinate frame, $\mathbf{x}$, to a new shape in the image coordinate frame, $\mathbf{y}$, by minimizing the following error function

$$E = (\mathbf{y} - \mathbf{M}\mathbf{x})^T \mathbf{W}^T (\mathbf{y} - \mathbf{M}\mathbf{x}) , \qquad (7)$$

where $\mathbf{M}$ represents the geometric transformation of rotation θ, translation $\mathbf{t}$, and scale s. For instance, if we apply the transformation to a single point, denoted by $[x,y]^T$, we have

$$\mathbf{M}\begin{bmatrix} x \\ y \end{bmatrix} = s\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} . \qquad (8)$$

After the set of pose parameters, {θ,$\mathbf{t}$,s} are obtained, the projection of $\mathbf{y}$ into the model coordinate frame is given as

$$\mathbf{x}_p = \mathbf{M}^{-1}\mathbf{y} . \qquad (9)$$

Finally, the model parameters are updated as

$$\mathbf{b} = \mathbf{\Phi}^T(\mathbf{x}_p - \bar{\mathbf{x}}) . \qquad (10)$$

Modeling a Local Structure

A statistical, deformable shape model can be built by landmark point's assignment, PCA, and model fitting steps. In order to interpret a given shape in the input image based on the shape model, we must find the set of parameters that best match the model to the image.

If we assume that the shape model represents boundaries and strong edges of the object, a profile across each landmark point has edge-like local structure.

Let $\mathbf{g}_i$, i=1,…,n, be the normalized local profile across the I-th landmark point, and $\overline{\mathbf{g}}$ and $\mathbf{S}_g$ the corresponding mean and covariance, respectively. The nearest profile can be obtained by minimizing the following Mahalanobis distance between the sample and the mean of the model as

$$f(\mathbf{g}_s) = (\mathbf{g}_s - \overline{\mathbf{g}})^T \mathbf{S}_g{}^T (\mathbf{g}_s - \overline{\mathbf{g}}) . \qquad (11)$$

In practice, we used a multi-resolution ASM technique because it provides a wider range for the nearest profile search.

Extending ASMs to Color Image Sequences

In gray scale image processing, the objective functions are determined along the normals for a representative point in the gray value distribution. This procedure can be extended to color images by first computing objective functions separately for each component of the color vectors. Afterwards, a "common" minimum has to be determined by analyzing the resulting minima that are computed for each single color component. One way of doing this consists of selecting the smallest minimum in the three color components as a candidate. If, however, one of the three color channels contains an outlier (compare Figure 3), this outlier might be selected as a minimum.

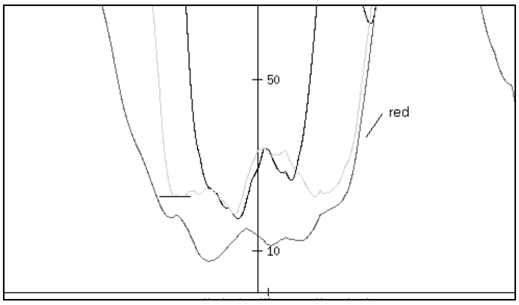

Figure 3. Example of objective functions for three color components with an outlier in the red component.

Another procedure consists of selecting the average of the absolute minima in all three color components. However, outliers in one color channel also lead in this case to a wrong result. Furthermore, the average value may represent a value that corresponds with none of the regarded energy functions. One way to overcome this problem is to use the median of the absolute minima in the three color channels as a candidate. Thereby the influence of outliers in the minima of the objective functions is minimized. However, further false values may arise during the alignment of the contours.

In the next section we will further address the question if a contrast-adaptive optimization may improve the ASM performance. For every single landmark point we will select the color channel with the highest contrast and minimize the corresponding objective function.

Experiments and Results

Two frames of an indoor color image sequence were used to determine the best searching method. The test images are shown in Figure 4. For this experiment, 57 shapes were used as the training set for PCA, and a 7 pixel-wide profile was used for each landmark point in three RGB color channels. After the modeling step, we got three profile models for each color channel and a shape model.

