ISSN 1000-0054　清华大学学报（自然科学版）2012 年 第 52 卷 第 4 期　29/32
CN 11-2223/N　J Tsinghua Univ (Sci & Tech), 2012, Vol. 52, No. 4　561-565,570

结合 Kinect 深度图的快速视频抠图算法

何　贝，　王贵锦，　林行刚

（清华大学 电子工程系，北京 100084）

摘　要：现有视频抠图算法主要存在人机交互繁琐、计算复杂度高的问题，为此，该文提出了一种利用 Kinect 深度图的新的快速视频抠图算法。首先结合彩色图信息改进区域生长算法，估计出三色图（原始图像被 3 种颜色标记出前景、背景和未知区域）以避免深度图中遮挡区域的影响。其次，提出前景和背景样本点集二次筛选机制，保证估计精度的同时大幅降低计算复杂度。最后，采用深度、彩色和置信度图对抠图结果进行加权滤波，减少不透明度图像中低置信度的像素点和不平滑区域。实验结果证明了该算法精度高、速度快且交互简单。

关键词：视频抠图；Kinect 深度图；三色图生成；样本点集筛选；加权滤波

中图分类号：TP 391　　**文献标志码**：A

文章编号：1000-0054(2012)04-0561-05

Quick matting for videos based on depth images of the Kinect

HE Bei, WANG Guijin, LIN Xinggang

(**Department of Electronic Engineering, Tsinghua University, Beijing 100084, China**)

Abstract: Video matting can be computationally expensive. This paper presents a quick matting algorithm for videos based on depth images of the Kinect. First, the color image information is used to improve the region-growth process to estimate the trimap (marked as the foreground, the background and unknown regions). This scheme avoids the effect of occlusion regions. Secondly, samples refinement of the foreground and background regions is used to preserve the accuracy of the matting results while reducing the computational cost. Finally, the depth, color and confidence image are combined into a weighting filter to smooth the matting results and reduce the number of low confidence pixels. Tests verify the accuracy and speed of this algorithm.

Key words: video matting; Kinect depth image; trimap generation; samples refinement; weighting filter

抠图(matting)即从图像或视频中精确提取出前景区域的过程，现在已经广泛应用到图像或视频编辑中，包括背景剔除、前景调色以及图像分层等[1]。在抠图过程中，通常认为像素点 $\boldsymbol{I}_i$ 由前景点 $\boldsymbol{F}_i$ 和背景点 $\boldsymbol{B}_i$ 线性组合得到，满足式(1)模型，

$$\boldsymbol{I}_i = \alpha_i \boldsymbol{F}_i + (1-\alpha_i)\boldsymbol{B}_i. \tag{1}$$

其中 α_i 为 i 点的前景不透明度，表示该点前景分量吸收辐射的强弱。由式(1)可看出，未知变量的数目（$\boldsymbol{F}_i$、$\boldsymbol{B}_i$ 和 α_i）大大超过方程的数目，因此抠图问题可以看作是一个不定方程的求解。现有的抠图算法利用用户提供的三色图（利用三种颜色分别标记出待抠图像已知的前景、背景区域和未知区域）来辅助计算[2-8]。由于已知的前景区域的不透明度 α 值为 1，而已知的背景区域 α 值为 0，因此仅需要计算未知区域的 α 值。然而该方法需要用户手动标注，特别在视频抠图中需要标注每一帧的三色图，工作量大，交互繁琐；同时，高计算复杂度也限制了现有视频抠图算法的实际应用。

