# 一种面向无人机视频的多尺度摘要的设计与实现

杨大慎[1]， 陈科圻[2,3]， 马翠霞[2,3]

(1. 中国石化销售有限公司华南分公司，广东 广州 510000；
2. 中国科学院大学计算机科学与技术学院，北京 100190；
3. 中国科学院软件研究所人机交互北京市重点实验室，北京 100190)

**摘 要：**无人机视频是利用无人机航拍得到的一类重要的视频资源，被广泛运用于地面目标的监测。但是，无人机视频的视野辽阔、不具有目标针对性的拍摄特点，使其存在大量时空冗余，传统的视频交互手段显得十分低效。为此，提出了一种面向无人机视频的多尺度螺旋摘要。首先，基于 YOLOv3 算法，训练能检测无人机视角的行人、车辆等目标的模型。然后，提出了基于关键帧的视频目标检测算法，根据改进后的基于颜色特征的关键帧提取算法提取涵盖视频关键信息的关键帧，并将检测模型应用于关键帧，高效获取整个视频的目标检测结果。之后，从关键帧中提取相应的关键区域，作为摘要的呈现单元，并以螺旋的形式从内向外地将摘要单元逐一呈现，辅以基于关键帧的视频定位和尺度缩放功能。最后，开发了草图注释、目标分布螺旋、双螺旋播放等新颖的交互工具，满足用户的潜在需求，共同实现面向无人机视频的高效交互。

**关 键 词：**无人机；视频摘要；视频目标检测；小目标检测；螺旋摘要；视频交互
**中图分类号：**TP 391 **DOI：**10.11996/JG.j.2095-302X.2020020224
**文献标识码：**A **文 章 编 号：**2095-302X(2020)02-0224-09

# Design and implementation of a multi-scale summarization for unmanned aerial vehicle videos

YANG Da-shen[1], CHEN Ke-qi[2,3], MA Cui-xia[2,3]

(1. South China Branch of Sinopec Sales Co., Ltd., Guangzhou Guangdong 510000, China;
2. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China;
3. Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract:** Unmanned aerial vehicle (UAV) videos, an important video resources captured by unmanned aerial vehicles, are now being widely used in ground target monitoring. However, there's usually a large amount of space-time redundancy in UAV videos due to their grand view and unspecified targets, making the traditional methods of video interaction inefficient to get usable details. To solve the problem, a multi-scale spiral summarization for UAV videos was proposed. Firstly, we trained a detection model based on YOLOv3 algorithm to detect the small targets including pedestrians and vehicles from the UAV's perspective. Then, we proposed a key-frame-based video object detection algorithm, by first extracting the key frames of the videos according to the improved color-feature-based key-frame-extraction algorithm, and then applying the model on the key frames to get the target detection results of the whole video. The key areas from the

**收稿日期：**2019-12-10；**定稿日期：**2019-12-16
**基金项目：**国家自然科学基金项目(2018YFC0809303)
**第一作者：**杨大慎(1979–)，男，山东德州人，工程师，本科。主要研究方向为管道管理、无人机图像研究等。E-mail：527667227@qq.com
**通信作者：**马翠霞(1975–)，女，山东高唐人，研究员，博士。主要研究方向为人机交互、媒体大数据可视分析。E-mail：cuixia@iscas.ac.cn

key frames were extracted as the displaying units of video summarization in a spiral form from the inside out with basic functions including key-frame-based video location and dynamic scaling. At last, some novel extended interaction tools were developed including sketch annotation, object distribution spiral and double spiral player, aiming to meet the users' potential needs, and help them interact with the UAV videos more efficiently.

