Eurographics Italian Chapter Conference (2007)
Raffaele De Amicis and Giuseppe Conti (Editors)

A practical vision based approach to unencumbered direct spatial manipulation in virtual worlds

Fabio Bettio, Andrea Giachetti, Enrico Gobbetti, Fabio Marton, Giovanni Pintore

CRS4, POLARIS Edificio 1, 09010 Pula, Italy.

Abstract

We present a practical approach for developing interactive environments that allows humans to interact with large complex 3D models without them having to manually operate input devices. The system provides support for scene manipulation based on hand tracking and gesture recognition and for direct 3D interaction with the 3D models in the display space if a suitably registered 3D display is used. Being based on markerless tracking of a user's two hands, the system does not require users to wear any input or output devices. 6DOF input is provided by using both hands simultaneously, making the tracker more robust since only tracking of position information is required. The effectiveness of the method is demonstrated with a simple application for model manipulation on a large stereo display, in which rendering constraints are met by employing state-of-the-art multiresolution techniques.

1. Introduction

In recent years, the large demand for entertainment and games has resulted in major investments in commodity graphics chip technology, leading to low cost state-of-the-art programmable graphics units (GPUs) able to sustain rendering rates of hundreds of millions of graphics primitives per second. In order to fully harness the power of these GPUs, specialized adaptive multi-resolution techniques have been introduced to guarantee high frame rates with massive geometric models (e.g, [CGG*04, YSGM04, CGG*05, BGB*05]. As a result of this hardware and software trend, it is now possible, by using only commodity components, to develop applications offering very high quality real-time visualization of complex scenes and objects on very large screens and/or immersive 3D displays.

We envision a direct 3D interface for these environments that does not require users to hold and manipulate input devices, but rather accepts 3D gestures directly. Computer vision provides a natural basis for this sort of interaction method, because it is unobtrusive and flexible. Many interaction techniques have been proposed in this field (e.g., see [PL03]). However, the complexity of markerless computer vision tasks for full 3D pose recovery and gesture recognition has been a barrier to widespread use in real-time applications based on direct 3D manipulation.

In this short paper we report on practical techniques enabling the realization of interactive 3D environments based on full 6DOF manipulation. The approach is centered on a combination of dynamic color/background segmentation, dynamic modeling of hand motion, stereo matching to recover the 3D position of hands, and a priori knowledge about the interaction area. 6DOF input is supported by simultaneously tracking the position of both hands and interpreting their relative motion. The resulting system is thus made more reliable since only tracking of position information is required by the vision based subsystem. Moreover, the system does not require an initialization step, making it possible to naturally enter and exit interaction space.

The vision-based markerless hand tracking system can be applied to several 3D manipulation tasks. Its features are illustrated here through the control of the rendering of complex 3D models on a large stereo display.

2. Vision based interaction

There is obviously a full body of research on interaction within 3D virtual reality environments. The prototype discussed here is meant to work as an enabling technology demonstrator of fully device-less 3D solutions using manipulation of detailed 3D objects as driving application. Markerless 3D hand tracking is a major component of this application. Several solutions have been recently proposed to track hand motions and recognize gestures for human computer interaction. The methods applied differ in several respects (skin segmentation method, dimensionality of the tracking, temporal integration model, hand model, etc) and the development of a visual interface of this kind requires both knowledge of recent advances in tracking methods and a careful analysis of user requirements. Recent results, e.g., [BKMM*04, SMC01], show that complex hand models with at least 26 DOF can be tracked nearly in real time using smart model pose estimation methods. These methods, however, usually require complex parameter initialization and the availability of high quality depth or disparity maps as input. Practical applications, i.e. experimental game interfaces or sign language recognition systems, are based on simpler approaches, performing 2D tracking of the hand Region Of Interest and aspect based gesture recognition. The first step is usually performed through skin color segmentation in a chosen component space [ILI98, Xu03, AKE*04, MVMP05] or motion residuals [YSA05, SWTF04] and regions/blob tracking based on prediction/update schemes often exploiting Kalman or particle filters [Xu03, YSA05, SWTF04]. This last step can be achieved through the use of simple regional features or more complex procedures such as statistical classification or sequence analysis with Hidden Markov Models [ILI98, CFH03, SWTF04, MVMP05].

