International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:11, 2011

Combining Skin Color and Optical Flow for Computer Vision Systems

Muhammad Raza Ali and Tim Morris

Abstract**—**Skin color is an important visual cue for computer vision systems involving human users. In this paper we combine skin color and optical flow for detection and tracking of skin regions. We apply these techniques to gesture recognition with encouraging results. We propose a novel skin similarity measure. For grouping detected skin regions we propose a novel skin region grouping mechanism. The proposed techniques work with any number of skin regions making them suitable for a multiuser scenario.

Keywords**—** Bayesian tracking, chromaticity space, optical flow gesture recognition

I. Introduction

COMPUTER vision based systems have wide ranging applications from surveillance, biometrics to human computer interaction..

Skin color proves very useful in applications involving human users. Many state of the art techniques have been developed however skin color based techniques cannot be overlooked. We combine skin color with optical flow for an accurate background subtraction. For tracking we combine, for the first time, skin color and optical flow in a Bayesian (CONDENSATION algorithm) framework.

Computer vision research in single user applications is at an advanced stage [20] and vision based TV sets and teaching aids will soon be commercially available. On the other hand, research aimed at applications involving multiple 'active' users is comparatively at the early stages. Systems developed at MIT CSAIL [21] and work done by Del Bimbo et. al.[3] is an important step. Our techniques can be used in a multiuser, unrestricted setting where users are free to move and do not require markers or data gloves.

II. Overview of the Paper

Any computer vision system involving human users can be divided into three components; a) location of users and background subtraction, b) tracking users over a period of time and c) classification or recognition e.g. action or gesture recognition. We describe novel techniques for the first two stages.

M.R. Ali is a PhD candidate with the Advanced Interfaces Group, School of Computer Science, University of Manchester, UK. (phone: 00 44 161 2113717 e-mail: alim@ cs.man.ac.uk).

T. Morris is a Lecturer in the School of Computer Science, University of Manchester, UK. He is part of the Advanced Interfaces Group (e-mail: tmorrris@cs.man.ac.uk).

Detection of Skin Regions (Section III): This is the background subtraction step where we discard the irrelevant image data and identify skin regions.

Tracking of Skin Regions (Section IV): The skin regions are tracked using a Bayesian framework, the skin regions are not only tracked but they are given priority based on a novel joint criteria of skin similarity and optical flow magnitude. We also propose a novel skin similarity measure. As an example we employ our segmentation and tracking techniques for recognition of static gestures (Section V).

Detected skin regions will appear as 'floating' blobs. We are tracking individual regions but it might be desirable for some applications to group skin regions belonging to the same person (e.g. studies involving human-human interaction where we want to keep track of gesture frequency). We propose a novel technique for grouping detected skin regions and report accuracy of above 90 % (Section VI).

III. Background Subtraction

Various techniques have been developed for skin color detection, [13] provide a comprehensive survey and comparison of these techniques. We have used the skin color distribution in normalized red, green space i.e. skin color locus:

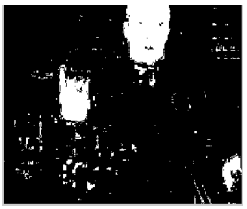
$$r = R/(R+G+B), \qquad g = G/(R+G+B) \tag{1}$$

Thresholds are specified and used to classify a pixel as skin or non-skin. Thresholding is conceptually straightforward; for a pixel to be classified as skin it should fall within a certain range of normalized red, green values. This is an extremely efficient method for detection of skin regions. Taking advantage of consistency in skin color distribution we suggested a simple method [1] for determining the skin color locus by taking image samples of hands and parts of face. This did not require camera calibration, knowledge of light source temperature as suggested in previous works [16], [17]. In [1] we identified credible thresholds for indoor applications requiring minimal adjustment. For red the range is 0.399-0.65 and for green 0.25-0.35. The technique showed high skin detection accuracy on publicly available sign language and web image datasets [1].

