# Three-dimensional tracking of multiple skin-colored regions by a moving stereoscopic system

Antonis A. Argyros and Manolis I. A. Lourakis

A system that performs three-dimensional (3D) tracking of multiple skin-colored regions (SCRs) in images acquired by a calibrated, possibly moving stereoscopic rig is described. The system consists of a collection of techniques that permit the modeling and detection of SCRs, the determination of their temporal association in monocular image sequences, the establishment of their correspondence between stereo images, and the extraction of their 3D positions in a world-centered coordinate system. The development of these techniques has been motivated by the need for robust, near-real-time tracking performance. SCRs are detected by use of a Bayesian classifier that is trained with the aid of a novel technique. More specifically, the classifier is bootstrapped with a small set of training data. Then, as new images are being processed, an iterative training procedure is employed to refine the classifier. Furthermore, a technique is proposed to enable the classifier to cope with changes in illumination. Tracking of SCRs in time as well as matching of SCRs in the images of the employed stereo rig is performed through computationally inexpensive and robust techniques. One of the main characteristics of the skin-colored region tracker (SCRT) instrument is its ability to report the 3D positions of SCRs in a world-centered coordinate system by employing a possibly moving stereo rig with independently verging CCD cameras. The system operates on images of dimensions 640 × 480 pixels at a rate of 13 Hz on a conventional Pentium 4 processor at 1.8 GHz. Representative experimental results from the application of the SCRT to image sequences are also provided. 

*OCIS codes:* 150.6910, 330.0330.

## 1. Introduction

Human beings have a remarkable ability to interpret the activities of other human beings visually. Developing machines with similar perceptual and cognitive capabilities constitutes an ambitious research goal. The accomplishment of this goal will have far-reaching implications in a wide spectrum of applications such as human–machine interaction, gesture tracking for surveillance systems, and development of tools for teaching by demonstration. The importance and the difficulty of solving this problem justify the great volume of research effort that has been devoted worldwide to providing a robust solution.

A fundamental building block of any system that is able to interpret activities is one that permits the three-dimensional (3D) tracking of a human operator as he or she performs a certain task. Several sensors and techniques to achieve this goal have been developed.[1] Nevertheless, vision-based methods are considered preferable because they are passive and not invasive, in the sense that they do not require modification of the environment or wearing of any special equipment by the human operator. A fundamental issue in human tracking is related to the modeling of a human operator. The human body is a complex, nonrigid structure with many degrees of freedom. Therefore the type and complexity of the proposed models vary dramatically,[2,3] depending heavily on the requirements of the application domain under consideration. For example, tracking people in an indoors environment in the context of a surveillance application has completely different modeling requirements from tracking the fingers of a hand for vision-based recognition of a sign language. In the research reported here, skin color is the fundamental visual cue employed for detection of the presence of a human being in a scene. Color offers many advantages over geometric models, such as robustness to occlusions, changes in scale and resolution, and geometric transformations. Additionally, the computational requirements of color processing

The authors are with the Institute of Computer Science, Foundation for Research and Technology—Hellas, Heraklion, Crete, Greece. A. Argyros's e-mail address is argyros@ics.forth.gr.

Received 15 May 2003; revised manuscript received 19 September 2003.

0003-6935/04/020366-13$15.00/0

are considerably lower than those associated with the processing of complex geometric models. For these reasons, color-based models have been applied to a broad spectrum of applications, such as content-based image retrieval and quality control.

### A. Previous Work

Vision-based methods for tracking skin-colored regions in three dimensions need to provide answers to several questions, each of which constitutes an open research problem: How is skin color modeled and how are instances of the model used detected in an image? How are detected instances associated temporally in sequences of images? How is 3D position information attained from the inherently two-dimensional (2D) observations of the tracked models?

What follows is a description of some representative methods for solving the problems listed above that have been proposed. It is important to note that the available options for solving partial problems should be evaluated with respect to several criteria, such as the quality of their results, their robustness, and their computational complexity.

#### *1. Skin-Color Modeling and Detection*

A major step toward providing a model of skin color is the selection of the color space to be employed. Several color spaces, including RGB,[4] normalized RGB,[5,6] HSV,[7] YCrCb,[8] and YUV,[9] have been proposed. Color spaces that efficiently separate the chrominance from the luminance components of color are typically preferable because, by employing only chrominance-dependent components of color, one may achieve some degree of robustness to changes in illumination. The choice of such color spaces is also justified by the fact that skin tones differ mostly in chrominance and less in intensity. Terrillon *et al*.[10] reviewed skin chrominance models and evaluated their performance.

When a suitable color space has been selected, the simplest methods define skin color by employing bounds in the coordinates of this space.[8] These bounds are typically selected empirically, i.e., by examination of the distribution of skin colors in a preselected set of images. Another approach to defining skin color is to assume that the probabilities of skin colors follow a distribution that can be learned. This learning is achieved through an off-line procedure, although on-line iterative methods have also been suggested.[7]

In contrast to the aforementioned nonparametric approaches, another paradigm is related to methods that make use of parametric models. These methods are based either on a unimodal Gaussian probability-density function[5,9,11] or on multimodal Gaussian mixtures[12–15] that model the probability distribution of skin color. Maximum-likelihood estimation techniques are used to derive the parameters of unimodal Gaussian density functions. Multimodal Gaussian mixtures require an expectation-maximization algorithm to be employed. According to Yang and Ahuja,[16] a mixture of Gaussians is preferable to a single Gaussian distribution. Still, histogram models provide better accuracy and lower computational cost than mixture models for the detection of skin-colored areas in an image.[6]

A few of the proposed methods involve some sort of adaptation to become insensitive to changes in the illumination conditions. For example, adaptation of a Gaussian mixture model that approximates the multimodal distribution of the object's colors based on a recent history of detected skin-colored regions has been suggested.[12,13]

Skin color is an important cue in detecting the presence of humans in a scene. However, it is often insufficient to separate skin objects from nonskin objects that appear to be skin colored. Therefore skin color is often fused with other cues such as motion, texture, shape, and 3D structure information. A recent survey[17] gave an interesting overview of the use of color and other visual cues in skin-color detection.

#### *2. Tracking*

As soon as skin-colored regions have been modeled and can be detected in an image, another major problem must be addressed, which concerns the temporal association of these observations in an image sequence. The traditional approach to solving this problem has been based on the original work of Kalman[18] and its extensions. If the observations and object dynamics are of a Gaussian nature, Kalman filtering suffices to solve the tracking problem. However, in many cases the distributions are non-Gaussian, and thus the underlying assumptions of Kalman filtering are violated.

As noted by Spengler and Schiele,[22] recent research efforts that deal with object tracking can be classified into two categories, those that solve the tracking problem in a non-Bayesian framework[19–24] and those that tackle it in a Bayesian framework.[25–31] In many cases,[25–27] the focus is on single-object tracking. These single-object approaches usually rely on sophisticated, powerful object models. In other cases[28–31] the problem of tracking several objects in parallel is addressed. Some of these methods solve the multiobject tracking problem by employing configurations of individual objects, thus reducing the multiobject tracking problem to several instances of the less-difficult single-object tracking problem. Other methods employ particle-filtering-based algorithms, which track multiple objects simultaneously.

Despite the considerable amount of research that has been devoted to tracking, an efficient and robust solution to the general formulation of the problem is still lacking, especially for simultaneous tracking of multiple targets.

#### *3. Three-Dimensional Reconstruction*

Tracking provides a mechanism for associating observations of models over time. Still, it involves 2D information regarding the location of the tracked model. Providing 3D position information requires at least two observations of the same object, from different viewpoints. Although techniques based on

multiple views acquired from more than two cameras have been proposed,[32] most of the existing approaches[33] involve a single, calibrated stereoscopic system.

To the best of our knowledge, all existing approaches use a static stereoscopic system because employing a moving stereoscopic system would considerably complicate the process of tracking. If the stereoscopic system moves, everything changes in the fields of view of both cameras. Therefore, background subtraction (i.e., detection of temporal change) cannot be used as a means of providing additional evidence regarding the presence of moving skin-colored regions (SCRs). Further complications related to 3D reconstruction are introduced when the geometry of the stereoscopic system does not remain constant over time, i.e., when cameras with independent pan and tilt control are employed. Employing a moving stereoscopic system is often desirable, however, because in this case cameras can be purposefully positioned in a way that facilitates the observation of a certain activity.