We follow this trend, proposing a simple approach that combines a hand region detection system based on a combination of skin color modeling and background subtraction, a stereo 3D tracker based on a prediction-update model and a simple aspect based hand shape recognition method. From the evolution of the two hands' states, it is possible to recognize and track different kind of gestures allowing an easy interaction with the virtual object. In this way we obtained a real time 3D tracking of two hands, with a rough gesture recognition, a combination of features that improves over current solutions, and that is suitable for the implementation of a variety of gestural interfaces. Another important peculiarity of the system is that does not need a complex initialization (we just assume that hands appear in front of the screen): the system detects the hands when they enter and then tracks them, automatically re-starting the hand detection loop when the tracking algorithm fails.

Therefore, the system also has the ability to recover from tracking failures due to noise, occlusions or illumination changes. We obtained this feature by keeping the hand model simple and exploiting a priori knowledge on the interaction context. This simple method is able to acquire in real time and without initialization tasks much more information on hand gestures than that used to implement the interface tested here. Moreover, with this kind of gestural interface, we only need to detect the hands' barycenter positions and to recognize whether the hands are open or closed, but the low level vision system is not forced to track hands orientation. This last point simplifies the image processing tasks, making them simpler and more robust at the same time.

3. A practical markerless approach to hand tracking

In order to support two-hand 3D interaction with a 3D environment, we do not need the localization of an articulated 3D model of both hands. However, we need a fully 3D real time tracking of the position of the two hands, and the recognition of a few simple postures. Furthermore, in order to naturally support 3D interaction, the system must work without complex initialization procedures and must be able to autonomously detect the entrance and exit of hands in the volume of interaction.

The solution adopted consists of a hand region detection system based on color thresholding and a stereo 3D tracking with a simple aspect based posture recognition. The hands' tracker setup is sketched in figure 1. A couple of calibrated stereo cameras are placed under the interaction volume with parallel optical axes pointing towards the ceiling and common x axes. This solution allows an optimal view of the gesture and a simple disparity computation (small perspective effects, horizontal search for disparity computation).

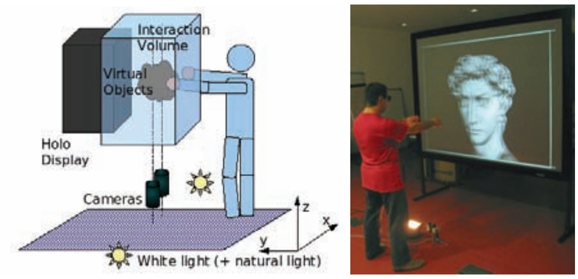

Figure 1: ***Vision based tracker setup.*** *Two calibrated stereo cameras are placed under the interaction volume with parallel optical axes pointing towards the ceiling and common x axes. This solution provides full coverage of the gesture space and simple disparity computation.*

The dynamic model used for tracking consists of two hand objects (left and right). Each object includes a status flag (active or inactive hand), 3D position and velocity, the region of interest and the orientation of the projected hands in the left image and the current posture label (open, closed or pointing hand). Once the system is activated, a continuous hand detection system starts searching for skin regions independently in the left and right part of the left stereo image.

A classical method to segment skin regions is based on ad hoc partitioning of a color space [PBC05, FG06]. Another approach consists in assigning skin class probability to color histogram entries (in a selected color space with a chosen quantization) given a distribution model and after a training phase [JR02, ZSQ99, SSA00, COB02, PBC05]. Our approach is similar to the latter. The overall hand detection function hp of an image pixel is given by a weighted sum of the following values:

- p_c is a color based pixelwise skin probability depending on r-g color components. This probability is precomputed on a training set using a multiple Gaussian model and stored in a look-up table;
- p_t is a probability component depending on the difference between the current image and an acquired background model;
- p_n is a term depending on the pixel neighborhood, which increases the hand probability of a pixel when it is surrounded by neighbors with high skin color probability.

We estimate hand region by thresholding this value.

If skin is detected, a tentative left/right hand region of interest (ROI) is created around the left/right skin pixel closest to the display, and if the skin area size in the ROI is compatible with a hand, the left/right hand tracker is started, all the hand parameters are computed (through binary mask analysis and disparity computation) and the corresponding hand detection loop is disabled.