深度图像对光照变化不敏感，且对其进行简单处理即可获得较好的前、背景分割结果，因此现有的视频抠图算法引入深度图像信息来辅助抠图[2-5]。这不仅能用于自动生成三色图，减少用户交互，还能约束相邻像素点间 α 值的变化。Wang[2] 对采集的低分辨率深度图进行插值、二值化和形态学操作得到三色图，再利用改进后的 Bayesian 抠图算法[6] 来估计 α 值。由于基于采样的 Bayesian 抠图算法没有考虑相邻像素间的连续性，因此难以得到平滑的抠图。Zhu[3] 对深度图进行 k 均值聚类及形态学操作从而估计三色图；在此基础上，将深度值作为像素点的第四维信息来改进闭式解抠图算法[7]。但闭式解抠图算法没有利用估计的 α 值，同时也会传播错误的 α 值，因此也无法获得精确的抠图。Cho[4]

收稿日期：2012-03-12

作者简介：何贝(1987—)，男(汉)，安徽，博士研究生。

通信作者：王贵锦，副教授，E-mail：wangguijin@tsinghua.edu.cn

主要通过二值化和形态学操作得到三色图,再利用鲁棒抠图算法[8]估计 α 值。对于前景和背景颜色值接近的区域,Cho 以深度图的二值化结果作为相应的不透明度。该算法综合考虑了式(1)中的线性模型和不透明度图的连续性,能够得到精确的抠图结果,但选择样本点对和全局优化的计算复杂度较高。

针对现有视频抠图算法计算复杂度高、交互繁琐的问题,本文提出了一种基于 Kinect 深度图的新的视频抠图算法。首先,为了避免深度图中遮挡区域的影响,根据彩色图信息对区域生长算法进行改进,自动生成三色图。其次,针对抠图计算复杂度高的问题,提出并设计了前景和背景样本点集的二次筛选机制,在保证抠图精度的前提下大大降低了计算复杂度。最后,利用深度、彩色和置信度信息对不透明度图像进行加权滤波,得到平滑、精确的抠图结果。公共图像集和 Kinect 采集得到的视频集上的实验均表明本算法精度高、速度快且交互简单。

1 本文算法

图 1 给出了本算法的流程图。首先,根据 Kinect 采集得到的彩色和深度图,利用改进的区域生长算法得到准确的三色图。其次,采集并二次筛选每个未知像素点的前景和背景样本点集,选择置信度最高的一组样本点对估计不透明度。最后,同时结合彩色、深度和置信度图对抠图结果进行平滑,细化不透明度,得到准确的抠图结果。

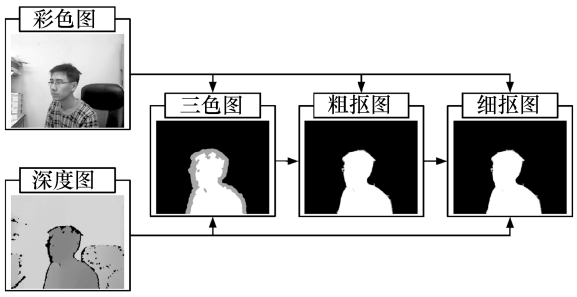


图 1 本文算法的流程图

1.1 三色图生成

与 ToF(time of flight)深度相机相比,Kinect 深度相机的优势在于深度图分辨率高、成本低,但采集的深度图中存在较多的遮挡区域。一方面,如果将遮挡区域均看作未知像素点,将会增大后续抠图的难度。另一方面,无反射区域(如黑色毛发等)在深度图上与遮挡区域表现相同,即深度值都为 0,那么根据深度值则无法进行分割。二值化[2,4]使得分割图中存在大量噪声点,增加了抠图的复杂度;深度值聚类[3]不能很好地处理遮挡区域。因此,本文融合彩色图对区域生长算法进行改进,自动计算得到三色图,能够避免遮挡区域的影响。首先用户手动指定待提取的区域,取其中心点为种子点,并利用改进的区域生长算法向邻域扩散。设像素点 i 为前景点,其 4-邻域像素点 j 是否属于前景区域的条件如式(2)所示:

$$f_j=\begin{cases}1, & \text{如果 } |D_i-D_j|<T_{\mathrm{d}};\\ 1, & \text{如果 } D_j=0 \text{ 且 } \|I_i-I_j\|_2<T_{\mathrm{c}};\\ 0, & \text{其他.}\end{cases} \tag{2}$$