An important implication of employing a stereoscopic configuration for computing 3D trajectories of tracked objects is that model detection and tracking should be performed in both views, thus increasing the computational requirements of the tracking system. In addition, an extra computational step is required that relates the two images of a stereo pair to each other. This is a crucial task because it permits the extraction of 3D information through standard 3D reconstruction techniques.[34,35]

### B. Proposed Approach to Three-Dimensional Tracking

In this paper we present our approach to 3D tracking of multiple SCRs observed by a moving stereoscopic system. This study was carried out in the context of a more-general research effort[36] toward developing a cognitive vision methodology to permit the interpretation of activities of people who are handling tools. Research and development are focused on the active observation and interpretation of the activities, on the extraction of the essential activities and their functional dependence, and on organizing the activities into their constituent behavior elements. The approach is active in the sense that the system seeks to obtain views that facilitate the interpretation of the activities observed. Therefore the ability to modify the viewpoint of observation of a certain activity is of utmost importance. Moreover, task and context knowledge is exploited as a means to constrain interpretation. Robust perception and interpretation of activities is the key to capturing the essential information that permits reproduction of task sequences from easy-to-understand representations.

The system that we propose is able to track and report the 3D trajectories of all SCRs that are present in a viewed scene. The proposed method for detecting SCRs has several attractive properties. A skin-color representation is learned through an off-line procedure. A new technique is proposed that eliminates much of the burden involved in generating training data. Moreover, the method adapts the skin-color model based on the recent history of tracked SCRs. Thus, without the need for complex models, the proposed approach is able to detect SCRs robustly and efficiently, even in conditions of changing illumination. The system employs a moving stereoscopic rig with cameras that have independent vergence control. To the best of our knowledge, this is the first method that is capable of tracking SCRs based on a moving stereoscopic system. Despite the motion of the cameras, the estimation of the 3D position of the detected and tracked SCRs is performed on a world-centered (i.e., extrinsic to the cameras) coordinate system. SCRs are tracked in time and associated with the images of each stereo pair by use of simple, computationally inexpensive techniques. The tracking system is implemented in C and can track multiple SCRs at a rate of 13 Hz on a Pentium 4 processor running Linux; the stereo stream employed consists of images with dimensions $640 \times 480$ pixels.

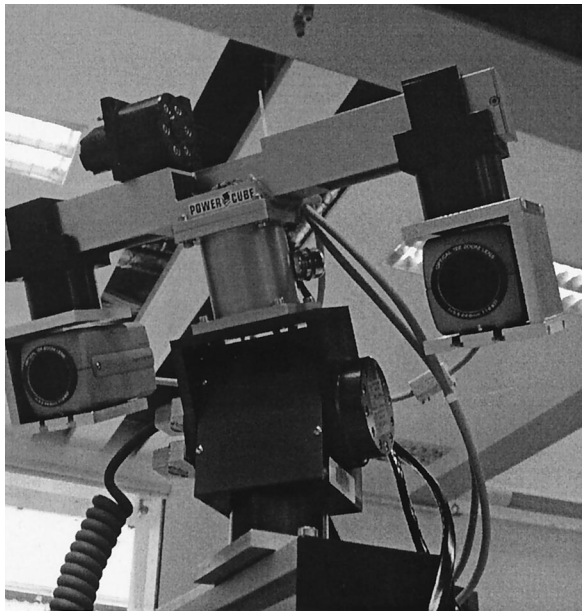


Fig. 1. Stereoscopic head (courtesy of Profactor GmbH) that is used to acquire stereo image pairs that are fed to the SCRT system.

The rest of the paper is organized as follows: In Section 2 we describe the skin-colored region tracker (SCRT). In Section 3 we provide sample results from operation of the SCRT in binocular image sequences. In Section 4, issues related to the computational performance of the SCRT are discussed. Finally, Section 5 provides the main conclusions of this research as well as information on its extension that is still under investigation.

## 2. Skin-Colored Region Tracker System

The SCRT is able to detect multiple SCRs and report their 3D positions by using images acquired by a moving stereoscopic head, such as the one shown in Fig. 1. Apart from providing raw stereo image streams, the stereoscopic head continuously provides

the position and orientation of each of the two CCD cameras with respect to a world-centered coordinate system. This is accomplished through the use of proprioceptive information provided by the motor encoders of the head. The SCRT exploits multiple cues toward achieving SCR tracking. These cues include color information and structure information as well as information regarding the camera positions and the epipolar geometry of the stereo system. In the remainder of this section we provide a brief outline of the SCRT system; more-detailed descriptions of its functional modules are provided in subsequent sections.

At each time instance $t$, the stereoscopic system acquires a synchronized image stereo pair, $I_L(t)$ and $I_R(t)$. Each of the pair's images is independently fed to a skin-color detection (SCD) module. SCD involves four key operations, specifically, (a) measurement of the probability of a pixel's being skin colored, (b) hysteresis thresholding on the derived probabilities map, (c) connecting components labeling to determine SCRs, and (d) computation of statistical information for each SCR (up to second-order moments). These SCRs, together with SCRs derived at the previous time instance $t - 1$, are then associated in time (AT module). The aims of using this module are (a) to assign a new, unique label to each new SCR (i.e., to a SCR that appears in the field of view for the first time) and (b) to propagate the labels of already detected SCRs in time. Then the SCRs detected in the left- and the right-hand images, along with the associated labels, are fed to a module that finds the correspondence between SCRs in the two images of the stereo pair (AS module). In fact, each SCR in the right image of the stereo pair is assigned the label of the corresponding SCR in the left image of the stereo pair if such a corresponding SCR actually exists. After completion of this type of association, the centroids of the corresponding SCRs are refined by a correlation-based stereo matching technique carried out by a centroid-matching (CM) module. This ensures that these points correspond to the same 3D scene point. The refined matches are then fed to a 3D reconstruction (3DR) module, which, taking into account the known geometry of the stereoscopic system as well as the intrinsic calibration parameters of the cameras, computes the 3D location of the centroid that pertains to each SCR. Finally, the 3D position that the system reports for each SCR is a weighted sum of 3D measurements in a sliding time window. The temporal smoothing (TS) module provides this type of functionality. A high-level block diagram that provides an overview of the SCRT system is illustrated in Fig. 2. In what follows, a more-detailed description of each of the aforementioned modules is provided.

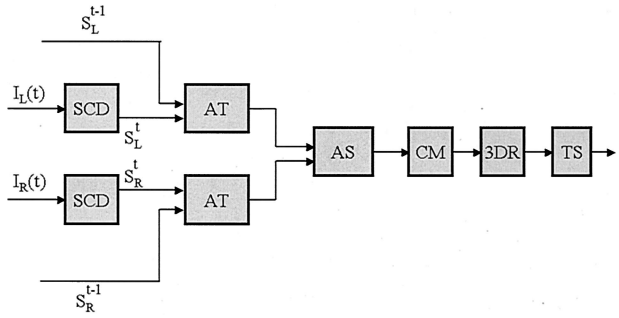

Fig. 2. Block diagram of the SCRT system.

### A. Skin-Color Detection Module

SCD is one of the fundamental building blocks of the SCRT system. The goal of the SCD module is to detect SCRs in an image. SCD adopts a Bayesian approach that involves an iterative training phase and an adaptive detection phase.

#### *1. Basic Supervised Training and Detection Mechanisms*

A set of training input images is selected upon which a human operator manually marks SCRs. The color representation used in this process[37] is YUV 4:2:2, which directly encodes the images acquired by the cameras used in the stereoscopic system of Fig. 1. However, the Y component of this representation is not employed for two reasons: (a) the Y component corresponds to the illumination of an image point, and therefore by omitting it the developed classifier gains some illumination-independent characteristics, and (b) employing a 2D color representation (UV), as opposed to a 3D representation (YUV), reduces the dimensionality of the problem and lowers the computational requirements of the overall system.