The hand tracker works as shown in Figure 2: on the basis of the current status, a prediction of the ROI position and size (depending on the z coordinate) is generated. If the ROI is outside the interaction region or there is a collision with the other hand, the ROI is automatically shifted in order to keep the hand in the ROI and separated from the other. The skin region is then re-segmented in the predicted ROI and centroid (and ROI) position, orientation and gesture are recomputed. In case of bad measurements (too few skin pixels in the region) hand parameters are cleared and the hand is set as inactive, otherwise the stereo disparity in the ROI is computed by a 1D sum-of-squared-differences (SSD) minimization algorithm on a subsampled window [Gia00], the 3D coordinates of the hand are recovered from calibration parameters and the posture label is evaluated. The posture label is computed by a simple classifier which uses shape descriptors (area, elongation) and the z coordinate as features.

In the ROI tracking we do not perform averaging between measurement and prediction as done in Kalman trackers. This procedure is not particularly useful in this kind of applications due to the large variability of the inter-frame motion and the lack of knowledge about measurement and model noise. Weighted averaging between prediction and measurement is applied only for the z coordinate estimation computed on the basis of the disparity evaluated inside the hand ROI. This is done because the matching algorithm can fail in some cases due to clutter.

The measurement model used to track hands' ROIs makes also useless a multiple hypotheses approach such as the one performed in classical particle filters [AMGC02]. The sort of "mean shift" algorithm here applied to the ROI, makes multiple hypotheses useful only in the case of large differences between samples (otherwise samples tends to collapse to a single one [AMGC02]).

In future we plan to introduce multiple hypotheses generating only few samples with large hand centroid variations: the ROI position in the particle samples will be derived from the analysis of 2D hand acceleration histograms acquired during a training phase.

Finally, in our tracking method occlusions and hands superimposition are handled introducing a constraint so that left and right ROIs cannot be superimposed. In this way, if a user crosses his hands, one hand simply disappears. The continuity of the detection loops ensures that, when hands are separated again, the two regions are correctly re-detected.

Figure 3 shows a frame of the processed stream captured live during an interactive session.

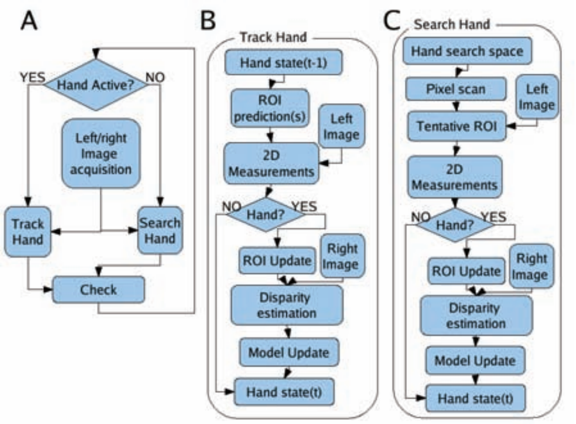

Figure 2: ***Hand tracking flowchart.*** *If hands are idle, the hand detection loop (B) is activated in the interaction region. If one hand is active, the tracking module (C) is iteratively performed instead.*

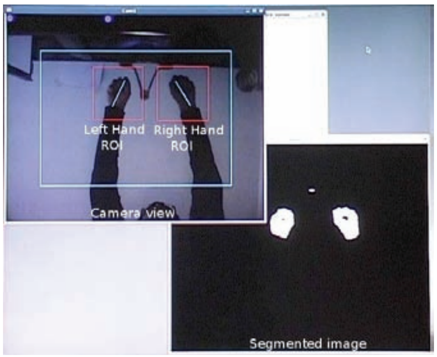

Figure 3: ***Hand tracking.*** *Frame captured live during an interactive session.*

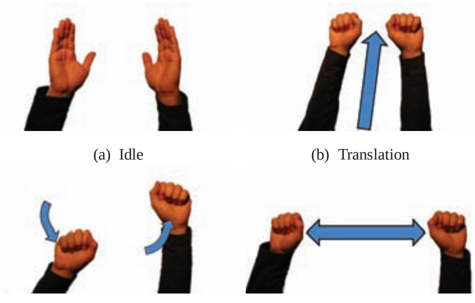
(a) Idle (b) Translation (c) Rotation (d) Scaling

