Journal of Information Hiding and Multimedia Signal Processing ©2014 ISSN 2073-4212
Ubiquitous International Volume **5**, Number **4**, October **2014**

Non-rigid Video Object Extraction Based on Semantic Multi-level Framework

Kun Zhao[1] , Jie Sun[2], Zhe-Ming Lu[3*], Jian Wang[4], and Li-Jian Zhou[5]

School of Communication and Electronic Engineering[1,2,5]
Qingdao Technological University
No.777 Jialingjiang Road, Qingdao, 266520, China
sterling1982@163.com[1]
sj1979419@163.com[2]
zhoulijian@qtech.edu.cn[5]

School of Aeronautics and Astronautics[3*]
Zhejiang University
No.38 Zheda Road, Hangzhou, 310027, China
zheminglu@zju.edu.cn

Cyber Physical System Research and Development Centre[4]
The Third Research Institute of Ministry of Public Security
Shanghai, 201204, China
wjconan@ieee.org

Received January, 2014; revised April, 2014
(Communicated by Zhe-Ming Lu)

Abstract. *This paper proposes a novel method to extract not only multiple objects but also interesting parts of these objects from video sequence with a semantic multi-level framework. The proposed method performs well when handling objects, especially non-rigid objects that are partial occlusive and deformed, by considering an object (rigid or non-rigid) as the combination of sub-objects (which are rigid ones at the highest level) at different semantic levels. Experimental results with different real-world video sequences demonstrate the validity and robust of our solution. The final semantic hierarchy of objects is looking forward to work effectively in video editing based on objects or parts of objects.*

Keywords: Video object extraction, Sub-object, GrabCut, Multi-level objects

1. **Introduction.** Occlusion and deformation handling are two challenges in the field of non-rigid video object extraction. Many methods[1,2]have been proposed to address these problems in the last decade. Approaches can be roughly classified into model-based approaches[1] and appearance-based approaches[2]. The former ones use priori models defined explicitly in terms of kinematics and dynamics. Their difficulties mainly comes from following parts: model initialization, fitting to image data, occlusion handling and singularity involved in inverse kinematics. However, the latter ones, using heuristic assumption on image properties such as colour, intensity, texture, edge, neighbour and so on, are quite suitable for maintaining and tracking image properties along the image sequences.

As a robust method of energy minimization, Graph Cut[3,4] has made great achievements in many vision tasks especially in image segmentation[3] and multiple objects tracking[4]. GrabCu[5], which extends Graph Cut to colour images with CGMMs (Colour

Gaussian Mixture Models), is a good appearance form combining colour, edge and area properties well. Owe to the presence of colour information, GrabCut can deal with problems such as convergence to LTIs (long and thin indentations) which troubled other models such as GVF (gradient vector flow) and GGVF (generalized gradient vector flow) based snakes[11] when using only edge information. In reference[6] , CGMMs was replaced by SCGMMs (Spatial-Colour Gaussian Mixture Models) to track multiple objects and handle occlusion. It assumes that the spatial covariance matrices of the SCGMM are constant when occlusion occurs. This assumption may not be valid in the case that the spatial layout of the colour feature undergoes severe change during occlusion.

In this paper, we propose a new method to extract not only multiple objects but also interesting parts of these objects from video sequences with a semantic multi-level framework. Through the way of taking an object, which can be rigid or non-rigid, as the combination of sub-objects at several higher semantic levels, the proposed method performs well when objects, especially non-rigid ones, are partially occlusive and deformed. Sub-objects at the highest level are regarded as rigid objects. Our method involves a two-phased approach. Firstly, we segment the key frame chosen by user from a video sequence using a modified GrabCut algorithm. Then we extract objects at each semantic level with the restriction between any of two adjoining levels based on the SCGMMs matching frame by frame. Occlusion and deformation of non-rigid objects at lower semantic levels are taken as regular movements of rigid sub-objects at higher levels. The final semantic hierarchy of objects is looking forward to work effectively in video editing based on objects or parts of objects.

The rest parts of this paper are organized as follow: two basic semantic restrictions are defined in section 2, as well as the key frame segmentation algorithm is given; the multi-level segmentation algorithm in subsequent frames is described in section 3, including occlusion and deformation handling approach; in section 4, results and discussions of experiments using proposed method are presented; at last, conclusions are obtained in section 5.

2. Multi-level Framework for Visual Object.

Visual object may contain some interesting areas as parts of it which are called *sub-objects* in[7]. In order to edit a sub-object with the restriction of its father object and other sub-objects, an image should not only be segmented as objects and background but also be taken as a semantic hierarchy. We use *father objects* and *son objects* iteratively to describe two kinds of objects at adjoining semantic levels. For example, in FIGURE 1, the doll is a son object when the whole image is taken as a father object. And his head is a son object while the doll is its father object.

