DYNAMIC WEIGHTING BASED ACTIVE CURVE PROPAGATION METHOD FOR VIDEO OBJECT SELECTION

Marwen Nouri[1,2], Emmanuel Marilly[1], Olivier Martinot[1] and Nicole Vincent[2]
[1]*Alcatel-Lucent Bell Labs, Paris, France*
[2]*LIPADE, Paris Descartes University, Paris, France*

Keywords: Interactive System, Video Object Selection, Scribbles based Segmentation, Scribbles Propagation.

Abstract: Improving video user experience is an essential task allowing video based algorithms and systems to be more user-friendly. This paper addresses the problem of video object selection by introducing a new interactive framework based on the minimization of the Active Curve energy. Prior assumption and supervised learning can be used to segment images using both color and morphological information. To deal with the segmentation of arbitrary high level object, user interaction is needed to avoid the semantic gap. Hard constraints such scribbles can be drown by user on the first video frame, to roughly mark the object of interest, and there are then automatically propagated to designate the same object in the remainder of the sequence. The resulting scribbles can be used as hard constraints to achieve the whole segmentation process. The active curve model is adapted and new forces are included to govern the curves evolution frame by frame. A spatiotemporal optimization is used to ensure a coherent propagation. To avoid weight definition problem, as in classical active curve based algorithms, a new concept of dynamically adjusted weighting is introduced in order to improve the robustness of our curve propagation.

1 INTRODUCTION

In the context of immersive video experience, some of the key questions are: What can immersion bring to communication, entertainment or human machine interaction? How to introduce immersion? One important research track related to those general questions is: how to better understand video in real-time? The main issue here is the semantic gap issue, that is, the way to transform low level description of the video (computer vision, signal processing) into high level understanding. In the case of object segmentation, the process of separating an image into foreground and background regions using a hard binary labeling, which can be also extended by finding a smooth alpha channel, known as image matting, some methods (Joshi et al., 2006) or (McGuire et al., 2005) introduce extra information coming hardware, such camera array, multi-focus imaging …, or also from learning technics to fill this semantic gap. The problem with these methods is their lack of ability to handle various types of object. When the objects are not rigid, the learning becomes difficult; the problem of object representation has to be addressed. Other approaches introduce interactive or user guided algorithms to take advantage of prior information on what or how the object is. This allows to deal with topology changes and to remove confusions. In recent years, a big progress has been achieved on interactive object segmentation and matting methods in the case of still images, which is well described in (Wang and Cohen, 2007). Two types of user interaction have been involved: the scribbles and silhouette. The scribbles are strokes placed roughly by the user to indicate the background and the foreground (Boykov et al., 2001). The silhouette is a coarsely tracing of the object's boundary which allows the construction of a trimap which is a three part image partitioning (foreground, background and unknown region) (Chuang et al., 2001.). High quality results can be achieved for fairly complex images. However, when dealing with videos, the user interaction is more difficult to get. (Wang et al., 2005) introduced a new way to let user make scribbles directly on 3D temporal video volume, this kind of interaction is not natural because visualizing and understanding information on video volume is not easy. In (Bai and Sapiro, 2007) the user is asked to act on different key frames. The choice of key frames is important

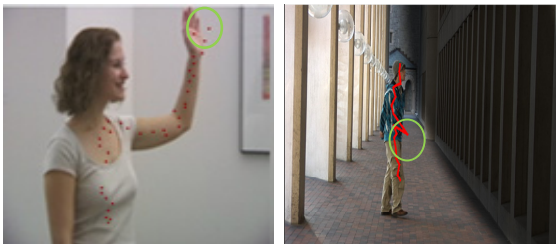
Nouri M., Marilly E., Martinot O. and Vincent N..
DYNAMIC WEIGHTING BASED ACTIVE CURVE PROPAGATION METHOD FOR VIDEO OBJECT SELECTION.
DOI: 10.5220/0003818500050011
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 5-11
ISBN: 978-989-8565-03-7

and often depends on the video. A complete segmentation of the object of interest is achieved on these frames, and then this segmentation is interpolated and propagated to extract the video objet.

