Robust Automatic Tracking of Skin-Colored Objects with Level Set based Occlusion Handling

Liang-Guo Zhang$^{1,2,3}$, Xilin Chen$^2$, Chunli Wang$^4$, and Wen Gao$^{1,2,3}$

$^1$ Key Lab of Intelligent Inf. Processing, Chinese Academy of Sciences (CAS), China
$^2$ Institute of Computing Technology, CAS, China
$^3$ Graduate University of the CAS, China
$^4$ School of IST, Dalian Maritime University, China

{lgzhang, clwang, xlchen, wgao}@jdl.ac.cn

Abstract. The automatic detection and tracking of multiple skin-colored objects in videos plays important roles in gestural human-computer interaction, such as in gesture and sign language recognition. In this paper, we propose a hybrid tracking framework, which consistently combines the blob based temporal data association and the level set based spatial occlusion handling through online shape learning, to track variable number of skin-colored objects. Blob based data association can provide rapid and accurate tracking when there’s no occlusion; while level set based occlusion handling can recover the occluded shape as precisely as possible. In this way, the robust complete region tracking to preserve and provide much more useful information for future high-level vision tasks is guaranteed. We demonstrate the effectiveness and efficiency of this approach by experimenting on several video sequences with complex motion patterns.

Key words: Objects tracking, occlusion handling, human-computer interaction, gesture recognition, sign language recognition

1 Introduction

The automatic tracking of skin-colored objects in videos is an important research topic. For example, the high performance tracking will definitely improve the gestural human computer interaction. Since our tracking work aims for the future visual gesture and sign language recognition (VGSLR), and especially SLR on medium or large vocabulary, some coarse tracking [9] (e.g., only extracting the location or/and rough geometry information of the targets) is not sufficient. Thus, to achieve the medium or large vocabulary visual sign language recognition, usually colored-gloves or markers are used to alleviate the targets segmentation (e.g., [8,22]). Therefore, during tracking, we want to preserve more rich information of the targets such as silhouette (i.e., complete region within the target’s boundary) for the future feature analysis. And we call it the complete object region tracking.

Although the work of contour tracking has been studied for decades, it’s still not possible to implement the contours tracking for multiple non-rigid objects
“economically” up to date. For example, Rathi et al. [18] proposed a generic framework, computationally very expensive, for tracking highly deformable objects by incorporating dynamic shape prior information into a particle filtering algorithm [10]. On the other hand, We do not want to impose any specific shape constraint for the object to track, which is especially important to perceive unknown (shape) patterns in PUI applications. Also, in the off-line training mode, lots of training samples are needed, as in the work [18].

In this paper, we propose a hybrid tracking framework, which consistently combines the blob based temporal data association and the level set based spatial occlusion handling through online shape learning, to track variable number of skin-colored objects. Blob based temporal data association can provide rapid and accurate tracking when there’s no occlusion; while level set based occlusion handling can guarantee the occluded shape recovery as precisely as possible. In such a way, the robust complete region tracking to preserve and provide more information for future high-level vision tasks is guaranteed. To the best of our knowledge, this is the first attempt to track variable number of skin-colored objects to obtain their complete regions even under (partial) occlusion.

The remainder of this paper is organized as follows. Section 2 briefly introduces some necessary background information. In section 3, we detail our proposed approach. In section 4, we present and analyze all experiments. Finally, in section 5, we conclude the paper and sketch our future directions of research.

2 Backgrounds

In this section, we briefly present some backgrounds relating to our work. To gain more insight into the object tracking and active contours models (or more specifically on level set theory), we refer the reader to some related papers.

2.1 Related Tracking Work

In general, the goal of rigid and/or non-rigid object(s) tracking is to obtain the targets between the consecutive frames in video sequences. Since the objects tracking are usually critical for video analysis, many research groups are involved in such active area and many tracking algorithms have been designed and implemented to overcome the difficulties arising from noise, occlusion, clutter and changes in the foreground and/or in the background.

Conventionally, there are several different methods to perform the temporal tracking, e.g., Kalman filter [15] and its variants, particle filtering [10] and even simple linear predictor (e.g., [1]). For a detailed review of the state of the art in objects tracking, the reader is referred to Yilmaz et al.’s recent survey paper [21].