**Keywords:** unmanned aerial vehicle; video summarization; video object detection; small object detection; spiral summarization; video interaction

无人机视频是利用无人机航拍得到的一类重要的视频资源，被广泛运用于地面目标的监测。其具有监控视频类似的特点：含有大量的时空冗余；反映特定时间、空间内的整体信息，不具有目标针对性。和监控视频的区别在于：无人机视频的镜头是移动的，不存在固定背景，且目标往往比监控视频的目标小得多。如何让用户与无人机视频进行高效交互，从中获取关键内容，是当前的一个重要研究课题。

传统的视频信息获取方式是顺着时间轴逐帧播放视频，并辅以快进、快退、拖拽进度条等交互手段。但是，无人机视频丰富的语义信息、不明确的目标十分考验人的认知水平，长时间的用眼负荷也会显著降低人的注意力，快进、快退等加速手段又存在着跳过关键信息的风险，这些因素导致获取视频信息的常规手段显得十分低效。因此，为了提高视频信息的获取效率，用于提炼视频关键信息的视频摘要技术便应运而生，摘要内容可作为视频交互的媒介。目前通用的视频摘要技术普遍是基于关键帧，而关键帧算法核心在于如何衡量视频帧的重要性。基于颜色、纹理等底层视觉特征，能够从图像的客观统计特性的角度将视频帧之间的差异量化；针对无人机视频的感兴趣目标提取语义信息，能够更精确地判断视频的时空冗余。

获得关键帧后，需要考虑以怎样的形式将这些摘要信息呈现出来，并提供给用户与这一摘要形式相搭配的交互手段，辅助信息获取。BARNES 等[1]提出将视频摘要以长矩形条的形式展示，LIU 等[2]将视频摘要以螺旋的形式由内到外排布，两者均基于关键帧实现视频定位与尺度缩放，且具有良好的可拓展性。相较而言，螺旋摘要在有限的空间内能展现更丰富的内容，且以螺旋线为时间轴保证了视觉上的连续性，在高效的同时兼顾了更加自然、美观的视觉效果。

因此，本文提出了一种面向无人机视频的多尺度螺旋摘要。首先，采用 YOLOv3 算法[3]训练了能检测无人机视角下的行人和车辆等小目标的模型。考虑到逐帧检测整个视频既耗时，检测结果也会大量重复，本文提出了基于关键帧的视频目标检测算法：运用改进的基于颜色特征的关键帧提取算法，从视频中提取出关键帧，再将目标检测模型应用于这些关键帧上，得到整个无人机视频的检测结果。之后，根据检测结果从关键帧中提取包含检测目标的摘要单元，并以螺旋的形式从内向外将摘要单元逐一呈现，绘制出完整的螺旋摘要，辅以基于关键帧的视频定位、尺度缩放等功能。最后，还设计了草图注释、目标分布螺旋和双螺旋播放等新颖的交互手段，共同实现对无人机视频的高效交互。

## 1　相关工作

### 1.1　视频摘要

视频摘要的主要研究是基于关键帧。经典的关键帧提取算法主要通过分析视频的底层视觉特征(包括颜色、纹理、运动等)来量化视频帧之间的差异。例如，WOLF[4]通过计算光流，筛选出运动强度较小的帧作为视频的关键帧；ZHANG 等[5]采用颜色等视觉标准，选择显著变化的帧作为关键帧；ZHUANG 等[6]对视频帧聚类，则从每一类中选择有代表性的帧作为关键帧。

近年来，AlexNet[7]，VGGNet[8]，ResNet[9]等算法的涌现，让图像语义理解的发展达到了空前的高度，并可直接作为视频摘要的选择依据。例如，Faster R-CNN[10]，YOLO[11]等网络在目标检测问题上表现优异，R-C3D[12]，TAL-Net[13]等网络则在时序动作定位问题上崭露头角。

### 1.2　视频交互

传统的视频交互手段单一，主要是通过拖拽视频的进度条和快进、快退实现定位。其交互比

较盲目，既难以定位自己想要的位置，又容易遗失重要信息。DRAGICEVIC 等[14]对界面交互的直接性做了分析，并提出了一种直接拖拽视频目标的交互手段。GOLDMAN 等[15]综合考虑了更多视频交互的辅助性手段，比如引入视频注释，包括描述性标签、说明性草图等。

除了在原始视频上直接进行交互外，也有学者利用视频摘要作为交互媒介，研究多样化的摘要呈现形式，并开发对应的视频交互手段：文献[1]将视频的关键帧以长矩形条的形式呈现，允许用户通过关键帧进行视频定位、自由缩放关键帧的尺度；文献[2]提出了螺旋摘要，同样支持基于关键帧的视频定位与尺度缩放，在兼顾自然与美观的视觉效果的同时，更充分地利用了空间。