(a) (b)

Figure 1: (a) Example of a scribble drawn by the user. (b) A set of curve's points representing discretized scribble.

This paper focuses on minimal user interactions that could enable video object segmentation and matting. We proposed an interactive method, which aims at tracking a moving object and designate it through video frames. The user is asked to designate an object of interest by drawing scribbles (or curves). Then, the problem is to propagate these scribbles in next frames while designating the same object. This allows to get user's hints in all video frames while reducing the user effort. We are taking the general problem when deformable object, moving in a non-constant environment and the camera is also a moving camera. Section 2 describes the curve propagation problem and the proposed active curve modeling. Section 3 presents and discusses the different forces we defined to govern the propagation, dynamic weighting is introduced and described in section 4. Experiments and results are presented in section 5.



(a) (b)

Figure 2: (a)(b) Optical flow based propagation in frame 10 from 'Amira sequence' and in frame 5 from 'walking man sequence'. Errors are indicated by the green circles.

2 SCRIBBLES PROPAGATION

Our method aims at propagating scribbles drawn on a video frame to the remainder of the sequence. For this purpose we chose an active curve modeling. First, we will explain the characteristics and the problems of propagating a curve drawn manually by the user to designate an object. Then, we will expose the active curve method, its properties and, based on it, how our problem is solved.

2.1 Curve Characteristics and Propagation

What are the characteristics and the functionality of the curve we want to propagate? This question is implicit to the user who wants to select specific regions to point out an object. What are the properties enabling to propagate the curve while preserving its functionality? This curve crosses several regions that compose the selected object. The curve is usually located at the middle portion of these regions as shown in Figure 1. User's drawing is usually coarse; security leads him to stay away from the object's edges.

Propagating a hand drawn curve along the different video frames consists in tracking a set of points. Standard point based motion estimation methods, such as (Baker and Matthews, 2004), do not allow a correct propagation of the curve's points through a large number of video frames. Indeed, to be correctly tracked, points must be illegible to criteria that can be found in interest points. This is the case of point of interest presented in (Shi and Tomasi, 1994). Moreover, due to the initial curve's properties (the way and where the drawing is done), tracking errors will be propagated and accumulated progressively frame after frame. This will decrease the tracking or the propagation quality. Point based tracking is sensitive to texture similarity or aperture problems.

In fact, the curve's points are not geometrically independent, as in some way they stay adjacent points along the video. Processing points independently can quickly become incoherent. We suggest processing all the points as a whole, as a curve, with a one dimension parameterization.

To point out the same object in next video frames, the curve must stay in the selected regions and must be located in the middle of these areas in order to reduce the drift. It is dangerous to move toward the crossed areas boundaries. In Figure 3, C_2 is more representative than C_1. Whatever the applied shifting C_2 (video temporal coherence i.e. little

shifting) is less sensitive to get away from the crossed region.

From these remarks, three constraints have been identified, allowing the curve propagation, they are stated as follows:

- If the selected object moves, the curve has to move accordingly to the movement of the different regions of the object.
- The characteristics of crossed regions must be conserved; however, the evolution of these regions must be taken into account as the environment may change along the video.
- In order to minimize errors, the curve has to move and to converge towards the middle of each region to give a good representation of the each one.

In our case, as the curve is manually drawn by the user, no initial assumptions can be done on the curve shape. As they are too global, some models like Bezier curves or splines don't seem flexible enough to enable curve propagation, considering our constraints. To perform a coherent propagation, we define forces and the associated energy functional to position the curve on the next frames. As several forces are involved, a dynamic balancing scheme is introduced to allow better sticking of the curve on the image data. Estimating the position of the curves in next frames is then done by an energy minimizing process based on dynamic programing (Williams and Shah, 1992). It is an active curve based approach presented here.