其中:状态 1 表示属于前景,而状态 0 表示属于背景。当深度值 D 之差小于阈值 T_{d},或像素点 j 的深度值为 0 且彩色值 $\boldsymbol{I}$ 之差小于阈值 T_{c} 时,点 j 被判为前景点。对于深度值不为 0 的像素点,遵循一般区域生长算法的原则,只要深度值没有突变就认为该点属于前景区域。对于深度值为 0 的像素点,一般区域生长算法认为是背景点,这将导致部分无反射和遮挡区域误判为背景点。本文在区域生长的过程中引入了彩色值的比较,只要彩色值变化不大,就将其视为前景点。因此,改进的区域生长算法根据无反射和遮挡区域的彩色值的连续性对其进行区分,最终得到较为精确的前景和背景的分割结果。

考虑到深度图自身精度和分割误差的影响,本算法利用形态学操作对改进的区域生长算法的分割结果进行处理,估计出三色图。由于前景区域的 α 值为 1,背景区域为 0,对分割结果进行窗口大小为 w_E 的腐蚀可得到三色图的前景区域 F;对分割结果进行窗口大小为 w_{D} 的膨胀可得到三色图的背景区域 B;剩余部分即为三色图的未知区域 U。

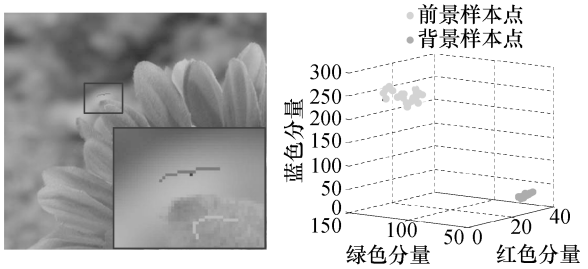
1.2 视频帧粗抠图

由于图像的连续性,在一个小的邻域中,可以粗略地认为图像彩色值不变[8]。因此,对每一个未知像素点来说,可以在其邻域采集最近的前景和背景点,根据式(1)估计 α 值。为了避免野点的影响,一般需要采集较多的前景和背景样本点,并从中选取置信度最高的样本点对。这里存在两个难点:一方面,需要定义置信度用于衡量每组样本点对,从而进行选取;另一方面,采集 n 个前景和背景样本点则构建出 $n\times n$ 个样本点对,直接选取点对的复杂度较高。

1.2.1 样本点集二次筛选

与 Cho 算法[4]类似,本算法采集与当前未知像

素点空间距离最小的 n 个前景和背景边界点作为样本点。但由于前景/背景图连续，采集得到的前景/背景样本点像素值通常变化不大，特别当前景/背景区域较为平坦的时候；因此，直接构建样本点对时，冗余的样本点集将会大大提高点对选择的复杂度。针对该问题，本算法设计了样本点集二次筛选策略以去除冗余样本点集。以前景样本点集为例：1）将 n 个样本点归为 n 个类，设定每个类的中心点为所含样本点的彩色值；2）当两个类中心点距离小于阈值 T_s，即 $\| \boldsymbol{I}_i - \boldsymbol{I}_j \| < T_s$ 时，两个类将被合并，且合并后新类的中心点为上述两个类中心点的加权和，权重为样本点数目；3）重复第 2 步，直到任意 2 个类均不能合并时停止二次筛选。背景样本点集重复同样的操作。如图 2 所示，采集的 20 个前景和背景样本点经过二次筛选后分别只剩下 12 个和 3 个，可构建的样本点对从 400 组下降到 36 组，大大降低了点对选择的复杂度。