## 2 无人机视频的目标检测

### 2.1 基于 YOLOv3 的小目标检测

无人机视频由于在拍摄时只遵循固定的路线、不针对具体目标，且拍摄角度高、视野辽阔，用户真正感兴趣的行人、车辆等小目标的信息总是被大量的时空冗余所淹没。为此，本文提出通过目标检测来提取关键信息，最终选择了权衡速度与精度的 YOLOv3 算法[3]。

无人机视角下的目标具有以下特点：外观上均为俯瞰视角；尺度上只有极端多的小目标。为了解决外观差异，采用了 VisDrone 数据集[16]对模型进行训练。该数据集由 7 000 多张无人机视角的图片组成，囊括了 pedestrian，car，van，truck，bus 5 类常见目标，包含了丰富的场景，且符合本文需求。为了解决尺度差异，使模型能更适应无人机数据集“小”的特点，本文主要做了 3 种处理：①扩大输入图片的尺寸，使模型更充分地学习到小目标的特征；②运用 k-means 聚类算法，根据训练集中目标的尺寸，确定 YOLO 模型的 Anchor；③对网络结构进行了微调，将模型中 FPN 架构的第二次上采样由 2 倍改为 4 倍，使网络能够将深层特征与更加浅层的特征结合，进而检测到更小的目标。

最后训练得到的模型在无人机视频帧上取得了不错的表现，具有良好的泛化能力，部分检测结果如图 1 所示。

### 2.2 基于关键帧的视频目标检测

获得能够检测无人机视角的目标模型后，虽然可直接逐帧进行检测，但这么做却过于耗时，且检测结果大量重复。为此，本文提出了基于关键帧的视频目标检测，顾名思义，即先从视频中提取关键帧，然后再对其进行目标检测，得到整个视频的检测结果。这样能大幅减少目标检测所需时间，也避免了重复性的检测结果，但同时对关键帧提取算法提出了 2 个要求：①计算简便，如耗时比 YOLOv3 逐帧检测视频长，则是得不偿失；②具有一定鲁棒性，保证算法在面对不同的视频时均能涵盖到大多数目标。提取关键帧的常用算法是基于视频每一帧的底层视觉特征，其中颜色信息相对而言最为稳定，且计算代价也很小。因此，本文采取了基于颜色特征的关键帧提取算法。