The purpose of the first experiment was to evaluate the performance of different combinations of color models. The used terminologies are summarized in Table 1.

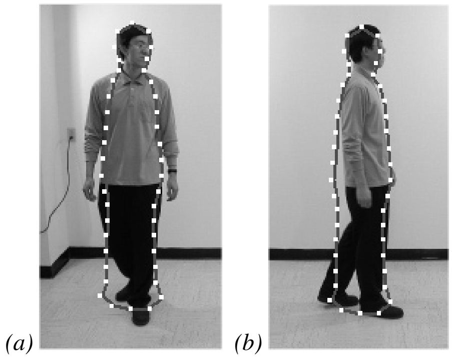

Figure 4. Test images with initial points for (a) the 57th image and (b) the 7th image.

Table 1. Terminologies

Intensity	The result using the intensity image with the intensity profile.
R	The result using the color image with the Red profile.
G	The result with the Green profile.
B	The result with the Blue profile.
Minimum	The result using the color image after selecting the minimum of the minima of the Mahalanobis distance in the three color channels.
Median	The result with the median of the minima.
Mean	The result with the mean of the minima.
Adaptive	The result using the intensity image with the adaptive profile model that is modeled with the strongest edge among three color channels for each point.

The initial landmark points were manually placed as shown in Figure 4. Hill, Taylor, and Cootes[5] suggested a genetic algorithm that determines the "best" form parameters from a randomly specified set of initial values. So far we did not examine this algorithm due to its computational complexity. We argue that a manual definition of the form parameters is suitable for our purpose since the initial form has only to be determined once for a class of similar-shaped objects. Our goal is to track persons and to ignore other moving objects.

Furthermore, we defined a maximum shift between two image frames for an object to be tracked. This limitation is due to a reduction of the computing time and does not restrict the algorithm in general. The maximum shift parameter depends on the size of the object, the distance between the camera and the object, the velocity of the object, and the moving direction of the object. For example, for tracking a person on an airport we can predict the maximum size of a person, the maximum velocity of a walking or running person, and the minimum distance between the camera and a person. To limit the moving direction of a person, we can further assume that only a few persons might move towards a camera that is mounted on a wall. In our investigation we limited the maximum shift to 15 pixels for the hierarchical approach.

Both hierarchical and non-hierarchical methods were tested for the image shown in Figure 4(a) because its initial contour was set smaller than the real object. On the other hand, only the non-hierarchical method was tested in Figure 4 (b). In the hierarchical approach, level 0 represents the original given resolution, level 1 the half-sized resolution, and level 2 the quarter-sized resolution. Three different levels are shown in Figure 5. We performed 5 iterations in level 2, another 5 iterations in level 1, and finally 10 iterations in level 0. For the non-hierarchical approach we performed 10 iterations. The hierarchical approach helps to enlarge the search regions and shows a better search result than the non-hierarchical approach. The model fitting error for each experiment is summarized in Table 2.

The result of the hierarchical approach to Figure 4(a) is shown in Figure 6. The result of the non-hierarchical approach is shown in Figure 7. The median method gives the best results in the sense of both visual and the objective error measurements. Results using the R, G, and B color channels show worse fitting than those method using intensity. Table 2 summarizes error measurements of different methods given in Table 1.

Table 2. The sum of distance between the estimated points by the different searching methods and the manually assigned points.