Assuming that image points $I(x, y)$ have a color $c = c(x, y) = (u, v)$, we use the training set to compute the following information:

- The prior probability $P(s)$ of having skin color in an image. This is the ratio of the skin-colored image points in the training set to the total number of image points.
- The prior probability $P(c)$ of the occurrence of each color $c$ in the training set. This is computed as the ratio of the number of occurrences of each color $c$ to the total number of image points in the training set.
- The prior probability $P(c|s)$ that color $c$ is a skin color. This is defined as the ratio of the number of occurrences of a color $c$ within the skin-colored areas to the number of skin-colored image points in the training set.

Based on the information extracted in the training phase, the probability $P(s|c)$ that a color $c$ is a skin color can be computed by use of the Bayes rule[38]

$$P(s|c) = \frac{P(c|s)\,P(s)}{P(c)}. \tag{1}$$

Then the probability that each image point $I(x, y)$ will be skin colored can be determined with the aid of a lookup table indexed with the point's color. The re-

sultant probability map is subsequently thresholded, and all image points with probability $P(s|c) > T_{max}$ are considered skin colored. These points constitute the seeds of potential SCRs. More specifically, image points with probability $P(s|c) > T_{min}$, where $T_{min} < T_{max}$, that are immediate neighbors of skin-colored image points are recursively added to the set of skin-colored points. The rationale behind this region-growing operation is that an image point with a relatively low probability of being skin-colored should be considered as such, when it is a neighbor of an image point with a high probability of being skin colored. This hysteresis thresholding type of operation has been successfully applied to edge detection[39] and also proves extremely useful in the robust identification of SCRs. Indicative values for thresholds $T_{max}$ and $T_{min}$ are 0.5 and 0.15, respectively.

A connected-components labeling algorithm is then responsible for assigning labels to the image points of various SCRs. Size filtering of the connected components is also employed to eliminate small, isolated blobs that are attributed to noise and do not correspond to interesting SCRs. Thus, connected components that consist of fewer than $T_{size} = 500$ image points are rejected from further consideration. Each of the remaining connected components corresponds to a SCR whose 2D image position is defined by its centroid.

*2. Adaptability*

The basic scheme for SCD described above has two major drawbacks:

- *Training*. Training is an off-line procedure that does not affect the on-line performance of the SCRT system. Nevertheless, it is a time-consuming process in the sense that a human operator should manually mark all skin-colored pixels in the chosen training set. Moreover, to obtain a training set that is capable of supporting tracking of various skin tones in images acquired from different cameras requires a large training set. Therefore, devising a method that will automate the processing of training data is considered quite important.
- *Detection*. When illumination conditions vary, the SCD module may produce poor results, despite the fact that the color representation employed has certain illumination-independent characteristics. Hence a method is required that adapts the representation of skin-colored image points according to the recent history of detected skin-colored points.

To cope with the first problem we developed an adaptive training procedure. Training is performed on an initial, small set of images for which the human operator provides ground truth by defining SCRs. Then detection, together with hysteresis thresholding, is used to update the prior probabilities $P(s)$, $P(c)$, and $P(c|s)$ continually in new images. The updated prior probabilities are then used to reclassify the full data set into skin-colored and non-skin-colored image points. When the classifier produces wrong results (false positives or false negatives), manual user intervention to correct these errors is necessary; still, up to this point the classifier has automatically completed much of the required work. The final training of the classifier is then performed based on the training set that results after user editing. This process for adapting the prior probabilities $P(s)$, $P(c)$, and $P(c|s)$ either can be disabled as soon as it is decided that the achieved training is sufficient for the purposes of the SCRT system or can continue as more input images are fed into the system.

At this point it is important to note that hysteresis thresholding is crucial for achieving the adaptation of prior probabilities described above. If hysteresis thresholding is not used, colors with probability $P(s|c) < T_{max}$ will never have the chance of being considered skin colors. Hysteresis thresholding with a threshold $T_{min}$ considerably smaller than $T_{max}$ allows colors with low probability of representing skin to be considered skin colors and permits the appropriate adaptation of their probabilities.

To solve the second problem, the SCD module maintains two sets of prior probabilities: $P(s)$, $P(c)$, and $P(c|s)$, which correspond to the training set, and $P_w(s)$, $P_w(c)$, and $P_w(c|s)$, which correspond to the evidence that the system gathers during the $w$ most-recent frames. Clearly, the second set better reflects the recent appearance of SCRs and is better adapted to the current illumination conditions. SCD is then performed based on

$$P_A(s|c) = aP(s|c) + (1 - a)P_w(s|c), \qquad (2)$$

where $P(s|c)$ and $P_w(s|c)$ are both given by Eq. (1) but involve prior probabilities that have been computed from the whole training set [for $P(s|c)$] and prior probabilities that have been computed from the detection results in the last $w$ frames [for $P_w(s|c)$]. In Eq. (2), $a$ is a sensitivity parameter that controls the influence of the training set in the detection process ($0 < a \leq 1$). If $a = 1$, then SCD takes into account only the training set, and no adaptation takes place; if $a$ is close to zero, then SCD becomes highly reactive, relying strongly on the recent past for deriving a model of the immediate future. Values of $a = 0.8$ and $w = 5$ gave good results in the tests that have been carried out.

A basic advantage of the proposed scheme lies in its simplicity. Other methods of adaptation have been proposed in the literature.[12,13] However, these methods require much more-complex modeling of the color characteristics of skin (i.e., modeling based on mixtures of Gaussians). An interesting study[6] has shown that, compared to mixture models, histogram models such as the one proposed in this paper provide better accuracy and lower computational cost for skin detection.

### B. Associating Skin-Colored Regions in Time: AT Module

As soon as a SCR is detected, it has to be tracked over time. This is a crucial function of the SCRT system

because it provides the temporal continuity of SCR observations.

We denote by $S_C{}^t$ the set of all SCRs detected at time $t$ in the image acquired by camera $C$ (left or right camera of the stereoscopic system). Also, let $S_C{}^t(i)$ be a particular SCR of this set with index $i$, $1 \le i \le N_C{}^t$, where $N_C{}^t$ is the cardinality of $S_C{}^t$. A distance measure $D_T[S_C{}^{t-1}(i), S_C{}^t(j)]$ between two SCRs, $S_C{}^{t-1}(i)$ and $S_C{}^t(j)$, that have been detected at times $t-1$ and $t$, respectively, is defined as follows:

$$D_T[S_C{}^{t-1}(i), S_C{}^t(j)] = |m_C{}^{t-1}(i) - m_C{}^t(j)|. \quad (3)$$

In Eq. (3), $m_C{}^t(k)$ denotes the centroid of SCR $S_C{}^t(k)$ and $||$ denotes vector magnitude. Equation (3) states that the distance between two SCRs is the Euclidean distance of their centroids. SCR $S_C{}^{t-1}(i)$ matches SCR $S_C{}^t(j)$, where

$$j = \arg\min_{1 \le k \le N_C{}^t} \{D_T[S_C{}^{t-1}(i), S_C{}^t(k)]\}. \quad (4)$$

The two SCRs, $S_C{}^{t-1}(i)$ and $S_C{}^t(j)$, are assumed to correspond to the same physical object if (a) $S_C{}^{t-1}(i)$ matches $S_C{}^t(j)$, (b) $S_C{}^t(j)$ matches $S_C{}^{t-1}(i)$, and (c) $D_T[S_C{}^{t-1}(i), S_C{}^t(j)] < T_D$ and $D_T[S_C{}^t(j), S_C{}^{t-1}(i)] < T_D$. Criterion (a) essentially states that, among all candidate SCRs at time $t$, the centroid of $S_C{}^t(j)$ is the closest to the centroid of $S_C{}^{t-1}(i)$. Symmetrically, criterion (b) states that, among all candidate SCRs at time $t-1$, the centroid of $S_C{}^{t-1}(i)$ is the closest to the centroid of $S_C{}^t(j)$. Criterion (c) provides an additional constraint on the actual distances of the centroids of matching SCRs. $T_D$ is a distance threshold that depends on the image-acquisition frame rate and on the velocity of the SCRs on the image plane. Assuming intrinsically calibrated images, it can be shown that

$$T_D > \frac{(|U_{\max}{}^C| + |U_{\max}{}^O|) f}{Z_{\min} \nu}. \quad (5)$$

In inequality (5), $U_{\max}{}^O$ is the maximum lateral component of 3D motion of a skin-colored object, $U_{\max}{}^C$ is the maximum expected lateral component of camera motion, $Z_{\min}$ is the minimum expected distance of this object from the camera, $f$ is the focal length of the camera measured in pixels, and $\nu$ is the image-acquisition frame rate. Then $T_D$ in inequality (5) corresponds to the maximum expected displacement between two successive images in time of a point on the image plane and is measured in pixels. The definition of $T_D$ above is very conservative in the sense that it corresponds to a lateral motion of the object with respect to the camera, with camera motion occurring in the direction opposite the object's motion and with the object located close to the camera. In practice, smaller values for $T_D$ are sufficient for associating SCRs in time.