2.1. Basic Semantic Relationships of Multi-level Framework: Inclusion and Exclusion between Objects.

Basic elements of our multi-level framework are obtained with this *father and son* conception iteratively. In[8], four basic semantic relationships are proposed to establish the whole framework which is inspired by[9]. The main difference between *composition* and *inclusion* (also *linking* and *exclusion*) is whether the two objects have a common contour. These contours may be invisible in some frames in video sequence. So the distinguish of *composition* and *inclusion* (also *linking* and *exclusion*) is meaningless. In this paper, four basic semantic relationships are combined into two relationships, namely *inclusion* and *exclusion.*

Definition 2.1. ***Inclusion****: O_s must be inside of O_f, perhaps with repulsion force between boundaries.*

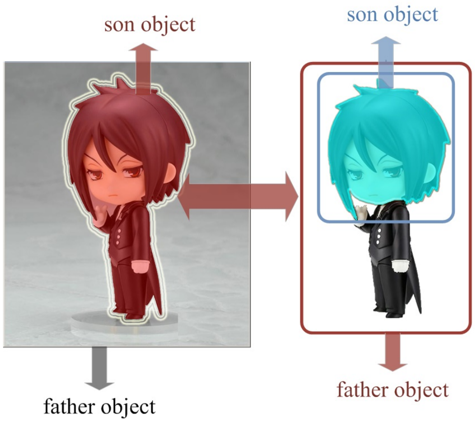


Figure 1. Father object and son object

Definition 2.2. ***Exclusion**: O_{s1} and O_{s2} cannot overlap at any pixel, perhaps with repulsion force between boundaries.*

In the definitions above, O_f denotes the *father object*, O_s denotes the *son object* and O_{si} *,i=1,2,3...*, denotes different son objects at the same semantic level, which are shown in FIGURE 2.

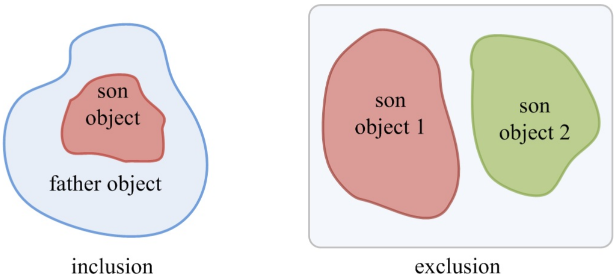


Figure 2. Inclusion and exclusion

The two basic relationships defined above impose restrictions on objects at adjoining semantic levels and objects at the same semantic level respectively. Once the user confirms the semantic levels of all objects, the relationships generate.

2.2. Segmentation Algorithm for The Key Frame.

We propose a twice-segmentation algorithm to segment the key frame of a video sequence and obtain intermediate results including interesting objects, their semantic levels and relationships. These intermediate results will be taken as reference information for segmentation in subsequent frames. In the first step of proposed twice-segmentation approach, user can locate the region of interesting objects in the rough and designate the semantic level for each object using our interactive interface. Then the system extracts the objects separately and analyses relationships between them by geometric information and semantic levels of them. Those relationships feed back and change the final result in the second step namely re-segmentation. FIGURE 3 shows the framework of our algorithm.

User Interaction and the First Segmentation

The first segmentation is based on GrabCut algorithm[5] which could deal with colour images. We improve Justin F. Talbot s interactive interface and make it possible to process multiple objects simultaneously. Besides of the interface, we also improve GrabCut

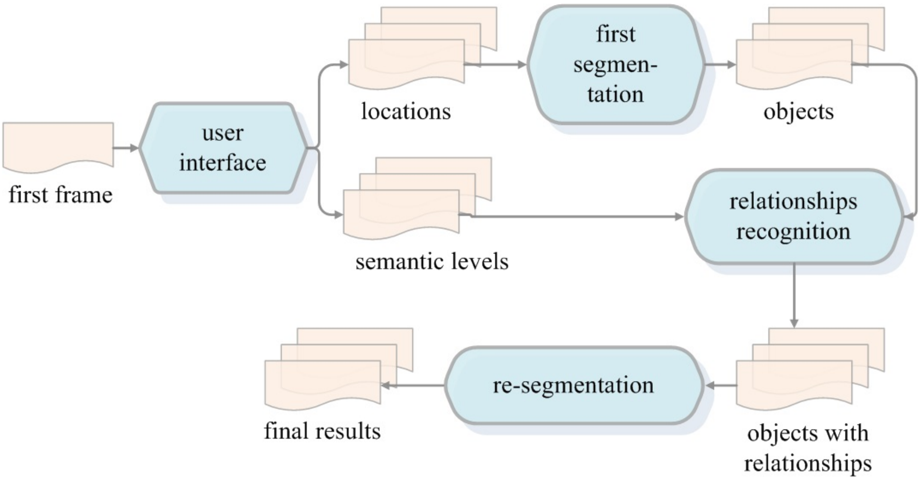


Figure 3. Framework of our twice-segmentation algorithm for the key frame of a video sequence

algorithm through a process of pre-clustering for the colour component of CGMMs. During the process of pre-clustering, a CLPSO algorithm is used to reduce the colour options from 768 RGB colours to 30 human perception-based colours in HSL space[12]. Owe to the reduction of initial colour options, the process of pre-clustering improves the computational efficiency of GrabCut algorithm greatly. With our interactive interface, users can label interesting objects in different semantic levels with different colour rectangles. The result of each object after the first segmentation is saved in a single layer. The user interactive interface is shown in FIGURE 4 where SL denotes the number of semantic levels and OL denotes the number of objects also the number of layers.