3 CONSTRAINTS MODELLING

To ensure a coherent propagation of scribble's points and to make them evolve while designating the same object -a set of heterogeneous areas- through video frames, we define a set of energies: internal and external energies. First, we model the forces, associated with the curve itself and then with image data, according to the defined constraints. In term of energies modeling, the global energy is composed from two terms: internal energies, related to internal forces that manage the liberty given the curve, and external energies, related to external forces that manage the environment of the curve. Internal forces are managing the intrinsic cohesion of the curve, and external forces are ensuring data attachment. Based on these forces we can define corresponding energies we want to globally minimize. Many elements may be considered such as the curvature, motion, texture, etc. Let's write the global energy as the sum of an intrinsic energy and an external one:

$$E_{global}(C) = E_{int}(C) + E_{ext}(C) \quad (1)$$

If the minimization has to be achieved in a global way, the forces are acting locally and we will present them when applied to each point.

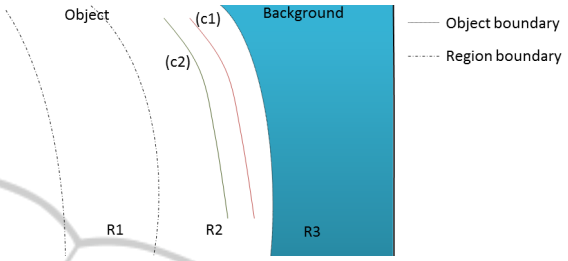


Figure 3: Curve's region representation: (C2) is less affected by object movements; (C2) is more representative to the region (r1) than (C1).

Let C be a curve discretized by N ordered points p. The global energy of curve E(C) can be estimated by:

$$\hat{E}_{global}(C) = \sum_{p \in R} E_{int}(p) + E_{ext}(p) \quad (2)$$

with $R = \{p_1, \dots, p_N\}$. The estimated solution correspond to the set R of points such that E(R) is minimum, that is to say rather more complete $(p_1, \dots, p_N)$ for which energy is minimal.

3.1 Internal Forces

Conventionally, the internal used forces are related to the curve bending. They are: uniformity and curvature forces. The uniformity force associated with its energy functional E_{unif} tries to maintain the cohesion of curve's points and therefore standardizes the distances between each pair of successive points. Notice avg_{dist} the average distance separating two successive curve's points p_i and p_{i+1}.

$$E_{unif}(p) = |avg_{dist} - \|p\, p_i\|| \quad (3)$$

The curvature energy E_{curv} influences the rigidity of the curve. Let p_{i-1}, p and p_{i+1} three successive curve's points. Let u_x the projection of p_{i-1}, p on the image's x axe and v_x the projection of p, p_{i+1} on the same axe. Let u_y and v_y the equivalent respectively of u_x and v_x projection on the image's y axe.

$$E_{curv}(p) = \left(\frac{(u_x + v_x)}{\|p_{i-1}, p\|\|p, p_{i+1}\|}\right)^2 + \left(\frac{(u_y + v_y)}{\|p_{i-1}, p\|\|p, p_{i+1}\|}\right)^2 \quad (4)$$

On each point p, the internal energy can be written as:

$$E_{internal}(p) = \omega_1 E_{unif}(p) + \omega_2 E_{curv}(p) \quad (5)$$

3.2 External Forces

The definition of external forces is more sensitive. It is directly related to the nature of the problem. To get a coherent tracking system, which selects the same object in the successive frames, we decided to model the constraints defined previously by three forces:

- One is related to the estimated motion of each point,
- Another is related to the local texture similarity,
- The last one indicates a privileged direction for propagation.