The works most closely relating to our work are those of Argyros and Lourakis [1] and Yilmaz et al. [20]. But our work differs from such two works. In [1], without explicit occlusion handling, accurate region tracking during occlusions
only based on blob based association can not be obtained, and some even resulted in wrong labeling. While in [20], pure contour tracking is computationally very expensive and can not deal with varying number of targets.

2.2 *Level Set Methods and Active Contour Models*

Active contour models (or snakes) were first introduced by Kass et al. [14] to detect objects in images. The basic idea is to evolve a curve, subject to constraints from a given image. Thus, the problem of object detection (segmentation) can be transformed into a problem of energy minimization. In practice, the level set method [16] has been used extensively dealing with problems of curve evolution.

**Standard Level Set Formulations.** Active contours implemented via level set methods can be formulated as the zero level set of a time dependent function $\phi$ that evolves according to the evolution equation [16], called *level set equation*, as follows:

$$\frac{\partial \phi}{\partial t} + F |\nabla \phi| = 0.$$  \hspace{1cm} (1)

Based on the level set evolution framework, Caselles et al. [2] proposed the well known geodesic active contour (GAC) model as follows

$$\frac{\partial \phi}{\partial t} = g(|\nabla I|)|\nabla \phi| \left( \text{div} \left( \frac{\nabla \phi}{|\nabla \phi|} \right) + \alpha \right) + \nabla g(|\nabla I|) \nabla \phi,$$  \hspace{1cm} (2)

where $\alpha$ is a constant, and $g(|\nabla I|)$ is the edge indicator function defined as

$$g(|\nabla I|) \triangleq \frac{1}{1 + |\nabla G_{\sigma} * I|^2},$$  \hspace{1cm} (3)

and $G_{\sigma}$ is the Gaussian kernel with standard deviation $\sigma$.

Chan and Vese [3] proposed a different active contour model, unlike methods above that detect objects with edges defined by gradient, with the stopping term based on Mumford-Shah segmentation techniques.

Moreover, as to the efficient implementation of the level set, Li et al. [13] presented a variational formulation for geometric active contours that forces the level set function to approximate the signed distance function so that the level set can be evolved without the need of re-initialization.

**Integrating additional constraints into the LSE.** This introduction of shape priors into segmentation models has been proposed in Leventon [12] and has been extended and modified in a large number of successive works (e.g., Rousson [19] and Paragios [17], Cremers et al. [4,5,6]). All these methods, in common, take offline training for shape priors modeling. The most disadvantage of the offline training is to need lots of training samples to sufficiently “representing” the desired shape priors [18]. Obviously, it may does not work well when some unknown shapes occur. However, one exception is the work of Yilmaz et al. [20], where an online shape model was learned from non-rigid contour deformations during the contour tracking.
3 Proposed Approach

Our approach to tracking the complete regions of skin-colored objects include (1) adaptive skin-colored blobs extraction, (2) blob-based temporal data association, (3) on-line shape prior learning, and (4) level set based spatial occlusion handling. A flowchart of our system is shown in Fig. 1.

![Fig. 1. System overview](image)

To extract the skin-colored blobs, we filter the frame to get the skin mask by histogram based Bayesian classification of skin color. To associate each blob with the existing or new tracks, inspired by Argyros’ work [1], we enhance the original association mode. To obtain the shape prior, we propose an approach to online shape learning, where each target’s shape is described as its “contour map” after distance transforming. To detect the occlusion, we just propose a simple hierarchical decision scheme by considering the overlap relationship among two or more blobs, i.e., from the bounding area level to the “actual” blob area level determined by data association before. As for the shape recovery, we apply level set based active contour method. After the occluded shape recovery, the corresponding hypothesis of the object will be updated.

3.1 Adaptive Skin Colored Blobs Extraction

To extract the skin-colored blobs in the current frame, we first filter the frame to get skin mask. Although there are many sophisticated skin classifiers, we choose Bayesian classifier here due to its effectiveness in skin detection [11]. As to the color space, in our experiments, we obtain the similar results with RGB and HSV spaces, so we choose the RGB space here for simplicity.