The labels at time $t-1$ of all corresponding SCRs are propagated to the current time instance $t$. All SCRs detected at time $t$ that do not correspond to SCRs at time $t-1$ are assigned new labels, because they constitute regions observed for the first time.

The AT module is employed independently on the left and the right images of the stereo pair. The related process is simple and computationally cheap. Moreover, it proves robust in all cases when SCRs do not overlap because of occlusions.

### C. Associating Skin-Colored Regions in a Stereo Pair: AS Module

To provide information regarding the 3D position of each SCR, the tracker should be able to define the relationship of SCRs in the images of a stereo pair. This purpose is served by the AS module. As has already been stated here, it is assumed that the position and the orientation of each camera of the stereo pair are known with respect to a world-centered coordinate system. Based on this information, it is possible to compute the rotation matrix $R$ and the translation vector $t$ of the relative rigid motion between the coordinate systems of the cameras of the stereoscopic system. Knowledge of $R$ and $t$, in turn provides the means to compute analytically the fundamental matrix $F$ that captures the underlying epipolar geometry of the stereo pair[40]:

$$F = \frac{1}{\det(A_R)} [e_R]_x H_\infty, \quad (6)$$

where

$$e_R = A_R t, \quad (7)$$

$$H_\infty = A_R R A_L{}^{-1}. \quad (8)$$

In Eqs. (6)–(8), $A_L$ and $A_R$ are the intrinsic calibration matrices of the left and right cameras, respectively, $H_\infty$ is the homography of the plane at infinity, and $e_R$ is the epipole in the right image. $[e_R]_x$ denotes the skew-symmetric matrix associated with the vector cross product; i.e., for each vector $y$, $[e_R]_x y = e_R \times y$. Assuming that $m_L$ and $m_R$ are two corresponding points in the left and the right images of the stereo pair, then $m_R$ is constrained to lie on the epipolar line $l_R$, defined[35] as $l_R = F m_L$. Similarly, $m_L$ is constrained to lie on the epipolar line $l_L$, defined as $l_L = F^T m_R$ (see Fig. 3). The AS module employs the epipolar constraint to relate the SCRs of the images of a stereo pair. As in the case of the AT module, we denote by $S_L{}^t$ and $S_R{}^t$ the sets of SCRs that have been detected at time $t$ in the left and right images of the stereo pair, respectively. Moreover, $S_L{}^t(i)$ and $S_R{}^t(j)$ denote specific SCRs with indices $i$ and $j$ detected at time $t$, $1 \le i \le N_L{}^t$ and $1 \le j \le N_R{}^t$. A distance measure $D_S[S_L{}^t(i), S_R{}^t(j)]$ is defined between two SCRs, $S_L{}^t(i)$ and $S_R{}^t(j)$, as

$$D_S[S_L{}^t(i), S_R{}^t(j)] = \max\{d[F m_L{}^t(i), m_R{}^t(j)], d[F^T m_R{}^t(j), m_L{}^t(i)]\}. \quad (9)$$

In Eq. (9), $m_L{}^t(i)$ and $m_R{}^t(j)$ are the centroids of SCRs $S_L{}^t(i)$ and $S_R{}^t(j)$, respectively, and $d(l, p)$ denotes the

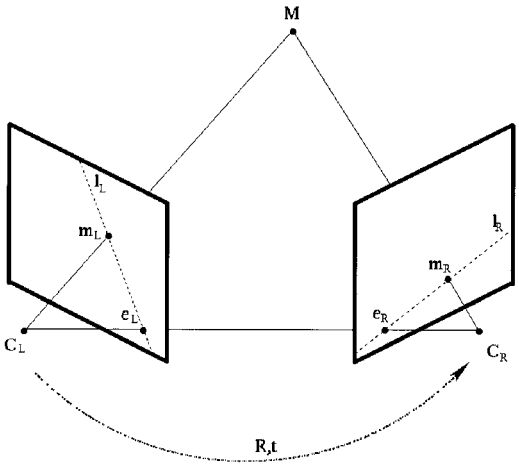


Fig. 3. Graphic illustration of the epipolar geometry of a stereo pair. Epipolar plane $C_L M C_R$ intersects the two image planes along epipolar lines $l_L$ and $l_R$.

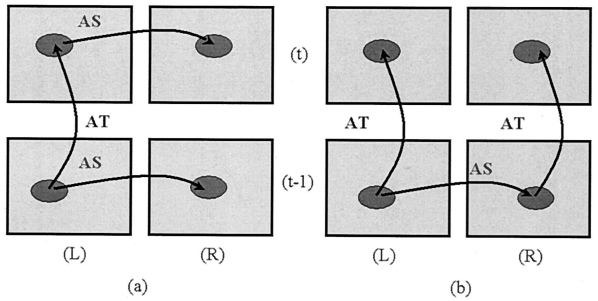


Fig. 4. Two methods of achieving the propagation of labels of SCRs both in time and between two stereo views. (a) The AS module is used to match SCRs of the left and the right images of the stereo pair at each point in time. The AT module is used to propagate labels in time in the image sequence of the left camera only. (b) The AS module is used to match SCRs only when a new SCR appears in the field of view. Separate AT modules are then used to propagate the SCR labels in time independently for the left and the right image sequences. As the AT module is typically more robust than the AS module, the second approach is adopted.

Euclidean distance of point $p$ from line $l$. SCR ${S_L}^t(i)$ matches SCR ${S_R}^t(j)$, where

$$j = \underset{1 \le k \le N_R{}^t}{\arg\min} \{D_S[{S_L}^t(i), {S_R}^t(k)]\}. \quad (10)$$

Symmetrically, SCR ${S_R}^t(j)$ matches SCR ${S_L}^t(i)$, where

$$i = \underset{1 \le k \le N_L{}^t}{\arg\min} \{D_S[{S_R}^t(j), {S_L}^t(k)]\}. \quad (11)$$

Two SCRs, ${S_L}^t(i)$ and ${S_R}^t(j)$, are assumed to correspond to each other if ${S_L}^t(i)$ matches ${S_R}^t(j)$, ${S_R}^t(j)$ matches ${S_L}^t(i)$, and $D_S[{S_L}^t(i), {S_R}^t(j)] < T_S$, where $T_S$ is a threshold that depends on the accuracy of the estimated epipolar geometry. Note that by definition this distance is symmetric; i.e., $D_S[{S_L}^t(i), {S_R}^t(j)] = D_S[{S_R}^t(j), {S_L}^t(i)]$.

For all corresponding SCRs, the label of the SCR in the left image propagates to the corresponding SCR in the right stereo image. All SCRs that have not been paired are excluded from further consideration in the subsequent process of 3D position estimation. Such SCRs typically correspond to skin-colored objects that are visible only in one of the two cameras of the stereo pair.

The method for matching SCRs described above becomes unstable if the epipolar geometry is not accurately computed. In this case the threshold $T_S$ has to be set conservatively to quite a large value, which leaves room for errors in matching the SCRs. For this reason, 3D position information from previous time instances can be used, if available. More specifically, when distances $D_S[{S_L}^t(i), {S_R}^t(j)]$ are computed, the 3D position of the SCR that results from the assumption that ${S_L}^t(i)$ actually corresponds to ${S_R}^t(j)$ is computed. If the resultant 3D position is invalid (in the sense that this position either is not plausible or differs substantially from the SCR's 3D position in the previous time step), a penalty term is added to the corresponding distance measure to guarantee that ${S_L}^t(i)$ and ${S_R}^t(j)$ will not be considered to correspond.