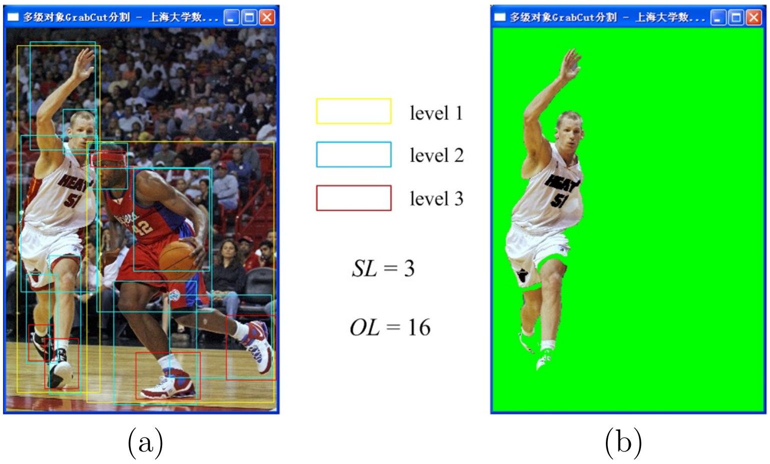


Figure 4. User input and the result after first segmentation. (a) shows our multi-level user input: interesting objects at semantic level 1 to 3 are located within yellow, blue and red rectangles respectively. (b) shows the result of object 1 at level 1 after the first segmentation.

User's input actually indicates the hypotaxis between an object and its sub-objects. FIGURE 5 shows this kind of relationship with a layer structure of input in FIGURE 4. Each sheet denotes a layer that saves one sub-object. The difference of size and colour indicates the semantic level of layers. Different groups with the biggest sheets as the base level denote different objects and their sub-objects at different semantic levels.

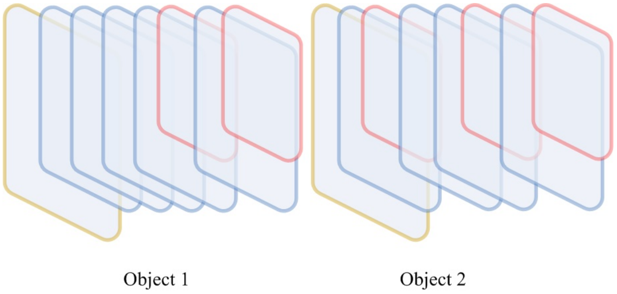


Figure 5. Layer structure of input in FIGURE 4.

Re-segmentation Based on Semantic Relationships

The purpose of re-segmentation is to refine the result by correcting mis-segmentations that may happen in the first segmentation. To achieve this goal, a new category of energy term S^{lm}, which depends on the relationships between objects, is added to energy function (Equation(12) in section 3). The details of our energy function will be discussed in next section. Here we just show how S^{lm} works with different relationships in TABLE 1.

In *inclusion*, l denotes a *father object layer* and m denotes a *son object layer*. From input frame in FIGURE 4, the contrast between the results, whether re-segmented or not, is clearly shown in FIGURE 6. One can easily find the refinement in the areas of feet.

Table 1. Energy Terms Corresponding to Different Semantic Relationships

Energy Term	**Semantic Relationships of Objects**							
	Inclusion				*Exclusion*			
f_a^l	0	0	1	1	0	0	1	1
f_b^m	0	1	0	1	0	1	0	1
S_{ab}^{lm}	0	∞	0	0	0	0	0	∞

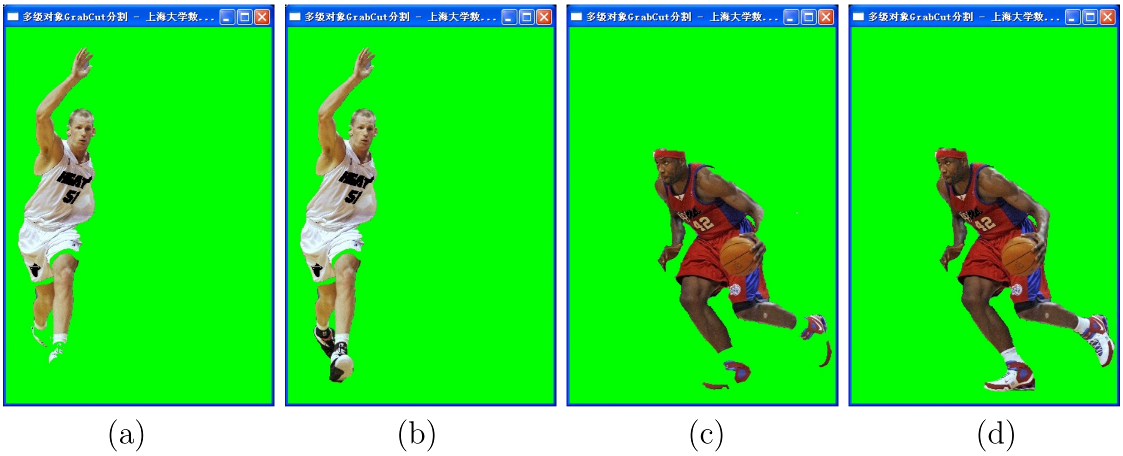

多级对象GrabCut分割 － 上海大学数...
多级对象GrabCut分割 － 上海大学数...
多级对象GrabCut分割 － 上海大学数...
多级对象GrabCut分割 － 上海大学数...
(a)
(b)
(c)
(d)

Figure 6. Results of two objects both at level 1 after (a), (c): first segmentation and (b), (d): re-segmentation.