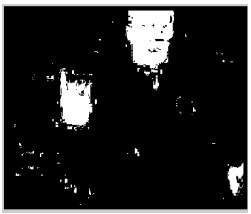
We encountered the issue of false positives that can occur in color based foreground detection. This issue can be avoided in application areas such as video games [19] and manipulation of virtual objects [15] where a cluttered environment may not play a major role. Some systems are evaluated only in a non-cluttered environment. However, for some applications like

gesture recognition the false positive regions can affect the accuracy of the system. Skin color based techniques identify false positives in their evaluation but do not present a solution. We have combined skin color with optical flow information to address the issue of false positives. Therefore, to be classified as a valid skin region and a candidate for our tracking system we propose a novel joint-threshold based on skin locus (i.e. normalized [red, green]) and optical flow magnitude. As shown in figure 1 the number of false positives is much less and the size of these regions is smaller when we use joint theresholding.

Image Frame[5]



Skin color based segmentation



Optical flow + skin color Based segmentation

Fig. 1 Comparison of segmentation techniques

Some studies [15], [19] have separately used skin color and optical flow the former is used for detection while the latter for motion estimation. This works if we assume that false positives are minimal and they can be easily removed by simple image processing techniques. This assumption will not hold in a cluttered environment, hence use of joint threshold is recommended.

After identifying the foreground, we have a binary image of detected skin/candidate regions. These candidate regions are to be tracked and used for gesture recognition (explained in detail in the next sections). Since we are using a joint threshold, every skin region may not qualify as a candidate region e.g. a hand static for prolonged period. We are only interested in tracking regions that are useful for gesture recognition.

IV. Tracking Mechanism

Our tracker combines skin color and optical flow in a novel way to track candidate regions. We propose a tracking mechanism based on the CONDENSATION algorithm which was proposed by Isard and Blake [10] based on the concept of factored sampling (appendix I). The CONDENSATION algorithm and its various extensions have been successfully applied to tracking of moving objects [11] and to gesture recognition [4]. Our tracking mechanism is fully capable of handling arbitrarily changing numbers of candidate regions. The challenge for our tracker is not only to track reliably the valid candidate regions but also to assign priority to a particular region for feature extraction and subsequent gesture recognition. Our proposed mechanism is novel in this aspect. This is useful in a multiuser application scenario where the system is expected to identify and track several skin regions and select a specific region for the feature extraction and classification. Each candidate region is treated as a particle for the Bayesian tracker.

A. Observation Vector

By direct processing of image frames and background subtraction we can express the image data in the form of an observation vector as:

$$z_t = [\alpha^- \quad \mu_{rg} \quad \sigma_{rg} \quad (x,y)^- \quad (w,h)^-] \tag{2}$$

Where α^- is the average optical flow magnitude, μ_{rg} is the average chromaticity value and σ_{rg} is the standard deviation of r,g values in the region. While the last two terms represent the centroid and dimensions of the bounding box around the region.

B. State Vector

Each candidate region is represented by a state vector x at time t*:*

$$x_t = [\alpha \quad \beta \quad (x,y) \quad (w,h)] \tag{3}$$

Where at a time step *t*: α is the average optical flow magnitude, β is the skin-similarity measure of the candidate region. The last two parameters indicate the centroid and dimensions of the candidate region/bounding box.

C. Novel Skin-Similarity Measure

Some sources of false positives are difficult to remove such as light colored hair or clothes, as they can not only be of similar color to skin but also have similar motion. These regions will be addressed through our tracking mechanism. The state-parameter β indicates the percentage of color values that lie within one standard deviation of the mean of [r, g] value for that region. This measure is very effective in determining whether a candidate region is a valid skin region or a false positive. This will assign low priority to difficult false positive regions thus improving the tracker performance. The human skin has an important characteristic that was observed during our skin detection experiments i.e. its chromaticity distribution is much more compact while a non-skin sample has a more scattered distribution. Therefore, a higher ratio of pixels is expected to lie within one standard deviation for a skin region. This novel similarity measure is simpler than traditional distance measures plus it is based on an important characteristic of skin color distribution. Although

we have used normalized red, green space for our tracker, this measure can be used in any color space or its respective chromaticity space as skin color occupies a certain part of a color or chromaticity space.