In general, the information on camera position and orientation that is computed from the encoders of the stereoscopic system is not reliable enough to permit an accurate estimation of the epipolar geometry of the stereo pair. For instance, experiments carried out with prerecorded image sequences indicated that the average distance of image points from their estimated epipolar lines was of the order of 15 pixels. The later error, in turn, affects the robustness of the AS module. To overcome this problem we apply AS only to SCRs that appear in the field of view for the first time. As soon as this is achieved, the AT module, which is more robust than the AS module, propagates the correct SCR labels in both images of the stereo pair, as is exemplified in Fig. 4. It should be mentioned at this point that a more accurate, image-based estimation of the stereo system's epipolar geometry[41] (as opposed to the currently employed encoder-based estimation) will considerably improve the robustness of the AS module. Still, such methods incur a significant computational overhead that is undesirable in the context of the SCRT system.

### D. Centroid Matching: CM Module

The matching of SCRs in the left and right images of a stereo pair leads to a rough correspondence between SCR centroids. This property can be used directly for deriving the 3D positions of SCRs. However, centroids are computed by the SCD module from the mean $x$ and $y$ coordinates of each SCR. Therefore it is not guaranteed that the left and right centroids of a SCR will correspond to the same 3D point. To refine this rough, initial correspondence we employ a correlation-based matching algorithm. Let $m_L$ and $m_R$ denote the centroids of a SCR in the left and the right images, respectively, of a stereo pair. Then a model template $W_M$ around $m_L$ and a search window $W_S$ around $m_R$ are defined. $W_M$ is

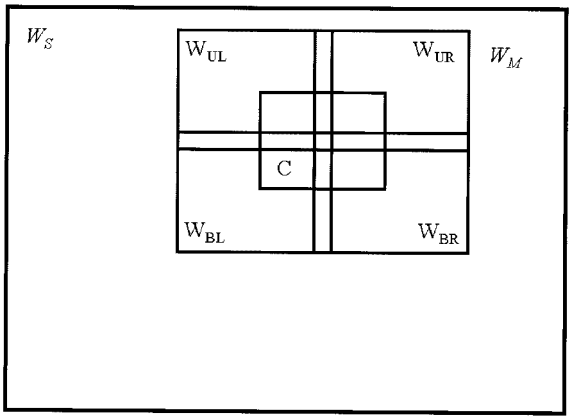


Fig. 5. Configuration of overlapping windows used in the correlation method proposed by Hirschmüller.[42]

placed over all possible positions in $W_S$, and a correlation measure $\Delta$ is computed. The location $m_R'$ in $W_S$ where correlation measure $\Delta$ is maximized ($\Delta_{MAX}$) is considered the refined right centroid of the specific SCR. We repeat the process symmetrically, by defining a model template around $m_R$ and a search window around $m_L$. If this search gives rise to a correlation score greater than $\Delta_{MAX}$ for some point $m_L'$ in the left image, then we consider the ($m_L'$, $m_R$) pair of centroid correspondences instead of the ($m_L$, $m_R'$) pair. This centroid refinement process is repeated for all pairs of corresponding SCRs. Note that if epipolar geometry has been computed accurately enough, search bands defined along epipolar lines can be used instead of search regions. Correlation measure $\Delta$ used in the CM module is inspired by the work of Hirschmüller[42] on dense stereo matching. Model template $W_M$ is divided into five overlapping subwindows, a central ($W_C$), an upper-left ($W_{UL}$), an upper-right ($W_{UR}$), a bottom-left ($W_{BL}$), and a bottom-right ($W_{BR}$). These five windows overlap as shown in Fig. 5. All subwindows have dimensions of $3 \times 3$ pixels, resulting in a $5 \times 5$ pixel model template. Inasmuch as in the work of Hirschmüller[42] a rectified stereo pair is assumed, search templates are essentially one dimensional. In our case, the dimensions of the search window are $25 \times 25$ pixels.

At each placement of model template $W_M$ in search window $W_S$, five correlation values, $\Delta_C$, $\Delta_{UL}$, $\Delta_{UR}$, $\Delta_{BL}$, and $\Delta_{BR}$, are independently computed. These values measure the correlation of each subwindow with the corresponding part in the search window. Then we can compute correlation value $\Delta$ for this placement by adding the values of the two best surrounding correlation windows, $\Delta_{max1}$ and $\Delta_{max2}$, to that of the middle window:

$$\Delta = \Delta_C + \Delta_{max1} + \Delta_{max2}. \tag{12}$$

In fact, in this approach a small central window is used, and the correlation decision is supported by four nearby windows. This formulation enables the refinement process to cope well with depth discontinuities and occluded and revealed regions that introduce errors when standard correlation is employed between the model window and the corresponding part of the search window. It should be noted that depth discontinuities and occlusions are common in this particular SCR tracking process, in which the SCRs are typically small image regions, closer to the cameras than to their immediate surroundings.

### E. Three-Dimensional Reconstruction of the Positions of Skin-Colored Regions: 3DR Module

The refined centroid correspondence of the SCRs are fed to a 3DR module, which computes the 3D position of each SCR. Two different reconstruction methods have been considered.

In the first method[40] the 3D position ($X$, $Y$, $Z$) of a point P, given its projections $m_L$ and $m_R$ in the left and the right images of a stereo pair, are computed as

$$\begin{aligned} Z &= -\frac{(m_R \times e_R)(m_R \times H_\infty m_L)}{\|m_R \times H_\infty m_L\|^2}, \\ X &= Z[A_L^{-1}(0)]m_L, \\ Y &= Z[A_L^{-1}(1)]m_L. \end{aligned} \tag{13}$$

In Eqs. (13), $m_L$ and $m_R$ are homogeneous vectors and $e_R$ and $H_\infty$ are defined as in Eqs. (7) and (8), respectively. Moreover, $[A_L^{-1}(r)]$ denotes the vector that corresponds to the $r$th row of the inverse of the intrinsic calibration-parameter matrix of the left camera. Equation (13) gives the 3D position ($X$, $Y$, $Z$) of point P with respect to the coordinate system of the left camera. The 3D position of this point with respect to the world-centered coordinate system can easily be computed through a rigid 3D transformation involving the position of the left camera with respect to the world-centered coordinate system.

In the second method the 3D position of a point is computed as the intersection of two 3D lines. More specifically, a 3D line is defined by 3D points $C_L$ and $M_L$, where $C_L$ is the optical center of the left camera and $M_L$ is the 3D position of the centroid of a SCR on the left image. Similarly, a second 3D line is defined by 3D points $C_R$ and $M_R$, where $C_R$ is the optical center of the right camera and $M_R$ is the 3D position of the centroid of the corresponding SCR in the right image. In the case of perfect, noiseless measurements these two lines should intersect at the desired 3D point. However, noise in the corresponding image coordinates $m_L$ and $m_R$ as well as inaccuracies in the calibration parameters will almost certainly result in these two lines' being skewed. Then the 3D location P of the SCR is[43]

$$P = \tfrac{1}{2}(C_L + \hat{v}_L s_L + C_R + \hat{v}_R s_R), \tag{14}$$

where

$$s_L = \frac{\det(M_R - M_L \hat{v}_R v_{LR})}{|v_{LR}|^2}, \quad s_R = \frac{\det(M_R - M_L \hat{v}_L v_{LR})}{|v_{LR}|^2}, \tag{15}$$

$$\hat{v}_L = \frac{M_L - C_L}{|M_L - C_L|}, \quad \hat{v}_R = \frac{M_R - C_R}{|M_R - C_R|}, \quad v_{LR} = v_L \times v_R. \tag{16}$$

If the 3D lines actually intersect, then P in Eq. (14) is their point of intersection. If the 3D lines are skewed, then P is the midpoint of the minimum-length line segment whose endpoints lie on the two 3D lines. Points $C_L$, $M_L$, $C_R$, and $M_R$ can be easily computed from the known 3D positions and orientations of the cameras with respect to the world-centered coordinate system and the knowledge of the refined centroids of corresponding SCRs.