Since there exists various “noise” effects (e.g., varying illumination conditions, background clutter), skin detection results may sometimes not be satisfying. Thus, we will adapt skin detection by incorporating the detection history of previous frames. Formally, the adaptive skin detection can be performed using

$$P(\omega_{\text{skin}}|x) = \alpha P(\omega_{\text{skin}}|x) + (1 - \alpha)P_{\text{history}}(\omega_{\text{skin}}|x),$$

(4)
where $P_{\text{history}}(\omega_{\text{skin}}|x)$ denotes the probability of a pixel $x$ being skin in the previous “history” frames.

After adaptive skin detection, we apply connected component analysis to form the skin blobs. Finally, we employ the morphological operations to remove small blobs or isolated holes caused by noise (See Fig. 2 for illustrative results). Once the skin-colored blobs are extracted, we can easily obtain the contour of each blob using approximate chain-code algorithm.

![Skin-Colored Blobs Extraction](image)

**Fig. 2.** Skin-Colored Blobs Extraction. From left to right: a) The input frame is first skin filtered and b) then through connected component analysis and c) morphological operations to form the skin-colored blobs, which are to be identified to specific targets.

### 3.2 Blob based Temporal Data Association

After well extracting the skin-colored blobs, one critical step for tracking is to assign the proper identity to each blob, i.e., associate the each blob with one of the existing or new tracks. Inspired by Argyros and Lourakis’ work [1], we employ similar but some different mode to implement the blob based data association.

The basic idea of the work [1] on blob-track association is to hypothesize the distribution of the object pixels to be an elliptic area (which can be called the object’s hypothesis ellipse) using ellipse fitting. Thus, the distance from any pixel to one object’s hypothesis ellipse can be used to determine whether a pixel lies within or out of the object’s area. Since there are three track-events relating to the blob-based temporal data association, i.e., object hypothesis generation, continuation and removal, it’s natural to establish the correspondence between blobs and objects hypothesis based on the spatial relationship of the one pixel against one blob hypothesis (See [1] for details). Then, the objects’ hypotheses are updated by ellipse fitting to their complete regions determined.

Different from the work [1], if occlusion occurs, we apply the process of ellipse fitting on the shape-recovered region after occlusion handling (as described in Sec.3.4), which will guide the hypothesis update more accurately and robustly in a consistent manner. As to the hypothesis prediction, Kalman filter [15] is em-
ployed to predict the center of the ellipse (hypothesis) at next instance (frame), since the simple linear prediction does not always work well.

3.3 Online Shape Prior Learning

To obtain the shape prior, we do perform online shape training. Similar to the work of modeling offline shape priors (e.g., [18]), we model the online shape prior as the signed distance function. Thus, we employ the fast distance transform [7]. Some illustrative results of distance transform are shown in Fig. 3.

![Fig. 3. (Signed) Distance transformation with object’s contour superimposed. Note: The intensity has been rescaled and inversed for better visibility.](image)

After distance transforming, we model the online Gaussian shape prior (i.e., in the form of the signed distance function) as follows:

\[
\mu^* = \rho \times \text{Dist}_{idx} + (1 - \rho) \times \mu, \quad (5)
\]

\[
\sigma^*^2 = \rho \times (\text{Dist}_{idx} - \mu^*)^2 + (1 - \rho) \times \sigma^2. \quad (6)
\]

where \(\rho\) is the learning rate; \(\mu\) and \(\sigma\) denote the Gaussian model parameters (in the form of SDF value) before update; \(\mu^*\) and \(\sigma^*\) represent the updated ones. Fig. 4 illustrates the map of the online shape prior models.

3.4 Level Set based Spatial Occlusion Handling

To deal with occlusion, we apply the active contours method. That is, we first detect occlusion after rough blob(s)-object correspondence. And then, if occlusion occurs, we employ a level set method to recover the occluded shape. Finally, after shape recovery, the corresponding hypothesis of the object also will be updated (Sec. 3.2).