3. **Objects Tracking and Segmentation in Subsequent Frames.** From the result of segmentation in the key frame, some noticeable parts and objects can be obtained at different levels. Each object is saved in a single layer. Hence, in a layer there is only

one object and the rest pixels are all labelled as background. In subsequent frames, these objects are tracked with their SCGMMs frame by frame and extracted with a modified GrabCut algorithm. Their semantic relationships do not keep the parts belonging to new ones or other objects but their father objects, which reduces mis-recognition and mis-segmentation when occlusion and deformation happen.

3.1. **Objects Tracking with SCGMMs.** We classify the pixels of each frame into K components. Noticing that the K components are not equal to K objects in a frame. Actually some components among K ones compose one object. Hence, the sum of objects O is always less than K. Then a SCGMM for each component is learned with the EM algorithm. With the location and the colour (after pre-clustering by the algorithm proposed in reference [12]) of a pixel as its joint feature, we get $\mathbf{z} \equiv (\mathbf{z}^S, \mathbf{z}^C) = (x, y, c)$. In frame t, the EM algorithm repeats the following steps until convergence to make all pixels labelled by K final spatial-colour Gaussian components. Let i=*1*, *2*, *3*, ...I be the number of iteration. Two steps of EM algorithm are as follows.

E-step: calculate the probability p of each pixel with its feature vector $\mathbf{z}$ belonging to Gaussian component K in the ith iteration.

$$p^i_{\mathbf{z},k} = p(\mathbf{z}|k)^i = C \cdot |\sigma^i_k|^{-1/2} exp(-1/2(\mathbf{z} - \mu^i_k)^T (\sigma^i_k)^{-1} (\mathbf{z} - \mu^i_k)) \tag{1}$$

C is a constant, μ^i_k is the mean of the kth Gaussian component of the SCGMM during the ith iteration, similarly, σ^i_k is the covariance of it. On the assumption that the spatial and colour features are independent of each other, we have Equation(2) and Equation(3) as below. Then Equation(4) to Equation(6) can be got easily.

$$\mu^i_k = (\mu^{S,i}_k, \mu^{C,i}_k) \tag{2}$$

$$\sigma^i_k = \begin{pmatrix} \sigma^{S,i}_k & 0 \\ 0 & \sigma^{C,i}_k \end{pmatrix} \tag{3}$$

$$p^i_{\mathbf{z},k} = p^{S,i}_{\mathbf{z},k} \cdot p^{C,i}_{\mathbf{z},k} \tag{4}$$

$$p^{S,i}_{\mathbf{z},k} = C_1 \cdot |\sigma^{S,i}_k|^{-1/2} exp(-1/2(\mathbf{z}^S - \mu^{S,i}_k)^T (\sigma^{S,i}_k)^{-1} (\mathbf{z}^S - \mu^{S,i}_k)) \tag{5}$$

$$p^{C,i}_{\mathbf{z},k} = C_2 \cdot |\sigma^{C,i}_k|^{-1/2} exp(-1/2(\mathbf{z}^C - \mu^{C,i}_k)^T (\sigma^{C,i}_k)^{-1} (\mathbf{z}^C - \mu^{C,i}_k)) \tag{6}$$

M-step: refine the parameters with the calculated probabilities (Equation(7) to Equation(10)), then back to E-step in the next iteration with new parameters.

$$\mu^{S,i+1}_k = \sum_{\mathbf{z} \in k} p^{S,i}_{\mathbf{z},k} \mathbf{z}^S \tag{7}$$

$$\mu^{C,i+1}_k = \sum_{\mathbf{z} \in k} p^{C,i}_{\mathbf{z},k} \mathbf{z}^C \tag{8}$$

$$\sigma^{S,i+1}_k = \sum_{\mathbf{z} \in k} p^{S,i}_{\mathbf{z},k} (\mathbf{z}^S - \mu^{S,i+1}_k)(\mathbf{z}^S - \mu^{S,i+1}_k)^T \tag{9}$$

$$\sigma^{C,i+1}_k = \sum_{\mathbf{z} \in k} p^{C,i}_{\mathbf{z},k} (\mathbf{z}^C - \mu^{C,i+1}_k)(\mathbf{z}^C - \mu^{C,i+1}_k)^T \tag{10}$$

We set μ^I_k and σ^I_k to be the final parameters of each SCGMM of the tth frame when it is convergent during the Ith iteration. Then these SCGMMs are delivered to the next frame and become the initial SCGMMs of the *I+1*th frame. From the result of EM steps mentioned above, the final SCGMMs will be easily found in each frame. K SCGMMs are tracked frame by frame.

Owe to our semantic multi-level framework, the system can track objects in different layers simultaneously with tracking their SCGMMs which are obtained when segmenting the key frame. SCGMMs of the key frame provides reference information when some objects missing in some subsequent frames. The system can keep on tracking them by the reference information when they appear again.