D. Modelling System Dynamics

In order to model movements/actions of users we have used the constant velocity motion model as described in [7,8]. State vectors at the current time step t are propagated from the previous time step as:

$$x_t = Ax_{t-1} + Bw_{t-1} \quad (4)$$

Where, the first tem in equation (4) represents the deterministic component of system, where A is the state transition matrix and second term is used to model uncertainties due to noise. The term w_{t-1} represents the normal random variables, that when combined with matrix B gives the stochastic component of our system. Matrices A, B for the motion model were determined from the training sequences of actions/gestures that our tracker is expected to accurately track. The *process noise* represented by $\boldsymbol{Bw}$ in (4) is caused mainly by the error is optical flow computation and to a small extent by drastic change in illumination conditions. Similarly for observation at time t, we have the following relation with the state vector:

$$z_t = Hx_t + v_t \quad (5)$$

Where, H is called the measurement matrix and v_t is the observation noise (caused primarily by sensor error).

D. Updating the Sample Weights

Each particle is weighted based on the observation data. The weight depicts the *importance* of a particle in a sample at t; this is potentially useful in our application as it can be used to assign priority to regions of interest. The weight of a particle at time-step t gives us the observation density as:

$$\pi_t = \exp(\alpha_t .(\mathbf{b}\beta_t)) \quad (6)$$

Weighting coefficient **b** is assigned a value of 0.25, 0.5, 1, 2 or 3. The value assigned to **b** depends upon the value of β. If β is less the 0.3 (less than 30% pixels lie within one standard) the coefficient **b** is assigned the lowest value 0.25. Similarly, if β is above 80%, **b** is assigned the highest value 3.

E. Evaluation of Tracker

The tracker was evaluated on four video sequences. Two clips are from the BOSS train surveillance datasets and two video sequences were recoded by the authors in a cluttered lab environment. We evaluate the tracker on the criteria that it should be able to 1) track all candidate regions and 2) assign top priority to a particular region based on the criteria described above. Surveillance sequences are good tests for our tracker, as they contain multiple users and periods of low and high activity. The evaluation results were encouraging. The error in computation of optical flow was kept in mind therefore the evaluation has been done using two algorithms; the Lucas Kanade [12] and Classic+NL [18]. Both algorithms, for both criteria produced accuracy of above 90%. The latter is ranked among the top ten in terms of error by Middlebury's evaluation of all major optical flow algorithms, for details of their methodology interested reader can refer to the link http://vision.middlebury.edu/flow/eval for details. The detailed breakdown of evaluation results is given in table 3. The comparison with previous work is difficult as to the best of our knowledge no previous implementation or extension of CONDENSATION algorithm has been specifically developed to track segmented skin regions. Also criteria set for multi-person tracker evaluation is, for the first time, from the point of a multiuser application setting.

It is worth mentioning that skin color and optical flow were combined in a Bayesian framework only to demonstrate that the established techniques can be adapted for multiuser applications by introducing the concept of priority assignment. As future work we can develop a novel, non-Bayesian tracker. The major strength of Bayesian framework is it's ability to track through clutter. Based on evaluation results for lab sequences our techniques can handle clutter, either removing non-skin image data or assigning it a low priority. Therefore, removing the reliance on Bayesian/ particle filter based tracking for cluttered backgrounds.

V. Application To Gesture Recognition

In this section we describe the gesture recognition results for the tracked candidate region. The region with the highest priority is to be used for gesture recognition at a particular time step. For the purpose of feature extraction we have implemented a Radon-descriptor recently purposed by Song et. al. [22]. This is defined as the projections of line integrals at various angles in an image. The most important characteristic of the radon transform is rotation invariance. For classification we have used the SVM-RBF Kernel of a widely used support vector machine library LIBSVM and 1 vs Rest classification scheme. Our gesture vocabulary consists of four gestures i.e. five, two, pointing and fist as shown in figure 2. For training we obtained around 500 images from various sign language datasets and images of four participants at three different locations making these gestures at least five times. As recommended in [9] we obtained the best parameters for our classification algorithm by cross validation on our training set. For testing we obtained more images of our participants under different settings plus images from DCU dataset [5]. We also included sequences from RWTH German finger spelling dataset [6] that contains images from 20 different users acquired at two camera angles at various times of the day. Our test set consists of around 2000 image frames. All the image frames are of size 320x240 pixels. The classification results we obtained are very encouraging and show the potential of using our proposed techniques with established classification algorithms; the results of our experiments are given in figure 3 as a confusion matrix.