(a) 欧式空间　　(b) 颜色空间

图 2　采集的前景和背景样本点在不同度量空间上的分布

1.2.2　不透明度和置信度估计

对于每组样本点对，根据式(1)，将当前未知像素点的彩色值投影到前景和背景样本点的彩色值张成的直线上即可估计出不透明度

$$\alpha_i^{(s,t)} = \frac{(\boldsymbol{I}_i - \boldsymbol{B}_i^{(t)})^{\mathrm{T}}(\boldsymbol{F}_i^{(s)} - \boldsymbol{B}_i^{(t)})}{\| \boldsymbol{F}_i^{(s)} - \boldsymbol{B}_i^{(t)} \|^2}. \tag{3}$$

其中 $\boldsymbol{F}_i^{(s)}$ 和 $\boldsymbol{B}_i^{(t)}$ 分别表示第 s 个前景和第 t 个背景样本点的彩色值。若样本点对和当前像素点的彩色值满足式(1)中的线性模型，或者当前像素点与样本点对的前景/背景彩色值接近，那么认为该组样本点对的置信度较高[8]。本文算法根据式(3)计算每组样本点对的置信度，选择置信度最高的样本点对估计不透明度。

$$c^{(s,t)} = \exp\left\{-\frac{r^{(s,t)} f^{(s)} b^{(t)}}{\sigma^2}\right\}. \tag{4}$$

其中 σ 用于调整权重。$r^{(s,t)}$、$f^{(s)}$ 和 $b^{(t)}$ 分别表示线性模型的误差、当前像素点与前景和背景彩色值的相似度，定义如下：

$$\begin{aligned} r^{(s,t)} &= \frac{\| \boldsymbol{I}_i - \alpha_i^{(s,t)} \boldsymbol{F}_i^{(s)} - (1 - \alpha_i^{(s,t)}) \boldsymbol{B}_i^{(t)} \|^2}{\| \boldsymbol{F}_i^{(s)} - \boldsymbol{B}_i^{(t)} \|^2}, \\ f^{(s)} &= \exp\{-\sigma_c^2 / \| I_i - F_i^{(s)} \|^2\}, \\ b^{(t)} &= \exp\{-\sigma_c^2 \| I_i - B_i^{(t)} \|^2\}. \end{aligned} \tag{5}$$

其中 σ_c 用于调整彩色值差的权重。

1.3　视频帧细抠图

选取样本点对进行粗估计能够得到每个未知像素点的不透明度，但逐个计算像素点的不透明度无法保证抠图结果的平滑，且存在较多的低置信度的点，因此需要对粗抠图的结果进行细化。像素点的不透明度主要受到邻域像素点的约束，但影响力各不相同。1）影响力随着与当前像素点空间距离的增大而减小；2）如果两个像素点的深度值相差较大，那么它们很可能属于不同的图像层，影响力较小；3）具有相似彩色值的两个像素点之间的影响较大；4）如果邻域像素点计算得到的不透明度置信度较低，那么应当降低该点的影响力。综合上述 4 个因素，本文设计了一种新的加权平滑策略来细化不透明度图像，像素点 i 的不透明度经过平滑后的结果 α_i' 为

$$\alpha_i' = \sum_{j \in N_i} \omega_{i,j}^{\mathrm{S}} \omega_{i,j}^{\mathrm{D}} \omega_{i,j}^{\mathrm{C}} c_j \alpha_j \Big/ \sum_{j \in N_i} \omega_{i,j}^{\mathrm{S}} \omega_{i,j}^{\mathrm{D}} \omega_{i,j}^{\mathrm{C}} c_j. \tag{6}$$

其中 $\omega_{i,j}^{\mathrm{S}}$、$\omega_{i,j}^{\mathrm{D}}$ 和 $\omega_{i,j}^{\mathrm{C}}$ 分别表示邻域中像素点 j 的空间、深度以及彩色距离权重，定义如式(7)。

$$\begin{aligned} \omega_{i,j}^{\mathrm{S}} &= \exp\{-\| x_i - x_j \|^2 / w_s^2\}, \\ \omega_{i,j}^{\mathrm{D}} &= \exp\{-\| D_i - D_j \|^2 / w_d^2\}, \\ \omega_{i,j}^{\mathrm{C}} &= \exp\{-\| I_i - I_j \|^2 / w_c^2\}. \end{aligned} \tag{7}$$