	Intensity	R	G	B
NH (57th)	196.25	241.33	157.51	190.40
HR (57th)	146.66	182.84	159.18	164.37
NH (7th)	161.33	171.09	171.19	225.47
	Minimum	Median	Mean	Adaptive
NH (57th)	192.57	178.81	185.56	185.54
HR (57th)	177.14	122.72	169.45	156.27
NH (7th)	158.94	152.08	170.46	162.96

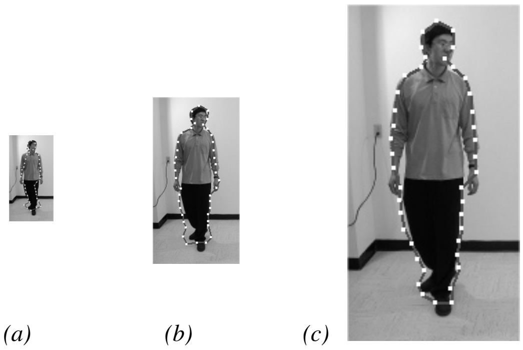
(a) (b) (c)

Figure 5. Three different resolutions used in the hierarchical approach: (a) level 2, (b) level 1, and (c) level 0.

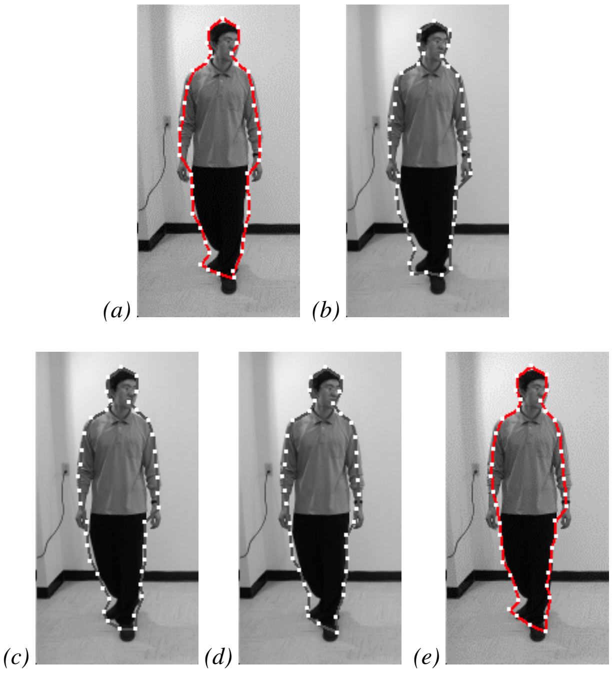
(a) (b) (c) (d) (e)

Figure 6. Hierarchical search results of the 6 different methods for the 57th image: (a) intensity, (b) minimum, (c) median, (d) mean, and (e) adaptive.

The second experiment used an outdoor sequence. We applied the ASM to each of the outdoor image frames and selected the mean, the minimum, and the median of the minima in the objective functions for searching. The results for selecting the median of the minima are shown in Figure 8. The ASM gives good results, even though the object is partially occluded by the bench.

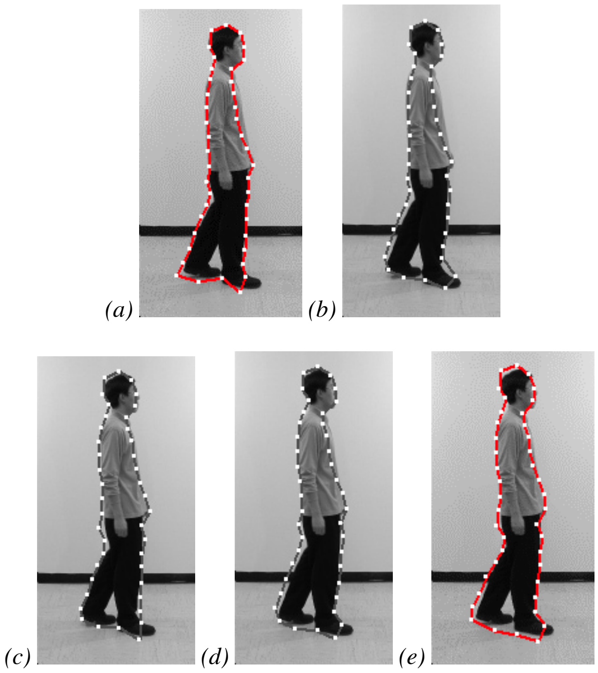

Figure 7. Non-hierarchical search results of six different methods for the 7th image: (a) intensity, (b) minimum, (c) median, (d) mean, and (e) adaptive.