VI. Grouping Skin Regions

The ability for the system to group skin regions will make it simpler to combine hand gestures with other visual cues. For example once we know which hands and face belong together it will be possible to associate gesture making hands with other cues like gaze, facial expressions etc. This can be extremely useful in developing a multimodal system.

Fig. 2 The gesture vocabulary

	Five	Two	Pointing	Fist
Five	99.0	0.0	0.0	1.0
Two	0.25	96.0	3.75	0.0
Pointing	0.0	2.99	97.0	0.01
Fist	0.0	0.0	0.0	100.0

Fig. 3 The confusion matrix

The skin region grouping technique proposed in [2] is perhaps the most important work done in this regard. They use a non-Bayesian technique to track and group skin regions. They have shown good results for a single-user and robustness for fast movements and overlapping of skin regions. Recently a simpler version of the algorithm has been implemented [19] for a gesture based game controller.

We propose a simpler, yet effective technique for grouping skin regions, for the first time in a multiuser scenario. Below we describe the core of our technique, and describe evaluation by testing on challenging video sequences and suggest an improvement in the technique.

A. Proposed Technique

The proposed method groups the candidate regions in two phases: firstly on the basis of difference in mean [red, green] values among regions and then on the basis of spatial location in the image frame. Both phases of the technique are outlined in the following figures.

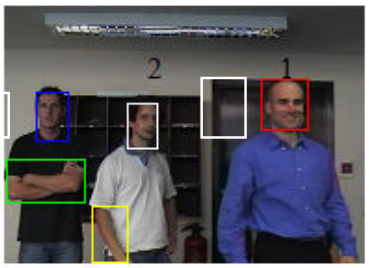


Fig. 4 Phase1 of skin region grouping technique

The frame in figure 4 shows five valid candidate regions. We create two matrices; one each for red and green chromaticity values that indicate minimum difference between regions in normalized red, green space. We extract information for region pairs that exist in both matrices. It may be noted that useful information may not be available from phase 1, as we only consider region pairs that feature in both matrices. In this example, regions marked 1 and 2 are considered to be belonging to one person as enough skin chromaticity information is available. Three regions are left unattached. Any groups formed here are ranked as unlikely (but possible), likely or highly likely based on spatial position checks. In this case regions 1 and 2 have been declared as an unlikely grouping. It is worth mentioning that in this step it is possible that no skin region groups are formed.

In the second phase we use the spatial information to group unattached skin regions. We look for vertical and horizontal overlap of the ungrouped regions. This is done by iteratively extending the width. The vertical overlap is given more importance. At this stage we obtain an *association matrix* that shows how strong the possibility of two regions belonging to a person is. Pairs of regions are ranked as 2(high), 1(medium) and 0.5(low).

The final grouping is created and any confusion is resolved by referring to information available at the end of the first phase or skin color information. As shown in the figure 5 below, regions have been correctly grouped. At the end of the first step regions 1 and 2 were declared an unlikely grouping, this has been verified.

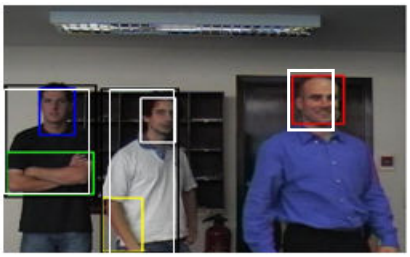


Fig. 5 Phase 2 of skin region grouping technique.

The proposed technique was evaluated on six video sequences. Three from the SPEVI dataset, one from the BOSS dataset (both datasets are available at http://www.multitel.be/cantata) and two that were recorded in our lab. The sequences had two to seven people in them. The

results are shown below in Table 1. Results for the lab sequences were significantly better, although participants moved freely there is a reasonable gap between participants as we would expect in a normal group. The results for other clips were satisfactory as these are very challenging and participants move randomly. However, we do not want the system performance to be affected by the way users move, we observed that by improving the heuristic used to resolve confusion we can improve the skin grouping results. We propose a modification in the final step of the technique whilst keeping the original technique as it is.

A. Modifying the Technique

We add a simple step of face detection in order to resolve any confusion and use faces as seeds to verify the final groupings. A straightforward check is performed to see if a group contains more than one region that is classified as 'face'; any such grouping is rectified. And for associating non-face regions to faces and to each other we use the *association matrix* as before. The modified technique improved our results as shown in Table II.