3.2. **Occlusion and Deformation Handling.** The proposed method implies three approaches to handle the situation of occlusion and deformation.

Firstly, the proposed appearance model uses not only the colour feature of pixels but also the spatial distribution of these colour blobs which make the objects compacter and easier to track.

Secondly, the proposed semantic multi-level framework presents a strong connection between sub-objects at different semantic levels and their father objects. Partial occlusion is considered as a kind of disappearance-appearance behaviour of sub-objects at high levels. The system can keep tracking them with the connection to their father objects. Contrarily, when deformation happens in non-rigid father objects, their rigid son objects can be tracked well and help system recognizing them effectively through the connection.

Thirdly, segmentation results and the SCGMMs of the key frame can be taken as reference any time when needed, especially in the case of whole occlusion. Because it is chosen by users from the video sequence, which shows all interesting parts of objects clearly, the key frame is the best choice as the reference frame.

3.3. **Multi-level Object Segmentation Algorithm.** The segmentation algorithm we use in subsequent frames is very similar to the one in the key frame. We will discuss it in detail in this section.

We still use a twice-segmentation algorithm to extract and refine objects saved in different layers. Different from the key frame segmentation, system can segment following frames automatically, rather than with the participation of user input.

Actually we get two kinds of important information from the intermediate results of key frame segmentation. One is the SCGMMs of the object and background in each layer, the other is the semantic level and hypotaxis of the layers. The former one can be taken as the initial mask for objects in each layer and help extracting them during the first segmentation. Well the later one affects the third category of energy term S^{lm} during the step of re-segmentation which refine the result by the rules described in TABLE 1.

As being discussed in [3], energy-based object extraction could obtain results which are more semantic and robust. We chose the form of energy function mentioned in [3] as the foundation of our energy function. In [3] a kind of energy function called *discontinuity preserving* was proposed which is composed of two components: data term and regularization term. A classical form of this kind of energy function is shown as below.

$$E(f) = \sum_{a \in P} D_a(f_a) + \sum_{a,b \in N} V_{ab}(f_a, f_b) \tag{11}$$

Graph Cut [3, 4] used the form above and obtained satisfactory result when extracting a single object from an image. When extracting objects in different semantic levels, we add a new category of energy term to describe the interactions based on two basic semantic relationships that we defined in section 2.1. Our energy function is shown as follow:

$$E(f) = \sum_{a \in P} D_a(f_a) + \sum_{l \in O} V^l(f^l) + \sum_{l,m \in O}^{l \neq m} S^{lm}(f^l, f^m) \tag{12}$$

$$V^l(f^l) = \sum_{a,b \in N^l} V^l_{ab}(f^l_a, f^l_b) \tag{13}$$

$$S^{lm}(f^l, f^m) = \sum_{a,b \in N^{l,m}} S^{lm}_{ab}(f^l_a, f^m_b) \tag{14}$$

In Equation(12), we define P as the set of pixel indices and O as the set of object indices. The binary variables are $f \in \mathbb{B}^{O \times P}$ in which we index f^l_a as the variable over pixels $a \in P$ and objects $l \in O$. We interpret f^l_a*=1* to mean that pixel a belonging to object l which is saved in the lth layer. The notation f_a denotes a vector of all variables that correspond to pixel a.

The proposed multi-level framework saves each object in a single layer, which simplifies a multi-label issue into a simple binary segmentation. Hence, there are only two objects in each layer, namely the foreground and the background. In the former two terms of our energy function, we set *l=0* to denote the background and *l=1* to be the foreground in each layer.For the first term $D_a(f_a)$ of background in the lth layer, tth frame, we have:

$$D^{l,t}_a(f^0_a) = -log \sum_{k \in bkg^{l,t-1}} p^I_{\mathbf{z}_a,k} \tag{15}$$

$p^I_{\mathbf{z}_a,k}$ is calculated by Equation(1) where I is the times of iteration that converge the function. All the probabilities of the pixel belonging to the components which are labelled as background in the lth layer, *t-1*th frame are also be computed.

Similarly, for $D_a(f_a)$ of foreground in the lth layer, tth frame, we have:

$$D^{l,t}_a(f^1_a) = -log \sum_{k \in fg^{l,t-1}} p^I_{\mathbf{z}_a,k} \tag{16}$$

The second term $V_{ab}(f_a, f_b)$ in the lth layer, tth frame denotes the N-link weights between pixel a and a pixel in its 26-neighbourhood, b. The expression of $V_{ab}(f_a, f_b)$ is as follow.

$$V_{ab}(f_a, f_b) = (\gamma / dist(a,b)) exp(- \left\| \mathbf{z}^C_a - \mathbf{z}^C_b \right\|^2 / (2\sigma^2)) \tag{17}$$

$$\sigma^2 = (1/n) \sum_{all(a,b) \in N^l} \left\| \mathbf{z}^C_a - \mathbf{z}^C_b \right\| \tag{18}$$

In Equation (17) and Equation (18) above, *dist(a,b)* is the pixel distance, $\mathbf{z}^C_a$ is the colour vector of pixel a, σ^2 is the average colour-difference over all pairs of pixels *(a,b)* neighbouring and n is the number of pairs in the lth layer. This smoothness term penalizes the labelling discontinuities of neighbouring pixels if they have similar colour. We set γ*=50* with suggestion in [10] and take place of the *8-neighbourhood* model with a *26-neighbourhood* model by adding the pixels in the *t-1*th and *t+1*th frame.