其中 $\boldsymbol{x}_i$ 表示像素点 i 的空间坐标，而 w_s、w_d 和 w_c 分别用于调整空间、深度和彩色距离的权重。由于本算法的加权平滑同时利用了空间、深度、彩色距离和置信度对每个邻域像素点进行评估，因此高置信度的邻域点能够对当前像素点的不透明度进行修正，从而得到更加精确的抠图结果。

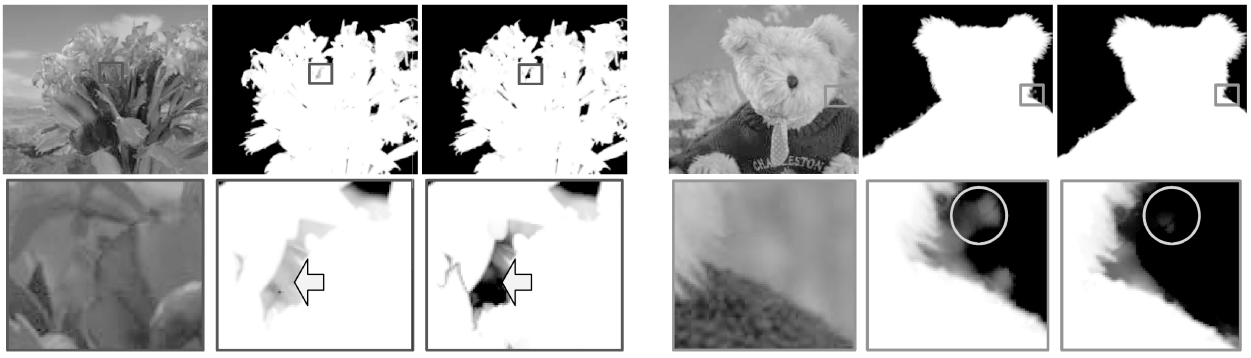
2　实验与讨论

为了评估本算法的性能，本文从图像和视频抠图两个角度与 Cho 算法[4]进行比较，两者都同时采用了样本点对的采样和利用了不透明度图像的连续性。实验环境为 Windows XP 的主机、Intel Core Ⅱ双核处理器、主频 2.0 GHz。首先，在文[9]提供的用于评估抠图算法性能的公共图像测试集上进行测试；其次，在利用 Kinect 采集的 3 组视频集上展开实验。

2.1 图像抠图

文[9]提供的公共测试集包含26组图像,分别提供了彩色图、三色图和不透明度的真值图。由于测试集中不包含深度图,因此Cho算法退化为鲁棒抠图算法[8],本文算法等效于细抠图模块中深度权重为1的情况。图像集的部分比较结果如图3所示。从图3a可以看出,Cho算法的全局优化能够平滑边缘,但孔洞区域处理结果不好。由于背景区域和前景区域的叶子部分彩色值相近,利用样本点对估计不透明度不能得到精确的抠图结果,而低置信度的结果会降低邻域点的估计精度。Cho算法将孔洞的背景部分误判为前景和背景的混合,如图3a中箭头所示。而本文算法结合空间、深度、彩色距离和置信度进行加权平滑,不仅不会受到这些低置信度点的影响,而且一些估计错误的像素点可以被邻域中高置信度的颜色相近的点纠正回来。如图3b中放大区域所示,背景区域灰白色部分在Cho算法的结果中依旧为前景和背景的混合,而本算法能够将该部分纠正为纯背景区域,从而获得更加精确的结果。