3.2.1 Motion

As no restriction has been expressed on the nature of the object, the movement of an object may be defined as the juxtaposition of different movements from different parts of it. For instance, the movement of the hand is not necessary the same as the movement of the head of some one. So, locally based **motion estimation** at each point is necessary to ensure that the overall curve is committed to the movement. Let p'_t be the optical flow [4] estimated image in frame *t+1* of a curve's point p_t from frame *t*. To adjust the curve to the movement, we look for p_{t+1} among the points *p* in frame *t+1* that have the lowest possible Euclidean distance to point p'_t. The energy to minimize is given in frame *t+1* by formula:

$$E_{motion}(p) = ||p\ p'|| \quad (6)$$

3.2.2 Texture

Between frames at time *t* and *t+1*, the point must remain in the same part of the object. The area may be **color** uniform, but it also can be characterized by a texture. Characterization of the texture would be too time-consuming, so we chose to limit the study to the average color in a neighborhood of each curve's point. To better model the temporal changes, we chose CIE Lab color space. It is composed from three components *L*, *a* and *b*. This color space allows us to better distinguish between luminance (*L*) and chromatic components (*a* and *b*) of color than in RGB space. When we deal with video, contrast change artifact is very common, so luminance is less discriminant than chromatic components. One curve's point p_t will then evolve to a new position with a similar color while including some tolerance to illumination change. Then we define color similarity energy using a weighted Euclidean distance in the Lab space, where less importance is given to the luminance than to the chrominance. The average color at point p is denoted *cm*. The average is computed in the neighborhood according the S_i segment described in the section. Three components: *cm(a), cm(b)* and *cm(l)*. Then the energy formula is:

$$E_{color}(p) = \left(cm_{pt}(a) - cm_p(a)\right)^2 + \left(cm_{pt}(b) - cm_p(b)\right)^2 + \frac{1}{4}\left(cm_{pt}(l) - cm_p(l)\right)^2 \quad (7)$$

The ¼ coefficient has been empirically set and is used in all the experiments we present.

3.2.3 Stability

The last energy comes from the fact that the user tends to draw scribbles while trying to stay well within the middle zones of the different object's regions (which prevents him from getting out of the object). The curve must also reproduce this effect. This will make the system more stable. To answer this third constraint, we first detect the locally homogeneous region around the curve, depending on the color of each of its points. The curve is extended to a confidence region R. The extension is computed from the set of p_i points. From each p_i the extension is computed in an orthogonal direction to the curve, using a region growing process. Each point is extended as a segment, denoted S_i. R is so defined as the convex hull of all S_i (Figure 4). The region growing is based on color similarity. To calculate this energy, we estimate the normal to the curve at each point p_i at the time t, and calculate a similarity line segment S_i. This segment associated with p_i is the maximum segment of uniform color. That is to say, all its points' colors are similar to the p_i color. The maximum concerns the length of the segment. No larger segment can be found with respect to p_i color. The extremities of this segment are the limit tolerance points. p_{lt} denotes the extremity which is farthest from p_i. This models the direction in which the curve is the farther from the region contour.

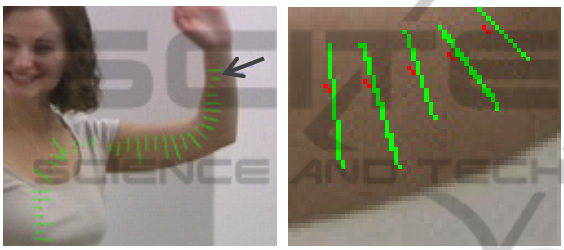
Our goal, starting from the state of a point on the curve at time *t*, is to **privilege the propagation** towards the middle of the region. This energy will push points to an area of greater homogeneity, as shown in Figure 5. To prevent curve's displacement from motion estimation errors and introduce more coherence and stability to our system, we look for

p_{t+1} among all possible position p in frame $t+1$ that have the lowest possible Euclidean distance to point p'_{lt} which is the optical flow image of p_{lt}, the point for each point p_i represents the limit of the similarity region R. The stability energy to minimize is given in by the following formula:

$$E_{Stability}(p) = \|p \; p'_{lt}\| \tag{8}$$

The external energy is composed from three weighted terms, and then we can finally write the global energy as:

$$E_{global}(C) = \sum_{p \in C} \begin{pmatrix} \omega_1 E_{unif}(p) + \omega_2 E_{curv}(p) + \\ \omega_3 E_{motion}(p) + \omega_4 E_{color}(p) + \\ \omega_5 E_{Stbility}(p) \end{pmatrix} \tag{9}$$



(a) (b)

Figure 4: (a) Similarity line segment drawn on each point of the curve. (b) Zoomed view of (a).