Objects in each layer are extracted by using a modified GrabCut algorithm (similar as the algorithm been used in key frame, but without user input) with their SCGMMs. Then, a process of re-segmentation is executed by using the term S^{lm}.

The third term in the proposed energy function is used between two objects which be put into different layers such as l and m. The mis-segmentation punishment term

$S_{ab}^{lm}(f_a^l, f_b^m)$ is set to be a binary term whose weight is either *0* or ∞. That means pixels have to be re-labelled in the opposite way in each layer once they are regarded as mis-segmentation ones. The proposed method judges the mis-segmentation by the two semantic relationships defined in section 2.1. How the term $S_{ab}^{lm}(f_a^l, f_b^m)$ works has been shown in TABLE 1.

By the proposed twice-segmentation algorithm, four kinds of information are obtained, which are the extraction result in each layer, the SCGMMs of each object, the semantic level of each object and the relationships between objects. These information spread to the next frame and help extracting objects frame by frame. At last, all frames in the video sequence are segmented.

4. **Experimental Results.** To demonstrate the capability of the proposed method in extracting and tracking multiple objects through occlusions, experimental results on several video sequences are reported in this section. In all the experiments, the image resolution of sequences is 4CIF (704 × 576). The parameters *K=8*, *C=0.01* and *I=5* are chosen. The proposed method is implemented on C++ platform with a 2.0G dual processor PC.

User input on the key frame, the SCGMMs of interesting objects at level 1 and the result of key frame has been shown in FIGURE 7. We suggest to choose the frames in which interesting objects are shown completely and clearly in order to get obvious features of each sub-object. In addition to this, there should be no occlusion or disappearance of the interesting objects in several frames (about 5 to 10) after the key frame for getting better SCGMMs for each sub-object.

Spread by the key frame, results of subsequent frames are obtained and have been shown in FIGURE 8. (b) and (e) of FIGURE 8 show wrong segmentation when interaction happens between the arms of the couple with only SCGMMs algorithm. The proposed method amends the results by tracking the arm and hand of the gentleman and the bright circles on the left arm of the lady as the rigid sub-objects at level 2 and level 3. The results have been shown in (c) and (f) of FIGURE 8.

FIGURE 9 shows the extraction and tracking results of objects at level 1 in subsequent frames with proposed method. Rectangles that generated by the centroid of objects are used to denote the locations of tracked objects. These results show the ability of proposed method in handling occlusion and deformation, especially in (c1) and (c5) when the boy in red has been occluded in large area.

5. **Conclusions.** A novel method is proposed in this paper to track and extract multiple objects and parts of them by a semantic multi-level framework. Through this method, objects and their interesting parts are input by users in a key frame of the video and initialized through a key frame segmentation algorithm and saved in a layer structure. Semantic relationships and SCGMMs spread frame by frame by using the proposed tracking method. In each frame, a twice-segmentation algorithm is used to obtain the refined extraction results of objects and interesting parts of them. Experiments on various video sequences show that the proposed method is effective in handling occlusion and deformation which bother non-rigid objects tracking all the time.

There is still a lot of work to do in future. The results of proposed method are very sensitive to the choice of key frame and the segmentation result of it. The proposed algorithm does not satisfy the demand of real time application due to its complexity. Hence, the proposed method and the results are looking forward to work effectively in succeeding applications such as video editing based on objects or parts of objects but not real-time monitoring.

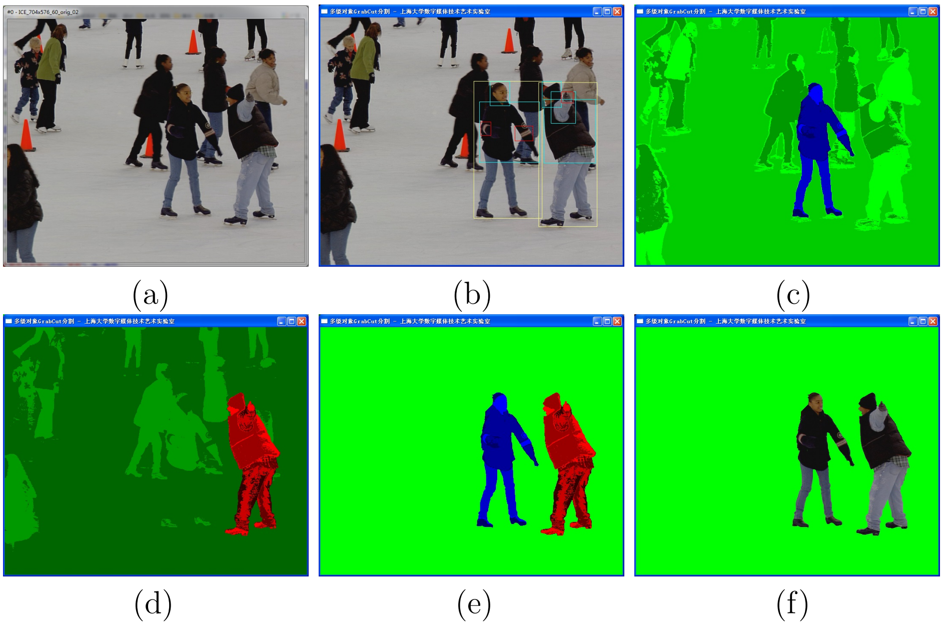

#0 - ICE_704x576_60_orig_02
多级对象GrabCut分割 – 上海大学数字媒体技术艺术实验室
(a)
(b)
(c)
(d)
(e)
(f)