不透明度图像的误差和计算时间的比较如图4所示。本算法和Cho算法性能持平,且在部分测试图像(如图4*a*中箭头所示)上获得更好的结果。由于本文算法提出了样本点集二次筛选和加权平滑代替全局优化的策略,大大降低了抠图复杂度。如图4b所示,本算法的计算速度比Cho算法提高了2~8倍。同时,对于n个未知像素点,本算法只需要存储一些临时图像,空间复杂度为$O(n)$;而Cho算法的全局优化则需要构建$O(n^2)$的空间用于存储Laplace矩阵[7]。因此,本文算法具有更低的时间和空间复杂度。



(a) 图像"GT21"[9] (b) 图像"GT23"[9]

(依次为原始图像(左)、Cho算法[4](中)和本文算法(右)的结果;第1行为整图,第2行为局部放大)

图3 公共测试集上抠图结果的比较

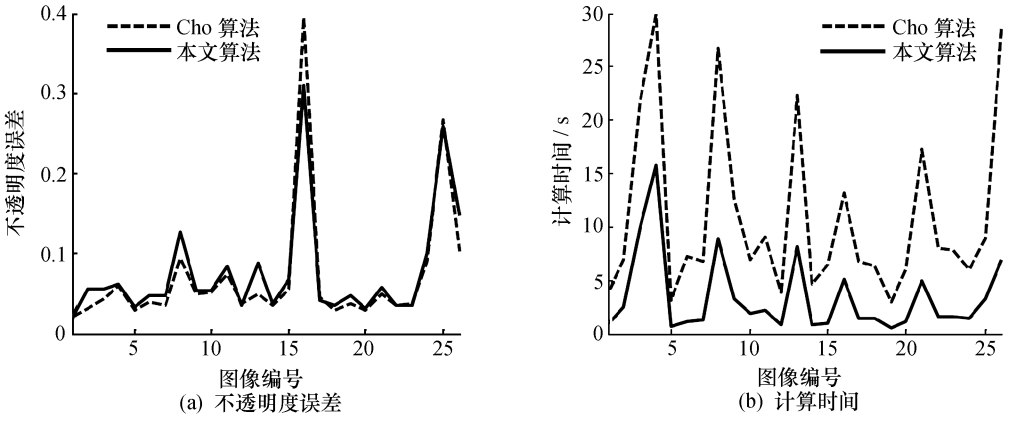
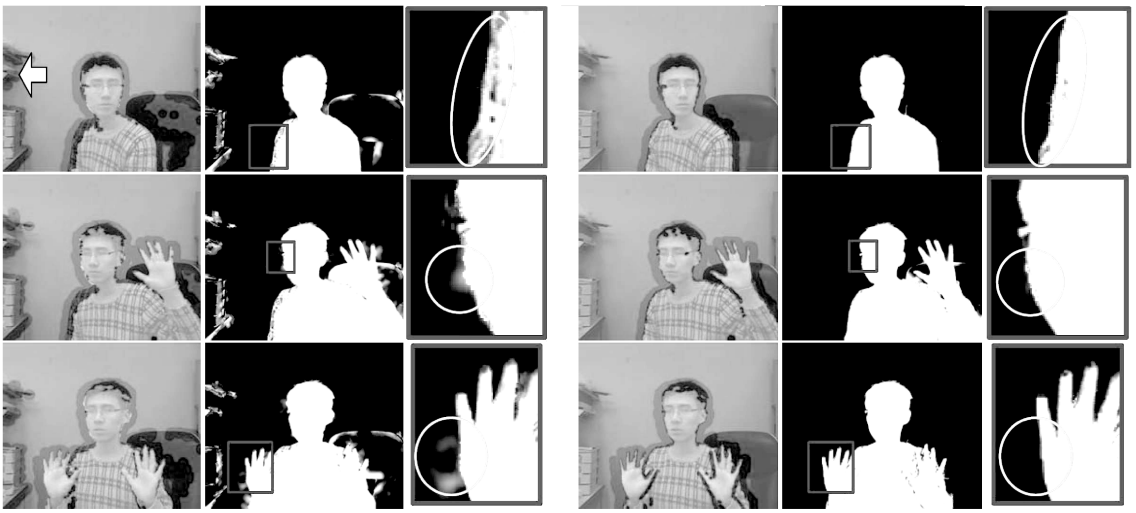