The first reconstruction method is based on accurately computed epipolar geometry between the cameras of the stereo pair and on the availability of point matches that satisfy this epipolar geometry. If the first condition is satisfied and the second is not, there exist methods[34] to refine the point matches so as to enforce that refined matches satisfy the prescribed epipolar geometry. If, however, the epipolar geometry has not been accurately computed, the 3D reconstruction is inaccurate even for perfect matches. In the context of the SCRT system, the second method provides more-accurate 3D reconstruction results than the first reconstruction approach and, therefore, has been adopted in the 3DR module.

### F. Temporal Smoothing: TS Module

The temporal smoothing (TS) module performs temporal filtering of the derived 3D position of each SCR, based on the assumption that the 3D trajectory of a SCR is smooth as a function of time. The current implementation considers 3D positions $P_{t-2}$, $P_{t-1}$, and $P_t$ of a SCR as they were computed in the last three time instances, $t - 2$, $t - 1$, and $t$, and reports 3D position P, defined as a weighted average of these measurements:

$$P = 0.6P_t + 0.3P_{t-1} + 0.1P_{t-2}. \tag{17}$$

Weights are appropriately adapted whenever 3D position measurements in time instances $t - 2$ or $t - 1$ are not available.

## 3. Sample Results

In this section we provide representative results of the application of the SCRT system to sequences of stereo images. All experiments reported here as well as several others that are not included because of space limitations were conducted by employing the same set of parameters in the different SCRT modules.

Experiments with two stereoscopic sequences, each of which has been acquired by a different stereoscopic system, are reported here. More specifically, the first sequence (Sequence1_head) was acquired by the stereo head shown in Fig. 1, and the second sequence (Sequence2_arm) was captured by a different set of cameras mounted upon a robotic arm. Despite the fact that in both experiments there are four cameras, each with different color response characteristics, a single training set has been established and SCD has been based on the same set of probabilities derived through Eq. (1) for the images of all four cameras. The initial, seed training set contained 40 images (10 from each camera) and was later refined in a semi-automatic way (as described in Subsection 2.B) by use of 160 additional images (40 from each camera).

Sequence1_head shows a human operator manipulating a CD player (the operator opens the tray, picks up a CD, places it in the tray, and closes the tray). The stereoscopic system does not move in this experiment. The full sequence consists of 146 left and 146 right frames. Figure 6 (top to bottom, left to right) shows characteristic snapshots from the tracking results obtained. Every tenth frame is shown in this figure. For purposes of illustration, each SCR appears as a color blob superimposed upon the right image of the stereo pair. A cross marks the centroid of each SCR. Moreover, an ellipse derived from the statistics of each SCR is shown surrounding each color blob. It can be seen that the system identifies three SCRs, namely, the head of the human operator, the skin-colored arm of the armchair, and the hand of the human operator. It can be verified that the labeling of the SCRs is consistent throughout the whole sequence, which means that SCRs are correctly tracked both in time and between the images of the stereo pair.

Figure 7 shows the 3D trajectories computed by the system for the three tracked SCRs. The upper facet of the CD player has also been reconstructed to serve as a reference. The trajectory of the operator's hand appears to be qualitatively correct. In this sequence the arm of the armchair does not move, and the head of the operator moves slightly. To measure the stability of the derived 3D coordinates we computed the 3D bounding box of all estimated 3D positions for each SCR. For the static arm of the armchair, the dimensions of this bounding box are 1.2 cm $\times$ 1.5 cm $\times$ 1.6 cm, close to zero, as expected.

In the Sequence2_arm a human operator again manipulates a CD player. In this sequence the cameras move in time. The sequence consists of 134 left and 134 right frames. Figure 8 (top to bottom, left to right) shows characteristic snapshots from the tracking results. Every tenth frame is again shown in this figure. It can be seen that the system identifies two SCRs, which correspond to the head of the human operator and to the hand of the human operator. It can be verified that the labeling of the SCRs is consistent throughout the whole experiment, which implies that SCRs are correctly tracked both in time and between the images of the stereo pair. Figure 9 shows the 3D trajectory computed by the system for the hand of the operator. Note that, despite the motion of the cameras, the hand's trajectory appears smooth, which serves as an indication that only small range errors, which are due to SCD detection, centroid estimation, and camera motion estimation, are introduced.

Videos related to the above experimental results

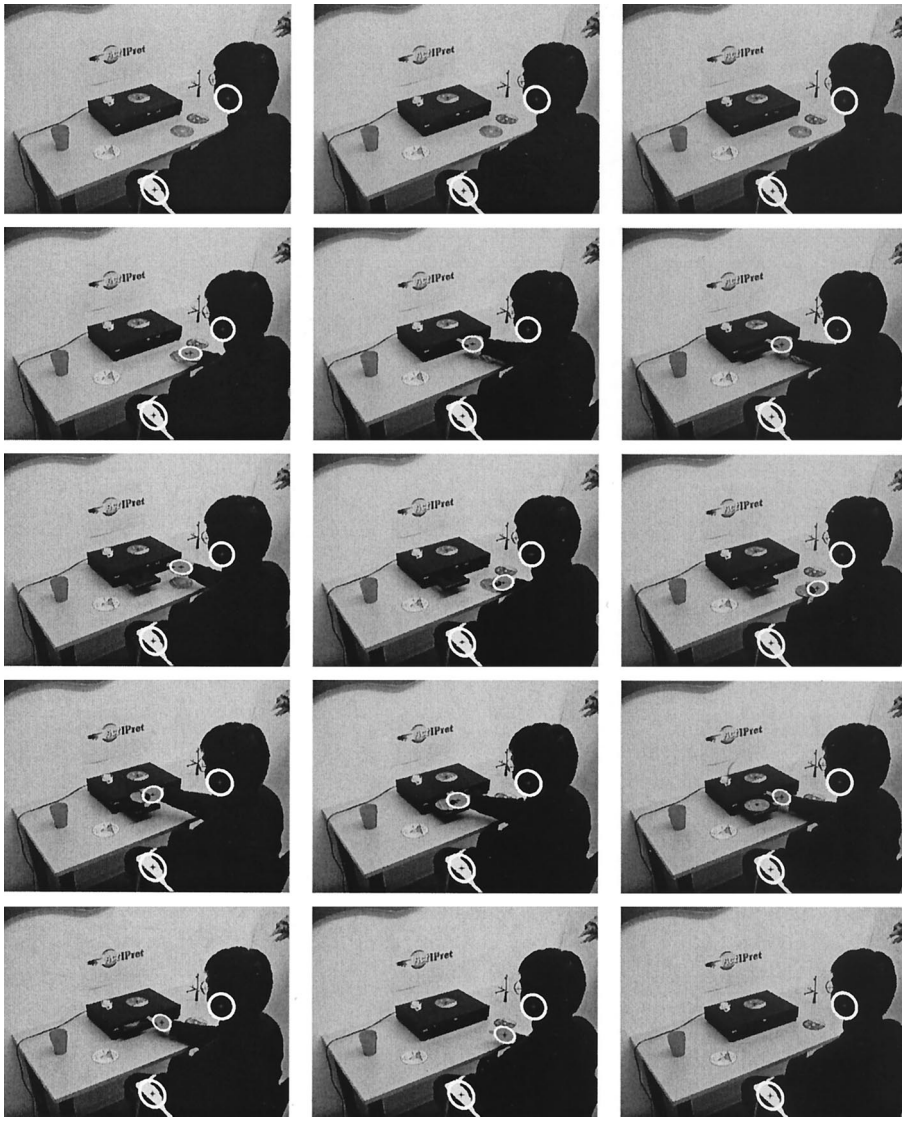

Fig. 6. Tracking results for Sequence1_head. Each SCR appears as a color blob superimposed upon the right image of the stereo pair.

can be retrieved at http://www.ics.forth.gr/cvrl/demos.html.