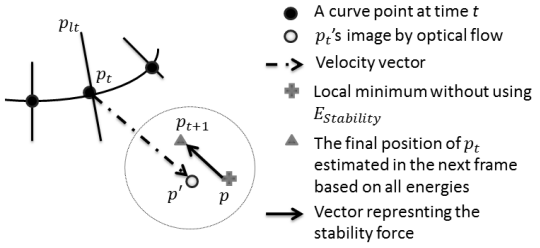


Figure 5: Stability force representation; the curve is pushed toward the middle of homogenous regions.

4 DYNAMIC WEIGHTING MECHANISM

One of the major problems in using active curves methods is how to tune the energies' related weights. There is no defined method to set them. Using different sets of weights, ones can get similar results. Usually, the weights are set from prior information or empirical estimations as in (Kass et al., 1987). In more recent studies (Etyngier et al., 2007), learning based approaches are used to adapt more robustly the weights to a pre-defined problem. Such an approach cannot be applied in our case due to the interactive nature of our application and the variety of object we want to handle. Moreover, we noted that even if one force has more importance than the others, this is not true all the time and for all points' position, so it could lead us to errors in later processed frames.

As shown above, the curve drawn, by the user, can cross many different high textured regions. A global parameterization is, so, not adapted in our case due to the lack of priors. In contrast of the active curve based previous works (Kass et al., 1987) or (Lefevre and Vincent, 2004), we introduce a dynamic weighting scheme based on the following observations: in the cases of points in regions containing many high gradient, one point can be easily tracked based on classical motion estimation methods with higher robustness. In the other cases motion estimator seems to be very errors prone and it would be better to give more importance to the others data terms. According to the position of the curve's point, for external forces, we want to dynamically change the weights to match the nature of the image, thus the area in which the point belongs to.

In the previous section, the global energy was formulated as fellow:

$$E_{global}(C) = min \sum_{1}^{n} \omega_i E_i(C) \tag{10}$$

While we consider dynamic weighting scheme, we can rewrite the global energy as follows:

$$E_{global}(C) = \sum_{p \in C} \begin{pmatrix} \omega_1(p) E_{unif}(p) + \omega_2(p) E_{curv}(p) + \\ \omega_3(p) E_{motion}(p) + \omega_4(p) E_{color}(p) + \\ \omega_5(p) E_{Stability}(p) \end{pmatrix} \tag{11}$$

We consider that the weight associated with each energy is related to the point characterization. To describe the area around the point, we have chosen the similarity segment S_i (described above) at each point of the curve. We may act individually on each point. Thus we can say, for example, the shorter the segment S_i is, the higher is the probability that the point is a contour point. Therefore we give more confidence (and thus more weight) to the estimated motion and so to E_{motion} by increasing ω_3 proportionally to the length of S_i denoted Ls_i . In the other case we want to increase the stability force effect. In practice, the lengths of all similarity segments Ls_i are normalized by the maximum length then we can write:

$$NLs_i = {}^{Ls_i}/_{max\,(Ls_i)}. \tag{12}$$

The global energy has to be rewritten according to this following formula:

$$E_{global}(C) = \sum_{p \in C} \begin{pmatrix} \omega_1(p)E_{unif}(p) + \omega_2(p)E_{curv}(p) + \\ (1 - NLs_i)\,\omega_3(p)E_{motion}(p) + \\ NLs_i\,\omega_4(p)E_{color}(p) + \\ NLs_i\,\omega_5(p)E_{Stability}(p) \end{pmatrix} \quad (13)$$

5 EXPERIMENTS AND RESULTS

Our system proposes a graphical user interface allowing user to draw one or many scribbles at any frame of a video to point out an object. Our algorithm is designed to handle dynamic backgrounds, as it is the case in Figure 6 and 8 sequences. There are two modes: an interactive mode allowing the user to designate the object of interest and an automatic mode where the scribbles are propagated automatically in a real time processing. Therefore user can interact at any time to point out a new region which was not visible at the beginning of the video Figure 8. Our final results are best seen in video form, though we show several still images for example in this section.