Figure 4: ***Manipulation gestures.*** *Closing both hands initiates the gesture. Object rotation, translation, and scaling are recognized by selecting the dominant relative motion.*

4. Two-handed manipulation

Three types of two-hand gestures are recognized: translation, rotation and scaling. A movement starts when both hands are in the working area of the two cameras and we detect that they are closed (as it would be done to grab a real object), and stops when the hands release the object or when they move out of the working area. At each moment from the starting time to the final time the object is moved according to three rules. Translation is accomplished by moving the two hands in parallel toward a desired point where we want to place the model. The model is moved by the relative translation between the starting point where hands have been closed and the current position. Rotation is performed by rotating both hands around their barycenter: the object is rotated around a predefined pivot (which is generally in the object center). The rotation axis is defined as the cross product of the two vectors 'connecting' the two hands at the starting time and at the current time, while the rotation angle is the angle between the two vectors. Scaling is done simply by moving the two hands apart, or moving a hand closer to the other one. The scaling factor is defined by the ratio between the initial and the current distance between the two hands.

When the two hands start to grab the object, we try to identify the type of movement that the user wants to perform by selecting the dominant type (translation, rotation or scaling). Each type of movement is measured and the first one whose measure is greater than a predefined threshold value is selected and will be used until the object is released. The translation measure is the distance of the barycenters of the two hands from the starting time to the current time; the rotation measure is given by the length covered by a point rotating around the two hand barycenter with a radius equal to half the mean distance between the two hands, and by an angle given by the rotation estimation, and the scaling is measured by the difference between the initial distance and the current distance between the two hands. While performing the selection no movement is applied to the object; the threshold used to identify the movement is 30mm, which is big enough to identify the type of movement properly, but not big enough to introduce unwanted response delay.

5. Implementation and Results

We have implemented a prototype hardware and software system based on the design discussed in this paper.

As a simple illustration of our tracker's current status and capabilities, we have tested it by integrating it into a system for large scale model visualization based on the Tetra-Puzzles technique [CGG*04] applied to high resolution laser scanned artifacts.

Camera acquisition and hand tracking is performed on a Pentium4 3GHz PC equipped with a mvSIGMA-SQ frame grabber connected to two PAL cameras. The PC is connected by an Ethernet 1Gb/s link to an Athlon64 3300+ PC with a NVIDIA7900GTX PC running the graphics application. A large scale stereoscopic display assembled from off-the-shelf components is used to show images to the users. The display consists in two 1024x768 DLP projectors connected to the two outputs of the graphics card, polarizing filters with matching glasses, and a backprojection screen that preserves polarization. Thanks to the performance of the multiresolution technique, a single PC is able to render two 1024x768 images per frame at interative rates (over 30Hz) while rendering multi-million datasets.

The interactive sequence depicted in figure 5 consists in a short free-hand manipulation of Michelangelo's David 1mm model (56M triangles; data courtesy of Stanford University). Using a calibrated large stereo display, objects appear floating in the display space and can be manipulated by translating, rotating, and scaling them with simple gestures. Under controlled lighting conditions, the hand tracker performance was satisfactory, as the tracking system was able to work at 10-15Hz rates, including frame acquisition times, while providing stable 3D positions and recognitions of open/closed states. Using two hands to input 3D transformations proved natural and reliable.

6. Conclusions and Future Work

We have presented a practical working implementation of an interactive display where the stereo visualization of a huge 3D scene can be controlled through a gesture-based interface. The system does not require users to wear any input or output devices. Vision-based tracking of a user's two hands provides for direct 3D interaction with the 3D models in the display space. The prototype discussed here is clearly meant to work as an enabling technology demonstrator, as well as a testbed for integrated 3D interaction, visualization, and display research. From the user interaction point of view, even the current simple hand tracking and gesture recognition system is already sufficient to obtain a simple deviceless interaction with the virtual scene without the need of