#### 4. Properties of the Skin-Colored Region Tracker's Computational Performance

Several tests aimed at assessing the functionality and the performance of a prototype implementation of the SCRT system have been carried out. Both off-line experiments (involving prerecorded image sequences) and on-line experiments were conducted. It turns out that one cycle of operations of the SCRT system takes approximately 75 ms to process 640 × 480 images on an Intel P4 processor at 1.8 GHz running Linux. The cycle includes all SCRT system functionality plus reading a stereo pair of 640 × 480 images from the hard disk. Approximately 40% of the cycle time is spent on image input–output, 40% on SCD, and the rest on the remaining modules. The SCRT system may be modified to operate on subsampled versions of the original images. More precisely, if the input images are subsampled by a factor of 2 (i.e., if 320 × 240 images are employed), the SCRT cycle time drops to 35 ms. The reason why the performance gain is not directly proportional to the reduction in input data (i.e., a factor of 4) is that SCRT always imports full-resolution images, which are then subsampled appropriately. Therefore, acquisition time is constant and independent of the operational image resolution. As an illustration of this claim, we measured image input–output and subsampling to account for almost 80% of the cycle time in the case of half-image resolution.

An important observation is that the SCR trajectories computed in full image resolution closely resemble the SCR trajectories computed at half-resolution. To illustrate this finding, we applied the SCRT system to Sequence1_head at both full and half resolution and computed the average distance between the reconstructed 3D positions of the hand in these two cases. This distance was of the order of 6

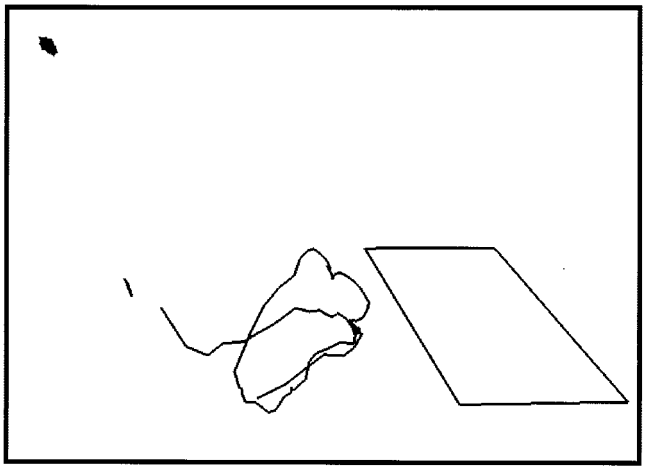

Fig. 7. 3D trajectories of the SCRs tracked in the experiment of Fig. 6. The top-left and middle-left isolated spots correspond to the motion of the operator's head and of the armchair's arm, respectively. The trajectory in the center of the image corresponds to the hand trajectory. The upper facet of the CD player has also been reconstructed, to serve as a reference.

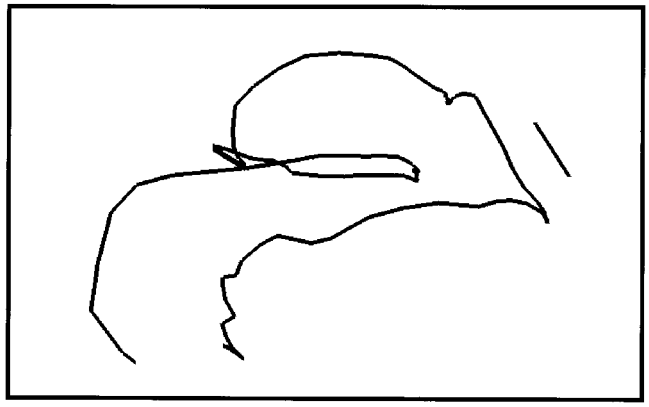

Fig. 9. 3D trajectory of the hand detected in the experiment of Fig. 8. The small straight-line segment that appears at the right corresponds to the tray of the CD player.

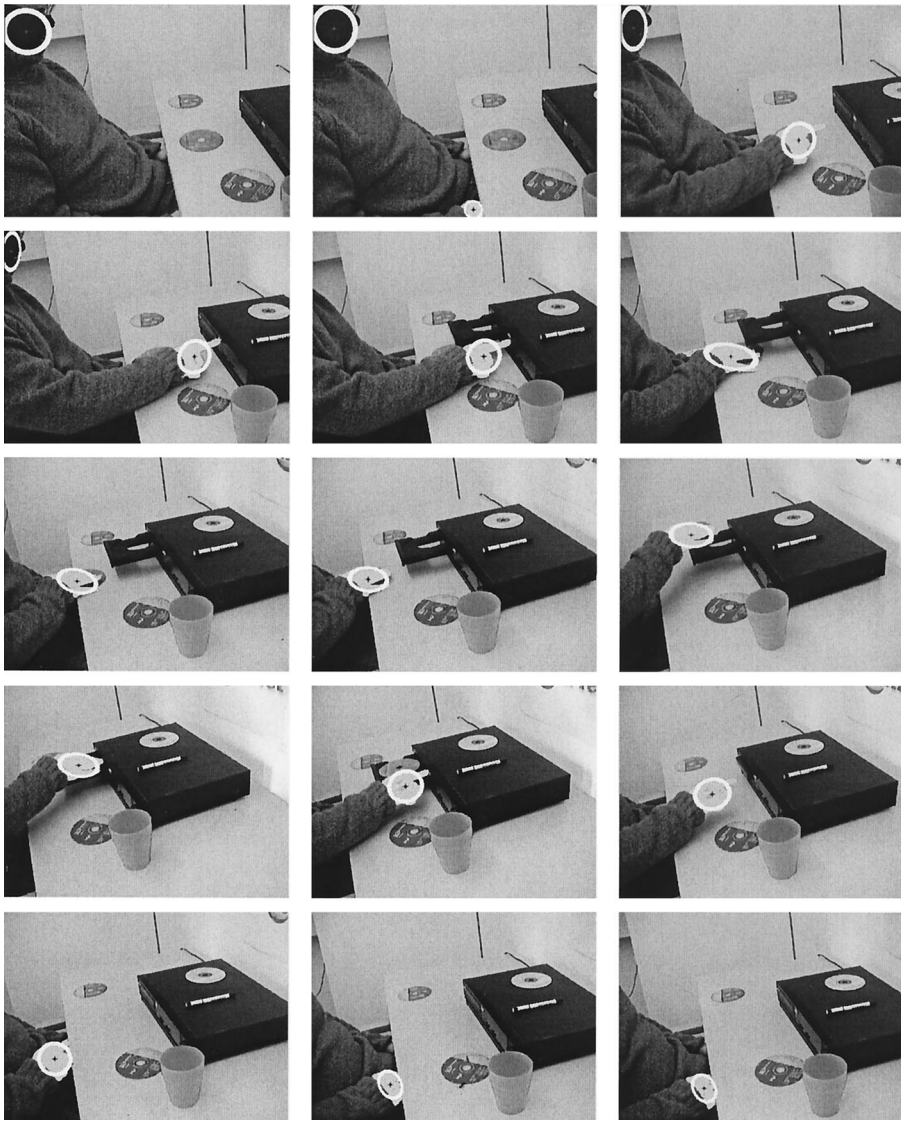

Fig. 8. Tracking results for Sequence2_arm. Each SCR appears as a color blob superimposed upon the right image of the stereo pair.

mm, thus demonstrating that a significant speedup can be obtained with little sacrifice of accuracy. Still, a rate of 13 Hz (for full resolution images) or 28 Hz (for half-resolution images) is considered sufficient for purposes of the SCRT.

## 5. Discussion

In this paper, a skin-colored region tracker system has been described. The SCRT is capable of detecting and tracking multiple SCRs in scenes viewed by a moving stereoscopic system in which each camera can move independently. The computational performance of the SCRT system is near real time when operating in full-resolution $640 \times 480$ images and can be considerably improved by subsampling the input images by a factor of two. In this case, the SCRT system operates in real time, without noticeable degradation of the quality of the computed 3D trajectories.

Ongoing research activities are focused on improving the robustness of the SCRT system in the case of overlapping SCRs. Tracking of each new SCR based on Kalman filtering[18] is expected to improve the capability of the SCRT system to handle occluded SCRs.

Currently, SCRs correspond to any skin-colored region in the employed images; moreover, each SCR is represented as a point in 3D space. However, in many cases it is desirable to focus attention on the activities of human hands, and it would also be desirable to provide information regarding the 3D position of each hand and the 3D positions of the fingertips. Another avenue of research aims at providing specialized hand models that will turn a SCRT system into a 3D hand tracker.