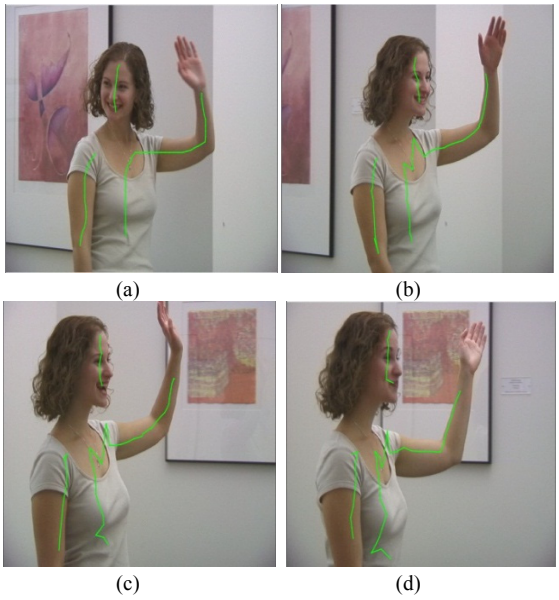


Figure 6: (a) Three scribbles drawn at frame 1 by the user to designate the women in the video. (b)(c)(d) The respective results of propagating scribbles from frame 'a' to frames 8, 23 and 30.

We tested our approach on three standard video sequences (640x480x30) from the literature (Bai et al., 2009) and (Wang and Cohen, 2005) (Figure 6, 7, 8). It is difficult to objectively measure a success of systems like our presented one. One possible approach is based on the number of frames in which the initial designated object continues to be pointed out by our algorithm. This information is got from the final user. To compare our method, we implement an optical flow based scribbles propagator, denoted OFBP. Figure 6a, 7a show two examples of scribbles drawn by user. We try to propagate these user's hints automatically to designate the object initially pointed by the user on the next video frames these scribbles can be used as input to (Levin et al., 2008) matting method applied frame by frame to extract to whole video object. OFBP approach fails even on basic cases as shown on Figure 2. Our results are shown on Figures (6, 7, 8) and compared to OFBP implementation in Table 1. We evaluate the improvement of the dynamic weighting mechanism. In the case of Amira sequence, the gain is 20%. In the Walking man sequence the gain is around 37%.

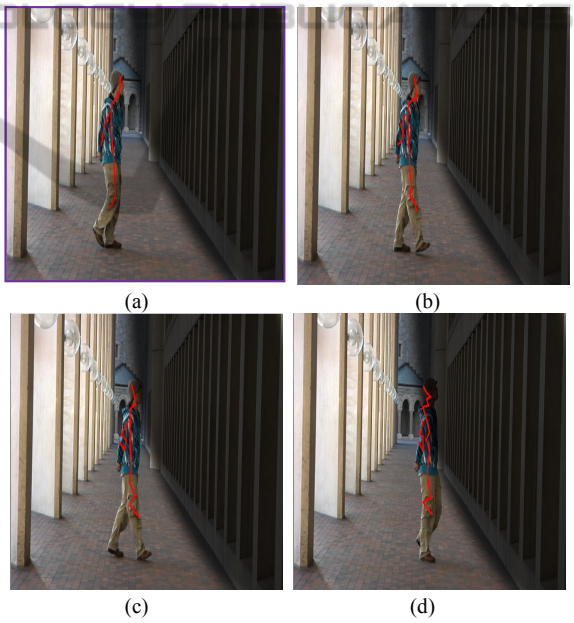


Figure 7: (a) The user points out the man by drawing two scribbles on the first video frame of walking man sequence. (b)(c)(d) The respective results of propagating the scribbles from frame (a) to frames 7, 14 and 25.