**Occlusion Detection.** After obtaining the blobs-object correspondence, we can easily determine whether there exists occlusion. Here, we just propose a simple hierarchical decision scheme by considering the overlap relationship among two or more blobs, i.e., from the bounding area level to the “actual” blob area level determined by data association before. The following is the pseudo-code:
1) Consider whether two bounding boxes (corresponding to two objects respectively) overlaps. If overlaps, goto 2); if not, goto 3).
2) Refine the decision in object region level.
   2-1) For each pixel $P_i$ belonging to object $I$
   2-2) For each pixel $P_j$ belonging to object $j$
   2-3) if ($P_i(x,y) == P_j(x,y)$) there’s an occlusion, and goto 4)
   2-4) else continue the loop.
3) There’s no occlusion.
4) Determinate the decision process.

When occlusion occurs, we recover the occluded shape within the specific area, i.e., the bounding box area determined by the tracked object’s elliptic area of the previous frame up-scaled with a constant (e.g., 1.5 in our experiment).

**Occluded Shape Recovery by Incorporating Shape Prior to Level-Set Evolution.** There are different level set formulations (e.g., [2,3]). In our work, we borrow the idea of the level set evolution without re-initialization proposed by Li et al. [13] and extend it by incorporating the learned online shape prior to minimize the energy for occluded part of the tracked object. Our goal here is to pull the contour towards a preferred shape (as assumed the “true” shape even under occlusion) mostly determined by the online learned shape prior. Also, employing shape prior can speed up the level set evolution.

We can simply model the shape prior (i.e., signed distance value of each pixel) using single Gaussian kernel:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right),$$  \hspace{1cm} (7)
Thus, we can define the shape energy $E_{\text{shape}}$ using the negative logarithm of the probability defined in (7) (and omitting some constants) as follows:

$$E_{\text{shape}}(\phi) = \int\int_{\Omega} \left( \frac{(\phi - \mu)^2}{2\sigma^2} + \log \sigma \right) \, dx \, dy,$$

with $\phi : \Omega \rightarrow \mathbb{R}$ is the level set function (i.e., the signed distance function of each pixel $(x, y)$) and $\Omega$ denotes the image domain.

When the shape energy is incorporated into the level-set evolution, similar to [13], we can obtain the total energy definition $E(\phi)$ as

$$E(\phi) = \alpha E_{dp}(\phi) + E_{\text{image}}(\phi) + \gamma E_{\text{shape}}(\phi),$$

with

$$E_{dp}(\phi) = \int\int_{\Omega} \frac{1}{2} \left( |\nabla \phi| - 1 \right)^2 \, dx \, dy,$$

$$E_{\text{image}}(\phi) = \lambda \int\int_{\Omega} g \delta(\phi) |\nabla \phi| \, dx \, dy + \nu \int\int_{\Omega} g H(-\phi) \, dx \, dy,$$

where $g$ is the edge indicator function as defined in (3); $\alpha$, $\lambda$ and $\nu$ are (positive) constants to weight the internal and image energies [13]; $\delta(\phi)$ is the univariate Dirac function and $H(\phi)$ is the Heaviside function as defined in [13].

As is known, the function $\phi$ that minimizes the functional $E(\phi)$ corresponds to the Euler-Lagrange equation $\frac{\partial E}{\partial \phi} = 0$. Thus, the level set evolution to minimize $E(\phi)$ (i.e., updating $\phi$ along artificial marching time $t$) through the steepest gradient descent can be derived as

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E}{\partial \phi} = -\left\{ \alpha \frac{\partial E_{dp}(\phi)}{\partial \phi} + \frac{\partial E_{\text{image}}(\phi)}{\partial \phi} + \gamma \frac{\partial E_{\text{shape}}(\phi)}{\partial \phi} \right\},$$

where $\gamma$ confines the relative weight of the shape prior. That is,

$$\frac{\partial \phi}{\partial t} = \alpha \left[ \Delta \phi - \text{div} \left( \frac{\nabla \phi}{|\nabla \phi|} \right) \right] + \left\{ \lambda \delta(\phi) \text{div} \left( g \frac{\nabla \phi}{|\nabla \phi|} \right) + \nu g \delta(\phi) \right\} - \gamma \frac{(\phi - \mu)}{\sigma^2},$$

(12')