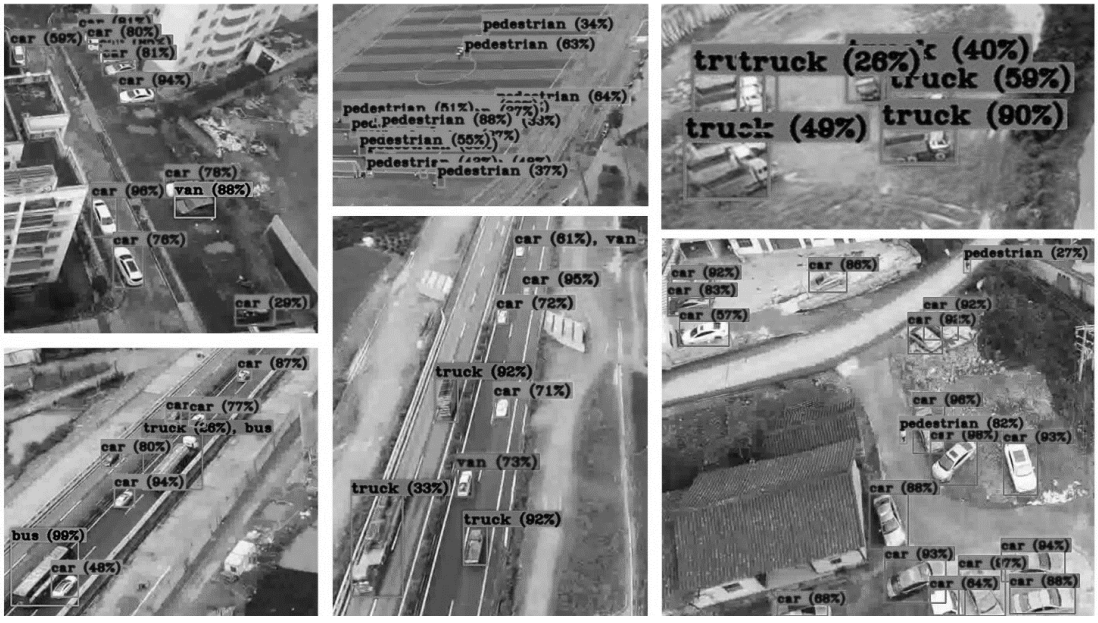



**图 1** 无人机视频的目标检测

基于颜色特征提取关键帧的传统算法主要有 2 种：①选择视频第一帧作为关键帧，依次计算下一帧与上一关键帧的颜色差值，若大于阈值，则选择此帧作为新的关键帧，以此类推，直到视频遍历完毕[5]；②对视频帧进行聚类，选取每一类中最具代表性的帧，组合得到整个视频的关键帧[6]。第 1 种算法是将视频拆分成片段，将每一片段的第 1 帧作为关键帧，因此其关键帧反映了视频的时序信息，但数量无法精确控制，只能通过改变阈值控制帧的密集程度。第 2 种算法的每一类视频帧具有跳跃性，因此提取出的关键帧不具备时序信息，但关键帧之间的差异整体更大，且数量能够根据设定的聚类数量实现精确控制(自适应聚类则另当别论)。

经过权衡，本文采取了第一种基于颜色特征的关键帧提取方法，主要原因有：①关键帧数量的精确控制并不具备实际意义，因为用户通常不知道从一个视频中提取多少关键帧才合适；②视频的时序信息很重要，能够反映前后关键帧所代表的视频片段之间的内在联系；③该算法能够通过改变阈值简单地实现关键帧数量和涵盖的信息量之间的权衡，而聚类算法的类别数和视频的时长、帧率息息相关，调节起来较烦索。

但是，该算法的主要缺陷为，阈值 $T$ 需用户给定，但具体取值多少，只能反复尝试。为此，本文提出了一种给定最小/最大间隔帧数的关键帧提取算法加以改进。新算法不再需要给定阈值 $T$，而是让用户给出相邻关键帧的最小/最大间隔帧数(MinFrames/MaxFrames)，且阈值 $T$ 可根据其值动态调整。此外，还引入了参数 Step，表示每隔 Step 帧进行一次视频帧的颜色计算和比较，这样能在对关键帧的选取影响较小的情况下，显著提高算法的运行效率。步骤流程如下：

**步骤 1.** 初始化阈值 $T$，定义缓冲变量 TempD 为零，定义关键帧集合 KeyFrames，将视频的第 1 帧作为关键帧加入 KeyFrames。

**步骤 2.** 从视频中获取第 Step 帧后的图像帧，并计算该帧和 KeyFrames 中最后一个关键帧的颜色直方图归一化后的差值 $D$。

**步骤 3.** 令 $N$ 为当前帧和最后一个关键帧的间隔帧数。当 $N<$MinFrames 时，若 $D>T$，则令 $T=D$；当 MinFrames$\leqslant N<$MaxFrames 时，若 $D\leqslant T$ 且 $D>TempD$，则记录当前帧为临时帧，且 Temp$D=D$，若 $D>T$，则令当前帧作为关键帧加入 KeyFrames；当 $N\geqslant$MaxFrames 时，则令临时帧作为关键帧加入 KeyFrames，且 $T=$Temp$D$。

**步骤 4.** 若视频第 Step 帧后不为空，则返回第 2 步，否则算法结束。

提取出关键帧后，将 3.1 节训练得到的模型应用于其中，检测图像中是否存在 pedestrian，car，van，truck，bus 这 5 类目标，最终得到的整个视频是基于关键帧的目标检测结果。