Conclusion

A technique was presented for recognizing and tracking a moving object or person in a video sequence. For this the objective function for active shape models was extended to color images. We evaluated several different approaches for defining an objective function considering the information from the single components of the color image vectors. This tracking technique does not require a static camera (except to initialize the landmark points for the object to be recognized). The median computation of the minima in the energy functions proved favorable in our indoor and outdoor experiments.

In general the error in fitting an ASM to the real contour of an object was lower when using color information than when just using intensity information. Furthermore, we showed that the fitting error can be further reduced when applying a hierarchical approach instead of a non-hierarchical one to the images. The performance of the algorithm was rather robust regarding partial object occlusions. The problem of outliers in the objective functions could be partly solved by the evaluation of color information. One way to further enhance these results might be a refined analysis of the objective functions, where the neighbors of one point are also considered. Thereby the number of outliers can be further reduced.

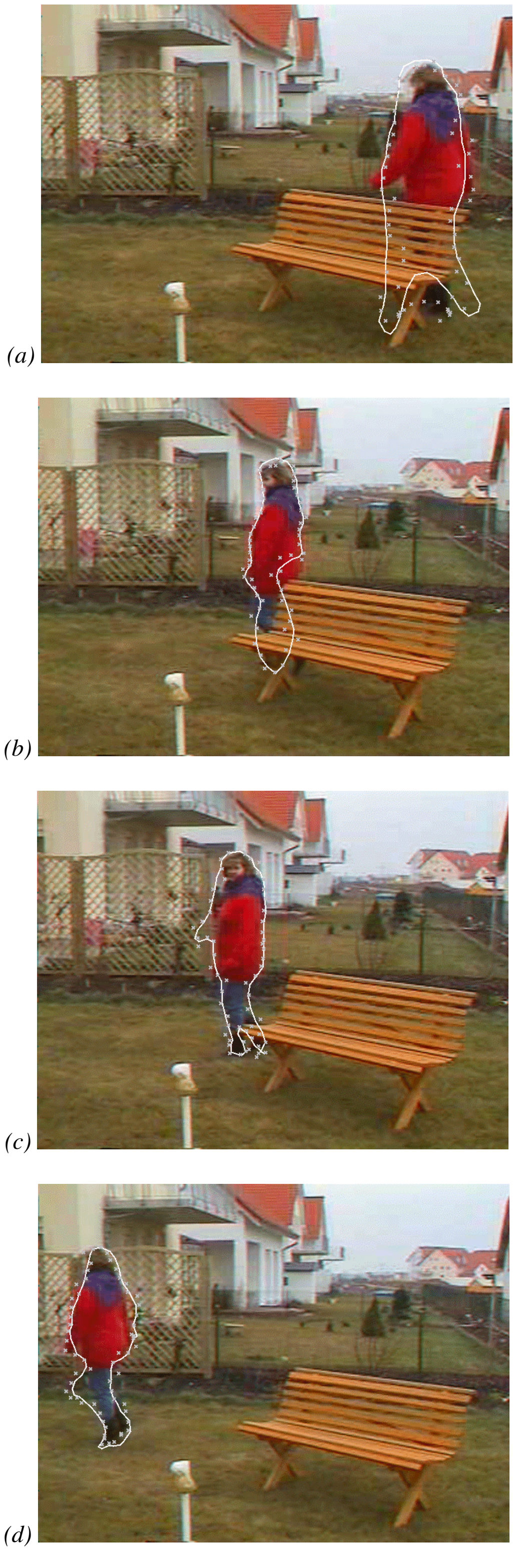

Figure 8. Search results for an outdoor sequence using the non-hierarchical approach for (a) the 1st frame, (b) the19th frame, (c) the 27th frame, and (d) the 33rd frame.