During the shape prior incorporated level set evolution, there are different strengths between image energy and shape prior energy, i.e., the last two terms in (12'). Therefore, we modify (12') to

$$\frac{\partial \phi}{\partial t} = \alpha \left[ \Delta \phi - \text{div} \left( \frac{\nabla \phi}{|\nabla \phi|} \right) \right] + \beta \left\{ \lambda \delta(\phi) \text{div} \left( g \frac{\nabla \phi}{|\nabla \phi|} \right) + \nu g \delta(\phi) \right\} - (1 - \beta) \left\{ \gamma \frac{(\phi - \mu)}{\sigma^2} \right\},$$

(13)

where $\beta$ is a weight introduced to balance the strengths between image energy and shape prior energy.

Finally, we obtain the following update equation of the level set function:

$$\phi^{t+1} = \phi^t + \tau \frac{\partial \phi}{\partial t},$$

(14)

where $\tau$ is the time step (which always takes 5.0 in all our experiments) and $t$ is the artificial time marching parameter.
4 Experimental Results

All our experiments were performed on a PC with Intel Pentium-IV 3.2GHz CPU, 1GB RAM. Our own test video sequences are captured by Sony DCR-PC120E in 25fps, where the image size is 320x240. And we have implemented all algorithms with non-optimized C++ code.

The proposed method has been tested on several video sequences. First, we compare our tracking performance with that of Argyros’ work [1] on his original test video (totally 3720 frames, omitting the last 106 meaningless frames). Then, we evaluate the method on the self-captured sequences “taiji” (totally 2929 frames) and “pui-sign” (totally 2495 frames) with complex motion patterns.

The system produces good tracking results of about 8∼12 fps with the frame size of 320x240. Here, the tracking speed also depends on the number and size of the targets existed in the field of view. Unlike the pure contours based multiple objects tracking, our hybrid tracking framework adaptively “select” the blob and contour processing, which makes the system more computationally efficient. Representative results to deal with occlusion are provided in Fig.5 and Fig.6. And some typical results on our self-captured video sequence “taiji” are shown in Fig.7.

It should be pointed out that one limitation of the continuous online Gaussian shape prior updating based occlusion handling is that the frame rate of capturing original video should be high enough (e.g., 25 fps or more) to keep the motion of the target smooth, or target’s motion should not be too rapid. And another limitation is that the current approach has not yet coped with the complete occlusion well.

![Fig. 5.](image-url) Comparision of the tracking result between our method and Argyros’ method on Argyros’ ECCV’04 sequence: (a) Argyros’ method (Result image is extracted from Argyros’ result demo video), (b) our method (Top-left corner of the image above show three skin-colored objects extracted and white bounding boxes indicates the area of occlusion handling).
Fig. 6. Typical occlusion handling result of our method on sequence “taiji” taking the frame 255 (a) and frame 271 (b) as examples (Top-left corner of the image above show three skin-colored objects extracted and white bounding boxes indicates the area of occlusion handling).

5 Conclusions and Perspectives

We have presented an approach to tracking the complete regions of variable number of skin-colored objects in the context of perceptual user interface. To deal with occlusion, we incorporate the online learnt shape priors into level set evolution to recover the occluded shapes. Our system is able to track long video sequences robustly and produces good results at about 10 fps with the frame size of 320x240, as demonstrated by our experiments performed on videos.

As for future work, we should discriminate the identity of right/left hand for future higher level vision tasks. And we will design more appropriate shape learning approach to address the occlusion when with the case of low capture frame rate or rapid target motion. Further, based on such results, we will focus on vision based sign large recognition (especially on medium or large vocabulary).

Acknowledgments. This work was partially supported by Natural Science Foundation of China under contract No.60533030 and No.60603023. The authors thank Dr. Chunming Li for his generous help on the level set method, Prof. Argyros for his test video data, and Ms. Yan Ma for her help in data capturing.

References

5. Cremers, D. Nonlinear dynamical shape priors for level set segmentation. In CVPR'07. 3
Fig. 7. Tracked skin-colored objects in self-captured video sequence “taiji”. The target areas being tracked are extracted and shown at the top-left corner (with scaling to the same rectangle area). Viewing the results above in color is recommended.