# 3　无人机视频可视分析

## 3.1　多尺度螺旋摘要

### 3.1.1　摘要单元提取

为了实现无人机视频的可视分析，需要将关键帧以螺旋的形式呈现。关键帧提取去除了视频的时间冗余，从关键帧中提取关键区域，去除空间冗余。传统提取图像关键区域的算法是基于图像的显著性检测[17]，对于普通图像的效果尚可，但是对无人机图像的效果却很差。究其原因，普通视频的镜头通常是精心把控的，考虑构图和色彩对比，关键目标很醒目。而无人机在拍摄时没有明确目标，只遵循固定的拍摄路线，感兴趣目标占的空间比例小，色彩对比也要微弱。为了能够更好地提取无人机视频的关键区域，必须引入语义信息。

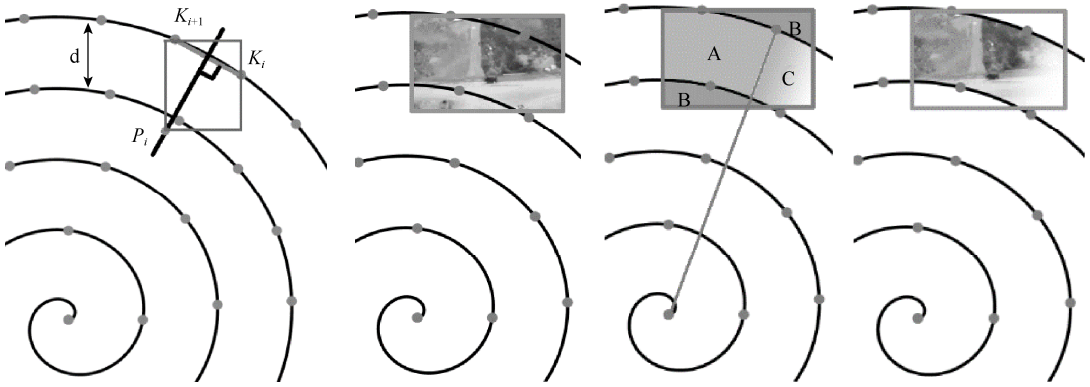
当获得了无人机视频的关键帧的目标检测结果，即可根据目标的方位，确定能够包围这些目标的最小矩形框，进而确定关键帧的关键区域。但最小矩形框不能直接作为最终的摘要单元，因为每一关键帧的目标均不同，矩形框的长宽比存在着很大差异，但摘要呈现时，每一单元的分辨率是固定的。如果直接将矩形框强行拉伸为需要的分辨率，将出现严重的图片变形。因此，本文提出将矩形框的长或宽进行一定程度的扩大，保证摘要单元有一个固定的长宽比，再将得到的关键区域统一为固定的分辨率，就得到了视频摘要的呈现单元。

### 3.1.2　螺旋摘要绘制

螺旋摘要的呈现是基于螺线的螺距 $d$ 恒定不变的性质。因此，在绘制螺旋摘要时，在螺旋线上每经过固定的弧长确定一个关键点，即可将螺旋划分为面积相似的区域，进而容纳摘要单元。

虽然螺旋摘要的呈现顺序是由内到外旋转，但考虑人的视觉习惯，摘要本身不能有任何旋转。

因此，将基于关键点确定正放的矩形框来容纳关键帧。以图 2(a)中的第 $i$ 个矩形框为例，将第 $i$ 和第 $i$+1 个关键点 $K_i$ 和 $K_{i+1}$ 相连，得到线段 $K_iK_{i+1}$，并作出中垂线。在中垂线上取靠近螺旋中心且和线段 $K_iK_{i+1}$ 的距离为 1.25 d 的点 $P_i$，根据 $K_i$, $K_{i+1}$, $P_i$ 这 3 个点，确定一个最小的矩形，即为第 $i$ 个摘要单元的矩形框。以此类推，根据关键点计算出螺线上所有的矩形框后，即顺利完成了螺旋区域的划分。再将摘要单元依次放入矩形框中，本文规定将摘要单元的中心和对应矩形框的中心位置相对应和将摘要单元整体进行缩放，保证其长或宽与矩形框的长或宽相等，且面积更大。最后，需对摘要进行边界处理，避免相邻的摘要彼此重叠或溢出螺旋线外，如图 2(b)所示。图中 B 区域的像素点因为溢出螺旋线外，需直接剔除。A，C 区域位于螺旋线内，被起始点和关键点相连的线分为两部分。按照螺线由内到外的旋转方向，A 区域往螺旋外，像素点保持不变；C 区域往螺旋内，属于过渡区域。过渡区域内，像素点随着离 A，C 区域分界线的距离越远，透明度逐渐线性过渡到零，完成平滑过渡。将所有经过边界处理后的摘要单元置于螺旋中，可得到了螺旋摘要的完整形状。螺旋摘要能够预览相应位置的原始关键帧，也支持基于关键帧的视频定位。