Finally, current research and development activities are targeted toward integrating vision-based camera tracking techniques[44] with the SCRT system. We expect that vision-based camera tracking will substantially improve the estimation of camera position and orientation, which will in turn improve accuracy in estimating the 3D position of a SCR. Additionally, this will alleviate the current dependence of SCRT on specialized equipment needed for continual monitoring of the position and orientation of the two cameras of a stereo rig.

This research was partially supported by EU Information Society Technologies project ActIPret (IST-2001-32184). The authors thank Stelios Orphanoudakis, Eleytheria Tzmali, Lena Gaga, Kostas Hatzopoulos, and Cedric Groyer (members of the Computational Vision and Robotics Laboratory of the Institute of Computer Science/Foundation for Research and Technology–Hellas) for fruitful discussions and contributions to the design and implementation of the SCRT.

The image sequences presented in this paper were acquired at the premises of Profactor GmbH, Steyr, Austria, whose contribution is gratefully acknowledged.

## References

1. K. Meyer, H. L. Applewhite, and F. A. Biocca, "A survey of position trackers," Presence **1,** 173–200 (1992).
2. D. M. Gavrila, "The visual analysis of human movement: a survey," Comput. Vis. Image Underst. **73,** 82–98 (1999).
3. Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with physical forces," Comput. (Vis. Image Underst. **81,** 328–357 (2001).
4. T. S. Jebara and A. Pentland, "Parameterized structure from motion for 3D adaptive feedback tracking of faces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1997), pp. 144–150.
5. S. H. Kim, N. K. Kim, S. C. Ahn, and H. G. Kim, "Object oriented face detection using range and color information," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1998), pp. 76–81.
6. M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1999), pp. 274–280.
7. D. Saxe and R. Foulds, "Toward robust skin identification in video images," 2nd International Conference on Automatic Face and Gesture Recognition (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1996), pp. 379–384.
8. D. Chai and K. N. Ngan, "Locating the facial region of a head-and-shoulders color image," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1998), pp. 124–129.
9. M. H. Yang and N. Ahuja, "Detecting Human Faces in Color Images," in *Proceedings of the IEEE International Conference on Image Processing* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1998), pp. 127–130.
10. J. C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *Proceedings of IEEE International Conference on Face and Gesture Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 2000), pp. 54–61.
11. J. Cai and A. Goshtasby, "Detecting human faces in color images," Image Vis. Comput. **18,** 63–75 (1999).
12. S. McKenna, Y. Raja, and S. Gong, "Tracking color objects using adaptive mixture models," Image Vis. Comput. **17,** 225–231 (1999).
13. Y. Raja, S. McKenna, and G. Gong, "Tracking and segmenting people in varying lighting conditions using color," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1998), pp. 228–233.
14. T. S. Jebara and A. Pentland, "Parametrized structure from motion for 3D adaptive feedback tracking of faces," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1997), pp. 144–150.
15. T. S. Jebara, K. Russel, and A. Pentland, "Mixture of eigenfeatures for real-time structure from texture," in *Proceedings of the Sixth International Conference on Computer Vision* (Narosa, Bombay, 1998), pp. 128–135.
16. M. H. Yang and N. Ahuja, *Face Detection and Gesture Recognition for Human–Computer Interaction* (Kluwer Academic, Dordrecht, The Netherlands, 2001).
17. M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," IEEE Trans. Pattern Anal. Mach. Intell. **24,** 34–58 (2002).

18. R. E. Kalman, "A new approach to linear filtering and prediction problems," Trans. ASME Ser. D. **82,** 35–45 (1960).
19. D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 2000), pp. 142–151.
20. O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *European Conference on Computer Vision* (Springer-Verlag, Berlin, 2002), pp. 343–357.
21. N. T. Siebel and S. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," in *European Conference on Computer Vision* (Springer-Verlag, Berlin, 2002), pp. 373–387.
22. M. Spengler and B. Schiele, "Multi-object tracking based on a modular knowledge hierarchy," in *International Conference on Computer Vision Systems* (Springer-Verlag Heidelberg, 2003), pp. 376–385.
23. C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 1999), pp. 246–252.
24. J. Triesch and C. von der Malsburg, "Democratic integration: self-organized integration of adaptive cues," Neural Comput. **13,** 2049–2074 (2001).
25. R. Fablet and M. J. Black, "Automatic detection and tracking of human motion with a view-based representation," in *European Conference on Computer Vision* (Springer-Verlag, Berlin, 2002), pp. 476–491.
26. M. Isard and A. Blake, "Icondensation: unifying low-level and high-level tracking in a stochastic framework," in *European Conference on Computer Vision* (Springer-Verlag, Berlin, 1998), pp. 893–908.
27. J. Vermaak, P. Perez, M. Gangnet, and A. Blake, "Towards improved observation models for visual tracking: selective adaptation," in *European Conference on Computer Vision* (Springer-Verlag, Berlin, 2002), pp. 645–660.
28. C. Hue, J.-P. Le Cadre, and P. Pérez, "Sequential Monte Carlo methods for multiple target tracking and data fusion," IEEE Trans. Signal Process. **50,** 309–325 (2002).
29. M. Isard and J. MacCormick, "Bramble: a Bayesian multiple-blob tracker," in *Proceedings of the International Conference on Computer Vision ICCV* (IEEE Computer Society, Los Alamitos, Calif., 2001).
30. E. Koller-Meier and F. Ade, "Tracking multiple objects using the condensation algorithm," J. Robot. Auton. Syst. **34**(2–3), 93–105 (2001).
31. P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," *Proceedings of the European Conference on Computer Vision* (Springer-Verlag, Berlin, 2002), pp. 661–675.
32. Y. Li, A. Hilton, and J. Illingworth, "A relaxation algorithm for real-time multiple view 3D-tracking," Image Vis. Comput. **20,** 841–859 (2002).
33. T. Inaguma, K. Oomura, H. Saji, and H. Nakatani, "Efficient Search Technique for Hand Gesture Tracking in Three Dimensions", in *International Workshop on Biologically Motivated Computer Vision* (Springer-Verlag, Berlin, 2000), pp. 594–601.
34. R. Hartley and P. Sturm, "Triangulation," Compu. Vis. Image Underst. **68,** 146–157 (1997).
35. O. Faugeras, Q.-T. Luong, and T. Papadopoulo, *The Geometry of Multiple Images* (MIT Press, Cambridge, Mass., 2001).
36. S. O. Orphanoudakis, A. A. Argyros, and M. Vincze, "Towards a cognitive vision methodology: understanding and interpreting activities of experts," ERCIM News, No. 53 (ERCIM EEIG, Sophia-Antipolis, France, 2003); http://www.ercim.org.
37. K. Jack, *Video Demystified: A Handbook for the Digital Engineer* (HighText, Solana Beach, Calif., 1993).
38. D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach* (Prentice-Hall, Englewood Cliffs, N.J., 2003).
39. J. F. Canny, "A computational approach to edge detection," IEEE Trans. Pattern Anal. Mach. Intell. **8,** 769–798 (1986).
40. L. Robert, C. Zeller, O. D. Faugeras, and M. Hebert, "Applications of non-metric vision to some visually guided robotic tasks," in *Visual Navigation: From Biological Systems to Unmanned Ground Vehicles*, Y. Aloimonos, ed. (Erlbaum, Hillsdale, N.J., 1997), Chap. 5, pp. 89–134.
41. Z. Zhang, "Determining the epipolar geometry and its uncertainty: a review," Int. J. Comput. Vision **27,** 161–195 (1998).
42. H. Hirschmüller, "Improvements in real-time correlation-based stereo vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, Piscataway, N.J., 2001), pp. 141–148.
43. R. Goldman, "Intersection of two lines in three-space," in *Graphics Gems*, A. S. Glassner, ed. (Academic, San Diego, Calif., 1990), Vol. 1, p. 304.
44. M. I. A. Lourakis and A. A. Argyros, "Efficient 3D camera matchmoving using markerless, segmentation-free plane tracking," Technical Report ICS/FORTH-TR-324 (Institute of Computer Science, Foundation for Research and Technology—Hellas, Heraklion, Greece, Sept. 2003).