TABLE I
EVALUATION RESULTS FOR SKIN REGION GROUPING TECHNIQUE

Dataset	Total Image Frames	Participants	Frames Currently Grouped (%)
SPEVI(multi_face_fast)	485	3	77%
SPEVI(multi_face_frontal)	1276	4	69%
SPEVI(multi_face_turning)	1000	4	70%
BOSS(No_Event Cam 7)	2200	7	78%
Lab sequence 1	1290	2	88%
Lab sequence 2	950	3	86%
		Average	**78%**

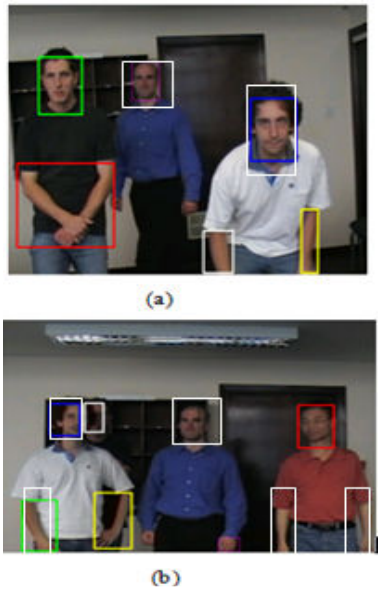


Fig 6.The regions in above examples were correctly grouped only when the modification to the original technique was made.

TABLE II
EVALUATION RESULTS FOR MODIFIED SKIN REGION GROUPING TECHNIQUE

Dataset/Clip	Frames Currently Grouped (%)
SPEVI(multi_face_fast)	94%
SPEVI(multi_face_frontal)	92%
SPEVI(multi_face_turning)	92%
BOSS(No_Event Cam 7)	88%
Lab sequence 1	95%
Lab sequence 2	95%
Average	**92%**

VII. Conclusion and Future Work

We propose an effective background subtraction technique that uses optical flow to minimize false positive regions. We have extended a Bayesian tracking mechanism for skin region tracking by combining optical flow and skin color distribution. The tracking mechanism also assigns priority to the skin regions. We also present a novel skin similarity measure. Any or all of the aforementioned techniques can easily be used in existing systems. The Radon-transformation based descriptor is used for feature extraction and subsequent gesture recognition through RBF-SVM. All these components form a gesture recognition system that has been evaluated on multiuser video sequences and have shown encouraging results.

It is worth mentioning that our techniques are the first skin color based techniques aimed at and evaluated from the point of view of multiuser applications. We would like to apply our techniques to an actual multiuser application. We are interested in evaluation of our techniques in a collaborative, problem solving scenario.

The development tool we have used is MATLAB. In future we would like to implement the system in C++ and OpenCV in order to achieve efficient performance on live video streams. Another option could be an implementation on the GPU. An optical flow algorithm has been developed for NVIDIA GPUs and has been used for optical flow based video games [14].

Appendix: Factored Sampling

The aim of the Bayesian tracking framework is to determine the posterior distribution $p(x|z)$. Using the same notation in the paper, x (state vector) represents our object of interest i.e. human skin region. And z is the observation data available to us at a particular time step. The posterior distribution is given by Bayes' rule:

$$p(x|z) = k\, p(z|x)p(x) \qquad (7)$$

The prior is given by $p(x)$ while likelihood term $p(z|x)$ is the observation density and k is the normalization constant. So our objective is to find an object x given the observation data z. However, the computation of posterior density is not straightforward as in practical computer vision applications p(z|x) is often multimodal thus no close form solution is

available for (7). To work around this problem we use factored sampling; the posterior distribution is computed at points. Suppose we are using n particles for our application. We choose a sample {s1 ,…….,sn } from the prior distribution p(x). The probability that a particle is chosen is determined by the conditional observation density (or likelihood) πi given as:

$$\pi^j = p(z|x = s^j) / \sum_{j=1}^{n} p(z|x = s^j) \qquad (8)$$

Increasing the number of particles improves the approximation of the posterior distribution.