Figure 7. Input and experimental results of a key frame. (a) is frame 0 of standard test sequence ICE 704×576 60 orig 02 yuv. (b) shows the user input of 3 levels, 10 layers. (c) and (d) show the SCGMMs of object 1 and object 2 at level 1. (e) and (f) are the results in mask of the SCGMMs image and the original frame.

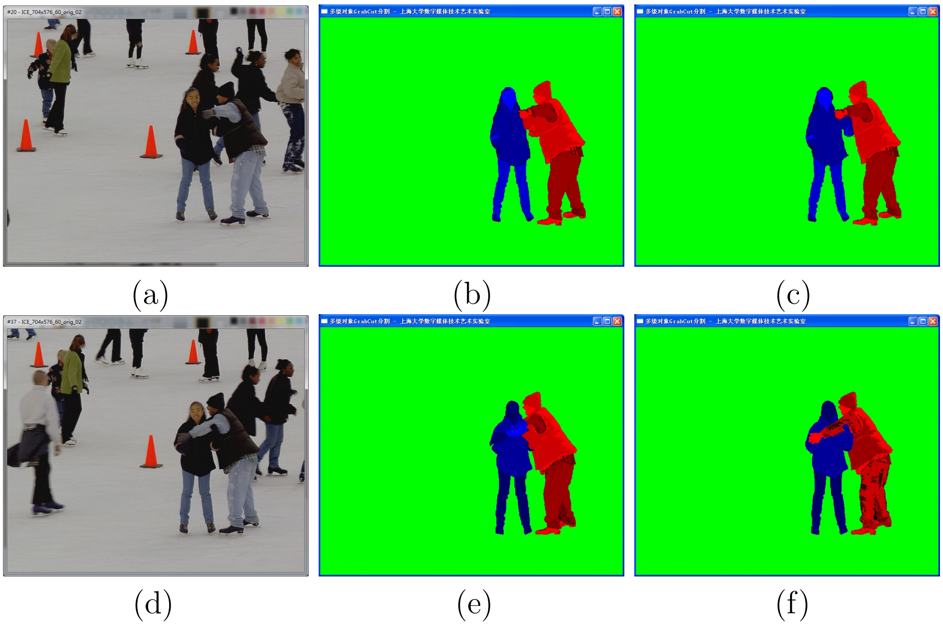
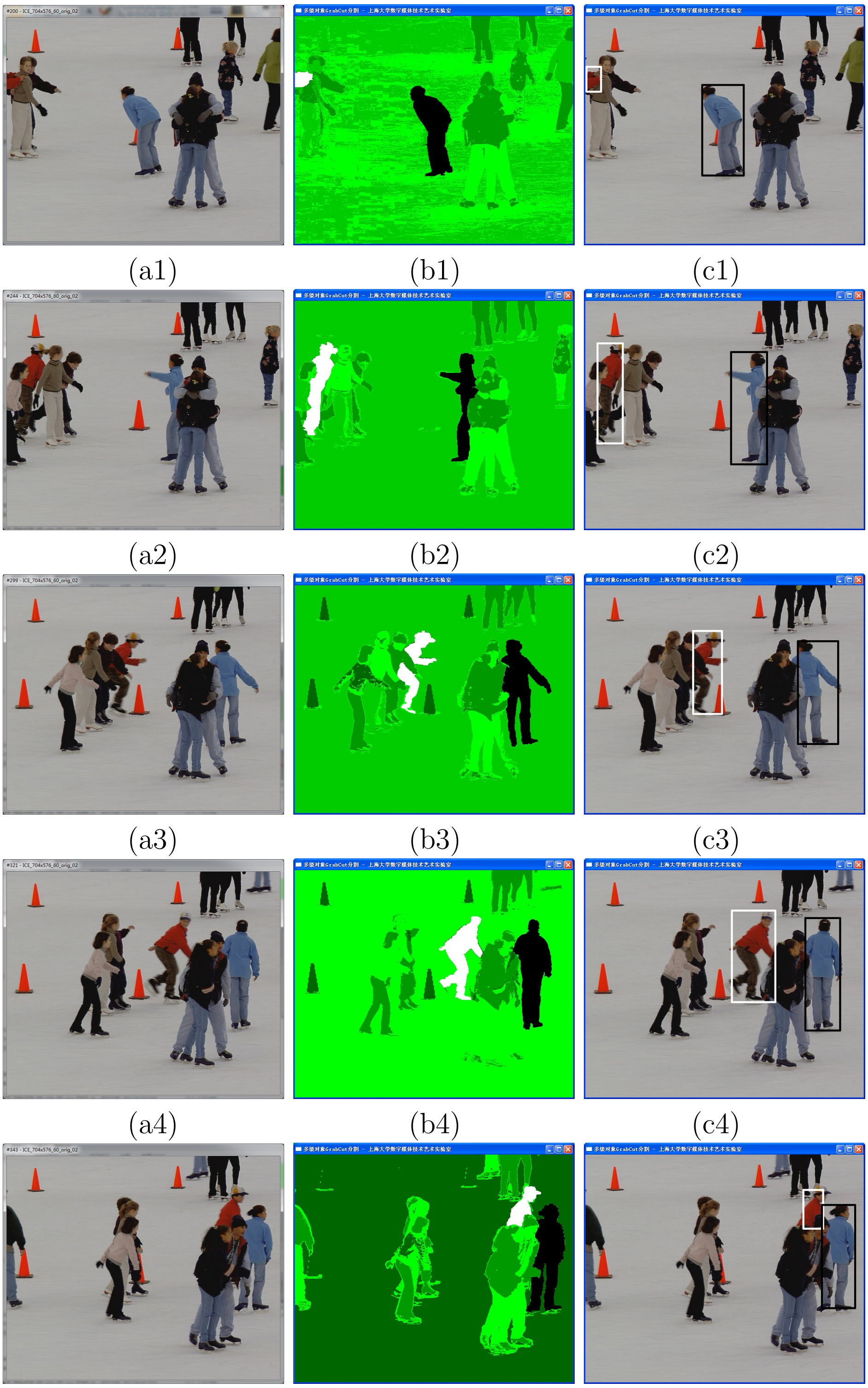