6 CONLUSIONS

In this paper, we showed a new approach for hand drawn curve propagation. It consists in using active curves model to formulate the problem of video object selection. In addition a new dynamic

weighting scheme has been introduced to let the curve stick better to the image data. Our framework allows user driven designation of objects in videos. Our contributions consist in the formulation of scribbles propagation as an active curve model and the definition of the different related energies functional combined with dynamic weights management in order to obtain a more accurate video tracking. Our algorithm can be further improved by adding features such as texture or by using a more recent and accurate optical flow estimator. We are currently focusing on this improvement and studying the potential of our approach in two fields: video matting and human actions classification.

Table 1: The number of frames in which the initial selected object is still designated.

Method \ Video	Amira (30 frames)	Adam Lib (29 frames)	Walking man (30 frames)
OFBP	11	26	5
our method without dynamic weights	24	29	14
our method	30	29	25

Figure 8: (a) The user adds a new scribble to point out a new region which was not visible in the beginning of the "Adam lib" sequence. (b) The propagation continue on based these two scribbles as shown in frame 29 (b).

REFERENCES

Bai, X. and Sapiro, G., 2007. A geodesic framework for fast interactive image and video segmentation and matting. In *Proc.of IEEE ICCV*.

Bai, X., Wang, J. Simons, D. and Sapiro, G., 2009.Video snapcut: Robust video object cutout using localized classifiers. In *SIGGRAPH 2009*, New York, NY, USA, ACM.

Baker, S. and Matthews, I., 2004. Lucas-Kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255.

Boykov, Y., Veksler, O., and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence 23*, 11, 1222–1239.

Chuang, Y.-Y., Curless, B., Salesin, D. H., and Szeliski, R., 2001. "A bayesian approach to digital matting," in *Proceedings of IEEE CVPR*, pp. 264–271.

Etyngier, P., Segonne, F., and Keriven, R., 2007. Active contour based image segmentation using machine learning techniques. in *MICCAI, ser. Lecture Notes in Computer Science*, vol. 4791. Springer, pp.891–899.

Joshi, N., Matusik, W., and Avidan, S., 2006. "Natural video matting using camera arrays," in *Proc. of ACM SIGGRAPH*, pp. 779–786, 2006.

Kass, M. Witkin, A. and Terzopoulos, D. 1987. Snakes: Active contour models. *IJCV*, 1(4):321–331.

Lefevre, S. and Vincent, N., 2004. Real time multiple object tracking based on active contours. In *International Conference on Image Analysis and Recognition*, volume 3212 of Lecture Notes in Computer Sciences, Springer, pages 606–613.

Levin, A., Lischinski, D. and Weiss, Y., 2008. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242.

McGuire, M., Matusik, W., Pfister, W., Hughes, J. F. and Durand, F., 2005, "Defocus video matting," in *Proceedings of ACM SIGGRAPH*, pp. 567–576.

Shi, J., Tomasi, C., 1994. Good features to track. *Proceedings of the IEEE CVPR*. p. 593–600.

Wang J., Bhat, P., Colburn, R. A., Agrawala, M., Cohen, M., 2005. Interactive video cutout. *SIGGRAPH'05*, 24(3):585–594.

Wang, J. and Cohen, M., 2005. "An iterative optimization approach for unified image segmentation and matting," in *Proceedings of ICCV*, pp. 936–943.

Wang, J. and Cohen, M., 2007. Image and video matting: A survey. Foundations and Trends. In *Computer Graphics and Vision* 3, 2, 97–175.

Williams, D. J. and Shah. M., 1992 A Fast Algorithm for Active Contours and Curvature Estimation, *VIGP Computer Vision Graphics Image Process Image Understanding*, vol. 55, n° 1, p. 14-26.