REFERENCES

[1] M.R. Ali and T. Morris. "Skin Locus Based Skin Detection for Gesture Recognition". BMVC Postgraduate Workshop, Aberystwyth, 3-Sep-2010.
[2] A. Argyros and M. Lourakis. "Real time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera". In Proceedings of the European Conference on Computer Vision (ECCV'04), Springer-Verlag, vol. 3, pp. 368-379, Prague, Czech Republic, May 11-14, 2004.
[3] A. Del. Bimbo, L. Landucci , and A. Valli. "Multi-User Natural Interaction System based on Real-Time Hand Tracking and Gesture Recognition". In Proceedings of ICPR (3). 2006, pp: 55-58.
[4] M. Black and A.D. Jepson. "A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions". In Proceedings of Fifth European Conference on Computer Vision Vol. 1, 1998, pp: 909–924.
[5] T. Coogan, G. Awad, J. Han, and A. Sutherland. "Real time hand gesture recognition including hand segmentation and tracking". In Proceedings of ISVC06, 2006 pp: 495 – 504.
[6] P. Dreuw, T. Deselaers, D. Keysers, D, and H. Ney. "Modelling image variability in appearance-based gesture recognition". In ECCV Workshop on Statistical Methods in Multi-Image and Video Processing (ECCV-SMVP), Graz, Austria, May 2006, pp: 7-18.
[7] A. Gelb. "Applied Optimal Estimation". MIT Press, 1996
[8] M.S. Grewal and A.P. Andrews "Kalman Filtering". Prentice Hall, 1993.
[9] C. Hsu, C. Chang, and C.J Lin. "A Practical Guide to Support Vector Classification". http://www.csie.ntu.edu.tw/~cjlin/
[10] M. Isard and A. Blake. "CONDENSATION -- Conditional Density Propagation For Visual Tracking". Int. Journal of Computer Vision, Vol: 29(1), 1998, pp: 5-28.
[11] E. Koller-Meier and F. Ade, "Tracking Multiple Objects Using the Condensation algorithm". Robotics and Autonomous Systems, vol. 34, 2001, pp. 93-105.
[12] B. Lucas and T. Kanade. "An Iterative Image Registration Technique With An Application To Stereo Vision". Proceedings of Imaging understanding workshop, 1981, pp: 121-130.
[13] L. Saigol , S. Sclaroff , and V. Athitsos. "Skin Color-Based Video Segmentation Under Time-Varying Illumination". IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 26, No. 7, July 2004.
[14] J. Santner , M. Werlberger, T. Mauthner, W. Paier , and Bischof H. "Flow Games". 1st Int. Workshop on Computer Vision for Computer Games (CVCG) in conjunction with IEEE CVPR, 2010.
[15] G. Shin. and J. Chun, "Vision-based Multimodal Human Computer Interface based on Parallel Tracking of Eye and Hand Motion". In Proceedings of International Conference on Convergence Information Technology , 2007, pp. 2443-2448.
[16] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen, "Adaptive Skin Color Modelling Using The Skin Locus For Selecting Training Pixels". Pattern Recognition,36(3), 2003, pp: 681-690.
[17] M. Storring , J.H. Andersen, and E. Granum, "Skin Color Detection Under Changing Lighting Conditions". In Proceedings of Seventh Symposium Intelligent Robotics Systems,1999. pp. 187-195.
[18] D. Sun, S. Roth and M. Black, Secrets of Optical Flow Estimation and Their Principles". In proceedings of CVPR, 2010. http://www.cs.brown.edu/~dqsun/pubs/cvpr_2010_flow.pdf
[19] N. Vo, Q. Tran, T.B. Dinh and Q.M. Nguyen, "An Efficient Human Computer Interaction Framework Using Skin Color Tracking and Gesture Recognition". International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2010.
[20] Y. Wu and T. Huang. "Vision Based Gesture Recognition: A review", In International Gesture Workshop (GW99),Gif-sur-Yvette, France, 1999.
[21] Y. Yin and D.R. Toward, "Natural Interaction In The Real World: Real-Time Gesture Recognition". International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10), Beijing 2010.
[22] S. Zhan, Y. Hanxuan, Z. Yanguo, and Z. Feng, "Hand detection and gesture recognition exploit motion times image in complicate scenarios". In Proceedings of the 6th international conference on Advances in visual computing - Volum. Part II. Springer-Verlag, Berlin, Heidelberg, 2010, 628-636.