However, the tracking of a person becomes rather difficult if the image sequence contains several moving persons with similar shape. In this case, a technique exclusively based on the contour of a person will have difficulties in tracking a selected person and the task may fail if the person is partially occluded. On the other hand, a technique exclusively evaluating the colors of a moving person (or object) may also fail. Any color-based tracker can lose the object it is tracking due, for example, to occlusion or changing lightning conditions. To overcome the sensitivity of a color-based tracker to changing lightning conditions, the color constancy problem has to be solved at least in parts. This is a non-trivial and computationally costly problem that can in general not be solved in video real-time.

Another solution to the problem mentioned above could consist of a weighted combination of a form-based tracking technique using, for example, ASMs and a color-based tracking technique using, for example, color indexing. By applying such a combination technique to image sequences we might be able to distinguish between a) objects of similar colors but with different forms and b) objects of different colors but with similar forms.

Acknowledgements

This work was supported by the University Research Program in Robotics under grant DOE-DE-FG02-86NE37968, by the DOD/TACOM/NAC/ARC Program, R01-1344-18, and by FAA/NSSA Program, R01-1344-48/49. Furthermore, the authors acknowledge the help of Klaus Curio of TU Berlin, Germany.

References

1. R. Plänkers and P. Fua, Tracking and modeling people in video sequences, Comp. Vision and Image Understanding 81, pg. 285-302 (2001).
2. S. J. McKenna, Y. Raja, and S. Gong, Tracking colour objects using adaptive mixture models, Image and Vision Computing 17, pg. 225-231 (1999).
3. I. Haritaoglu, D. Hartwood, and L. S. Davis, W4: Real-time surveillance of people and their activities, IEEE Trans. on PAMI 22, pg. 809-830 (2000).
4. S. A. Brock-Gunn, G. R. Dowling, and T. J. Ellis, Tracking using colour information, Proc. ICARCV ´94, pg. 686-690 (1994).
5. T. F. Cootes, D. H. Cooper, C. J. Taylor, and J. Graham, Active Shape Models - Their training and application, Comp. Vision and Image Understanding 61, pg. 38-59 (1995).
6. Y. Li, A. Goshtasby, and O. Garcia, Detecting and tracking human faces in videos, Proc. ICPR'00 vol. 1, pg. 807-810 (2000).
7. F. Marqués and V. Vilaplana, Face segmentation and tracking based on connected operators and partition projection, Pattern Recognition 35, pg. 601-614 (2002).
8. D. Comaniciu and V. Ramesh, Robust detection and tracking of human faces with an active camera, Proc. Visual Surveillance 2000, pg. 11-18 (2000).
9. H. Roh, S. Kang, and S.-W. Lee, Multiple people tracking using an appearance model based on temporal color, Proc. ICPR'00 vol. 4, pg. 643 -646 (2000).
10. W. Lu and Y.-P. Tan, A color histogram based people tracking system, Proc. ISCAS 2001 vol. 2, pg. 137 -140 (2001).
11. M. J. Swain and D. H. Ballard, Color indexing, Int. Journ. of Comp. Vision 7, pg. 11-32 (1991).
12. Q. Tian, N. Sebe, E. Loupias, and T. S. Huang, Image retrieval using wavelet-based salient points, Journ. of Electronic Imaging, 10 (4), pg. 935-849 (2001).
13. P. Sozou, T. F. Cootes, C. J. Taylor, and E. D. Mauro, A nonlinear generalization of point distribution models using polynomial regression, Image and Vision Computing 12 (5), pg. 451-457 (1995).
14. A. Hill, C. J. Taylor, and T. F. Cootes. A generic system for image interpretation using flexible templates, Proc. ECCV`94, pg. 276-285 (1994).

Biography

Andreas Koschan received his Diplom (M.S.) in Computer Science and his Doktor-Ing. (Ph.D.) in Computer Engineering from Technical University Berlin, Germany in 1985 and 1991, respectively. Currently he is a Research Associate Professor at the University of Tennessee, Knoxville. His work has focused primarily on color image processing and 3D computer vision including stereo vision and laser range finding techniques. He is a coauthor of two textbooks on 3D image processing and he is a member of the IS&T and the IEEE.