#20 - ICE_704x576_60_orig_02
#37 - ICE_704x576_60_orig_02
多级对象GrabCut分割 – 上海大学数字媒体技术艺术实验室
(a)
(b)
(c)
(d)
(e)
(f)

Figure 8. Experimental results of subsequent frames. (a) and (d) in the first column are frame 20 and frame 37 of standard test sequence ICE 704 ×576 60 orig 02 yuv respective. (b) and (e) in the second column are results of the two frames with only SCGMMs segmentation. (c) and (f) in the last column are results with the proposed method.



(a1) (b1) (c1)

(a2) (b2) (c2)

(a3) (b3) (c3)

(a4) (b4) (c4)

(a5) (b5) (c5)

Figure 9. Segmentation and tracking results of subsequent frames. (a1)∼(a5) in the first column: frame 200, 244, 299, 321 and 343 of standard test sequence ICE 704 ×576 60 orig 02 yuv respective. (b1)∼(b5) in the second column: segmentation results of corresponding frames at level 1. (c1)∼(c5) in the last column: tracking results of corresponding frames.

Acknowledgment. This work was partially supported by following grant: Technical Cooperation Project of Qingdao Technological University (No. B2-2013-0058); Natural Science Foundation of Zhejiang Province (No. R1110006); National High-technology Research and Development Program of China (863 Program) (No. 2013AA01A603); National Science and Technology Support Projects of China (No. 2012BAH07B01); Program of Science and Technology Commission of Shanghai Municipality (No. 12510701900, No. 13ZR1410400). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] E. Polat, M. Yeasin, and R. Sharma, A 2D/3D model-based object tracking framework, *Pattern Recognition*, vol.36, no. 9, pp.2127–2141, 2003.

[2] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[3] Y. Boykov, G. Funka-Lea, Graph cuts and efficient ND image segmentation, *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[4] J. Malcolm, Y. Rathi, and A. Tannenbaum, Multi-object tracking through clutter using graph cuts, *IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp. 1-5, 2007.

[5] J.F. Talbot, X.Q. Xu, Implementing grabcut, *Brigham Young University*, 2006.

[6] M.J. Wu, X.R. Peng, Q.H. Zhang, and R.J. Zhao, Segmenting and tracking multiple objects under occlusion using multi-label graph cut, *Computers and Electrical Engineering*, vol. 36, no. 5, pp. 927–934, 2010.

[7] B. Price, W. Barrett, Object-based vectorization for interactive image editing, *The Visual Computer*, vol. 36, no. 9-11, pp. 661–670, 2006.

[8] K. Zhao, W.J. Zhang, and Y. Jiang, Semantic interactions in multi-Level objects segmentation, *IEEE 2010 International Conference on Computational and Information Sciences*, Chengdu, China, pp.665-668, 2010.

[9] A. Delong, Y. Boykov, Globally optimal segmentation of multi-region objects, *IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, pp.285–292, 2009.

[10] Y. Boykov, M.P. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images, *IEEE 8th International Conference on Computer Vision*, Vancouver, Canada, pp. 105-112, 2001.

[11] L.M. Qin, C. Zhu, Y. Zhao, H.H. Bai, and H.W. Tian, Generalized gradient vector flow for snakes: new observations, analysis and improvement, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 883–897, 2013.

[12] P. Puranik, P. Bajaj, A. Abraham, P. Palsodkar, and A. Deshmukh, Human perception-based color image segmentation using comprehensive learning particle swarm optimization, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 2, no. 3, pp.227-235, 2011.