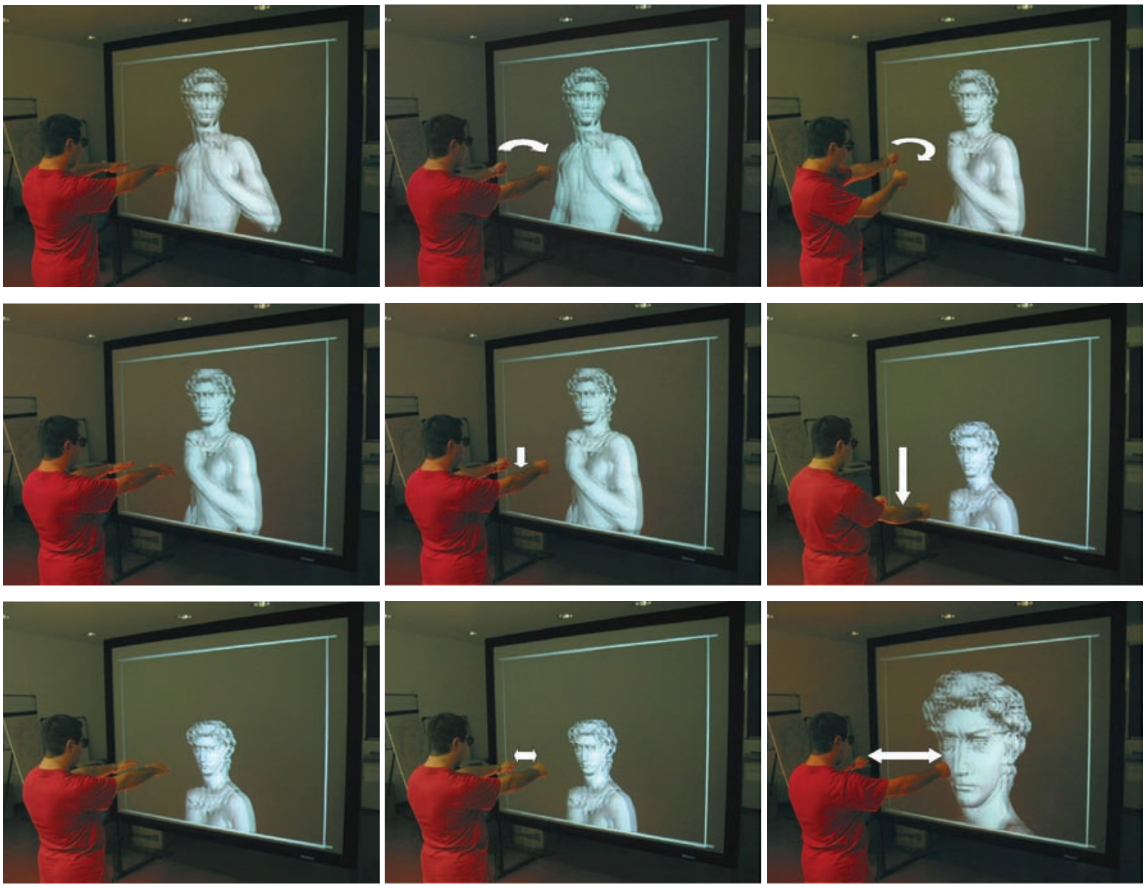

Figure 5: ***Interaction sequence.*** *These images illustrate successive instants of interactive manipulation of the David 1mm dataset (56M triangles) using a stereoscopic display coupled with the vision based hand tracker.*

relevant user training. We are currently working to improve its performances in different respects: hand model tracking will be improved by testing an adaptive change of the skin color model in order to make the system adaptive to variable illumination conditions and by testing different methods to perform multiple hypothesis tracking; the posture dictionary will be expanded and the number of recognized gestures will be increased after a task analysis for the different required interactions and usability tests on different proposed gestures; automatic calibration methods will be developed to make the system easily adaptable to different positionings, environments and displays. There is obviously more to user interaction than simple object manipulation. Devising general user interfaces that leverage the unique features of immersive 3D displays and computer vision methods is a challenging area for future work.

Acknowledgments. We are grateful to the Stanford Graphics Group and the Digital Michelangelo project. This research is partially supported by the COHERENT project (EU-FP6-510166), funded under the European FP6/IST program.

References

[AKE*04] Askar S., Kondratyuk Y., Elazouzi K., Kauff P., Schreer O.: Vision-based skin-colour segmentation of moving hands for real-time applications. In *1st European Conference on Visual Media Production (CVMP)* (March 2004), Chambers A., Hilton A., (Eds.), IEE.