图4 Cho算法[4]和本文算法在公共测试集上的量化比较

2.2 视频抠图

本文利用Kinect采集3组视频,分辨率为640×480,图像帧数为500,都提供了彩色和深度信息。3组视频分别包含3种动作:目标身体转动、手臂左右移动以及张开手掌的缩放。Cho算法和本文算法的比较结果如图5所示。1) Cho算法将干扰区域判断为未知区域,如图5a中箭头所示;

而本算法利用前景区域的连通性，能够避免该情况的发生，不仅减少了抠图的计算复杂度，而且避免了与前景颜色接近的背景目标的影响。2）当目标身体转动时，Cho 算法中的二值化不能很好地分割前景和背景，如图 5a 中第一行放大区域所示，从而不能得到精确的抠图结果。3）本文算法的细抠图模块利用邻域中高置信度像素点的 α 值更新低置信度的点，因此能纠正一些错误估计的 α 值；而 Cho 算法的全局优化可能会将低置信度点的 α 值传播出去，从而引起累积误差，如图 5 中第 2 行和第 3 行的放大区域所示，本算法获得更精确的抠图结果。



(a) Cho 算法　　(b) 本文算法

（依次为三色图（左）、不透明度图结果（中）和放大区域（右）；从上到下依次为视频 1 的 279 帧、视频 2 的 491 帧和视频 3 的 315 帧）

图 5　Kinect 视频集上抠图结果的比较

Cho 算法和本文算法在 3 组测试视频上的抠图时间如表 1 所示，可以看出，本文算法比 Cho 算法速度平均提高了 12 倍左右，达到了 0.4 s/帧。由于本算法设计了样本点集二次筛选的机制，大大减少了样本点对的数目。相对于公共图像测试集来说，Kinect 的 3 组视频背景颜色较为单一，样本点集在彩色空间上更集中，因此本文算法更加有效。另外，本文的加权平滑模块在获得较好性能的同时，能够避免 Cho 算法的全局优化算法的高时间复杂度；Cho 算法的二值分割与本文的改进的区域生长算法相比，产生的未知像素数目更多，进一步增加了抠图的计算时间。

表 1　Kinect 视频集上 Cho 算法[4]和本文算法的时间比较

视频	计算时间/(s·帧$^{-1}$)	
	Cho 算法	本文算法
视频 1	5.73	0.45
视频 2	5.91	0.47
视频 3	5.78	0.42

3　结　论

本文提出了一种结合 Kinect 深度图的快速视频抠图的算法，解决了现有视频抠图算法中存在的交互繁琐和复杂度高的问题。首先，结合彩色图信息对区域生长算法进行改进，用于生成三色图，避免深度图中遮挡区域的影响。其次，对采集的前景和背景样本点集提出二次筛选的机制，大大降低了不透明度估计的复杂度。最后，本算法结合深度、彩色和置信度图，进一步细化抠图结果，避免了不平滑区域和低置信度的不透明度的影响。实验证明本算法交互简单，抠图结果精度高且计算复杂度低。

参考文献　(References)

[1] Wang J, Cohen M F. Image and video matting: A survey [J]. *Foundations and Trends(R) in Computer Graphics and Vision*, 2007, **3**(2): 97 - 175.

[2] Wang O, Finger J, Yang Q, et al. Automatic natural video matting with depth [C]// Proc of Pacific Graphics. Hawaii, USA: IEEE Press, 2007: 469 - 472.