[AMGC02] Arulampalam M. S., Maskell S., Gordon N., Clapp T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE TRANSACTIONS ON SIGNAL PROCESSING 50* (2002), 174–188.

[BGB*05] Borgeat L., Godin G., Blais F., Massicotte P., Lahanier C.: GoLD: interactive display of huge colored and textured models. *ACM Trans. Graph 24*, 3 (2005), 869–877.

[BKMM*04] BRAY M., KOLLER-MEIER E., MUELLER P., GOOL L. V., SCHRAUDOLPH N. N.: 3d hand tracking by rapid stochastic gradient descent using a skinning model. In *1st European Conference on Visual Media Production (CVMP)* (March 2004), Chambers A., Hilton A., (Eds.), IEE, pp. 59–68.

[CFH03] CHEN F.-S., FU C.-M., HUANG C.-L.: Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and Vision Computing 21* (2003), 745–758.

[CGG*04] CIGNONI P., GANOVELLI F., GOBBETTI E., MARTON F., PONCHIO F., SCOPIGNO R.: Adaptive TetraPuzzles – efficient out-of-core construction and visualization of gigantic polygonal models. *ACM Transactions on Graphics 23*, 3 (August 2004), 796–803. Proc. SIGGRAPH 2004.

[CGG*05] CIGNONI P., GANOVELLI F., GOBBETTI E., MARTON F., PONCHIO F., SCOPIGNO R.: Batched multi triangulation. In *Proceedings IEEE Visualization* (Conference held in Minneapolis, MI, USA, Oct. 2005), IEEE, pp. 207–214.

[COB02] CAETANO T., OLABARRIAGA S., BARONE D.: Evaluation of single and multiple-gaussian models for skin color modeling. In *Proc. XV Brazilian Symposium on Computer Graphics and Image Processing* (2002).

[FG06] F. GASPARINI R. S.: Skin segmentation using multiple thresholding. In *IS&T/SPIE Symposium on Electronic Imaging 15-19 January 2006 San Jose, California USA* (2006).

[Gia00] GIACHETTI A.: Matching techniques to compute image motion. *Image and Vision Computing 18* (2000), 245–258.

[ILI98] IMAGAWA K., LU S., IGI S.: Color-based hands tracking system for sign language recognition. In *ICFGR, Nara, Japan* (1998).

[JR02] JONES M. J., REHG J. M.: Statistical color models with application to skin detection. *International Journal of Computer Vision 46*, 1 (2002), 81–96.

[MVMP05] MANRESA C., VARONA J., MAS R., PERALES F. J.: Hand tracking and gesture recognition for human computer interaction. *Electronic Letters in Computer Vision and Image Analysis 5* (2005), 9–104.

[PBC05] PHUNG S. L., BOUZERDOUM A., CHAI D.: Skin segmentation using color pixel classification: Analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 1 (2005), 148–154.

[PL03] P. LEMOINE F. VEXO D. T.: Interaction techniques: 3d menus-based paradigm. In *AVIR* (2003).

[SMC01] STENGER B., MENDONCA P., CIPOLLA R.: Model-based hand tracking using an unscented kalman filter, 2001.

[SSA00] SIGAL L., SCLAROFF S., ATHITSOS V.: Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. *cvpr 02* (2000), 2152.

[SWTF04] SHAN C., WEI Y., TAN T., F.OJARDIAS: Real time hand tracking by combining particle filtering and mean shift. In *Proc. Sixth IEEE Int. Conference on Automatic Face and Gesture Recognition (FGR'04)* (2004).

[Xu03] XU L.-Q.: Simultaneous tracking and segmentation of two free moving hands in a video conferencing scenario. In *Proc. Sixth IEEE Int.* (2003).

[YSA05] YUAN Q., SCLAROFF S., ATHITSOS V.: Automatic 2d hand tracking in video sequences. *wacv-motion 01* (2005), 250–256.

[YSGM04] YOON S.-E., SALOMON B., GAYLE R., MANOCHA D.: Quick-vdr: Interactive view-dependent rendering of massive models. In *VIS '04: Proceedings of the IEEE Visualization 2004 (VIS'04)* (2004), IEEE Computer Society, pp. 131–138.

[ZSQ99] ZARIT B., SUPER B., QUEK F.: Comparison of five color models in skin pixel classification. In *proc. Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems* (1999), pp. 58–63.