（下转第 570 页）

[7] Yang M, Zhang L, Feng X, et al. Fisher discrimination dictionary learning for sparse representation [C]// Proc ICCV. 2011: 543-550.
[8] Zhang L, Yang M, Feng X. Sparse representation or collaborative representation: Which helps face recognition? [C]// Proc ICCV. 2011: 471-478.
[9] Gao S H, Tsang I W, Liang T, et al. Local features are not lonely—Laplacian sparse coding for image classification [C]// Proc CVPR. 2010: 3555-3561.
[10] Yang J C, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification [C]// Proc CVPR. 2009: 1794-1801.
[11] Huang K, Aviyente S. Sparse representation for signal classification [C]// Proc NIPS. MIT Press, 2006: 609-616.
[12] Mairal J, Bach F, Ponce J, et al. Supervised dictionary learning [C]// Proc NIPS. MIT Press, 2008: 1033-1040.
[13] Lee H, Battle A, Raina R, et al. Efficient sparse coding algorithms [C]// Proc NIPS. MIT Press, 2006: 801-808.
[14] Lee K, Ho J, Kreigman D. Acquiring linear subpsaces for face recognition under variable lighting [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005, **27**(5): 684-698.
[15] Georghiades A, Belhumeur P N. From few to many: Illumination cone models for face recognition under variable lighting and pose [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(6): 643-660.
[16] Hull J J. A database for handwritten text recognition research [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994, **16**(5): 550-554.

(上接第 560 页)

[9] Allauzen C, Riley M, Schalkwyk J, et al. OpenFst: A general and efficient weighted finite-state transducer library [J]. *Computer Science*, 2007, **4783**: 11-23.
[10] Sangwan A, Hansen J H. Automatic analysis of Mandarin accented English using phonological features [J]. *Speech Communication*, 2012, **54**(1): 40-54.
[11] Siniscalchi S M, Reed J, Svendsen T, et al. Exploring universal attribute characterization of spoken languages for spoken language recognition [C]// Proceedings on Interspeech. Brighton, United Kingdom: International Speech Communication Association, 2009: 168-171.
[12] Siniscalchi S M, Svendsen T, Lee C H. Towards bottom-up continuous phone recognition [C]// Proceedings on ASRU. Kyoto: IEEE Press, 2007: 566-569.
[13] Hermansky H, Ellis D P W, Sharma S. Tandem connectionist feature extraction for conventional HMM systems [C]// Proceedings on ICASSP. Istanbul: IEEE Press, 2000: 1635-1638.
[14] Johnson D, Ellis D, Oei C, et al. ICSI QuickNet software package [EB/OL]. [2011-04-25]. http://www.icsi.berkeley.edu/Speech/qn.html.

(上接第 565 页)

[3] Zhu J, Liao M, Yang R, et al. Joint depth and alpha matte optimization via fusion of stereo and time-of-flight sensor [C]// Proc of CVPR. Florida, USA: IEEE Press, 2009: 453-460.
[4] Cho J H, Ziegler R, Gross M, et al. Improving alpha matte with depth information [J]. *IEICE Electronics Express*, 2009, **6**(22): 1602-1607.
[5] Sun W, Au O C, Xu L, et al. Adaptive depth map assisted matting in 3D video [C]// Proc of ICME. Barcelona, Spain: IEEE Press, 2011: 1-6.
[6] Chuang Y Y, Curless B, Salesin D H, et al. A Bayesian approach to digital matting [C]// Proc of CVPR. HI, USA: IEEE Press, 2001: 264-271.
[7] Levin A, Lischinski D, Weiss Y. A closed form solution to natural image matting [C]// Proc of CVPR. NY, USA: IEEE Press, 2006: 228-242.
[8] Wang J, Cohen M F. Optimized color sampling for robust matting [C]// Proc. of CVPR. Minneapolis, USA: IEEE Press, 2007: 1-8.
[9] Rhemann C, Rother C, Wang J, et al. A perceptually motivated online benchmark for image matting [C]// Proc of CVPR. Florida, USA: IEEE Press, 2009: 1826-1833.