(a) 根据关键点确定矩形框　　(b) 对放入矩形框的关键帧进行边界处理

**图 2**　绘制螺旋摘要

### 3.1.3　螺旋摘要尺度缩放

虽然螺旋摘要已经极大地利用了空间，但在面向较长的视频时，要想完整地将上百张关键帧同时呈现，显然是不现实的。因此，需引入尺度缩放的操作，实现多尺度螺旋摘要。当用户想要以尽可能少的关键帧概括视频的整体情况时，可以在一定范围内将尺度放大，做法是将每间隔一帧的关键帧隐藏；当用户想要仔细察看更详细的关键内容时，可在范围内将尺度缩小，即将帧与帧之间未显示的隐藏帧显示出来。

为了实现尺度缩放的平滑的动态过程，本文设计了相应的动态缩放算法。首先根据缩放范围，记录缩放前、后关键帧的移动路程 Num，计算关键帧的移动步长 StepNum=0.1×Num。对于缩放前的移动关键帧，将螺旋摘要擦除并重绘 5 次，每次重绘时关键帧会移动步长 StepNum 的距离。从第 6 次开始，重绘缩放后的移动关键帧，同样是相对于上一次绘制的关键帧位置移动步长 StepNum，直到第 10 次重绘结束后，关键帧相对于缩放前移动了距离 Num，整个动态缩放过程完成。图 3 展示了全局尺度缩小的动态过程，其中阴影部分为动态缩放过程中显现的隐藏帧。

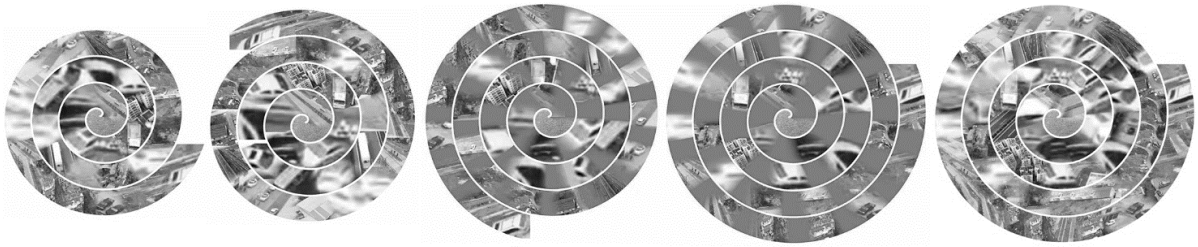


**图 3**　螺旋摘要全局动态尺度缩小，阴影部分为隐藏帧的显现

可以看出，螺旋摘要的动态缩放过程是一个将摘要反复擦除再重绘的过程，这就引出了一个关键问题：每一次重绘摘要时，关键帧在螺线上的位置均在变化，因此需重新进行像素级别的摘要边界处理，十分耗时。一旦重绘无法在 0.1 s 内完成，将会被肉眼察觉到断断续续的缩放过程，无法得到需求的平滑效果。因此，考虑到螺旋上的所有关键帧的分辨率均是相同的，不同关键帧在螺线的同一位置的边界处理均是套用相同的模板，本文提出了以空间换时间的模板加载摘要的算法。所谓模板，即螺线上某一位置的关键帧区域在进行边界处理时的 Alpha 值的集合。初次运行程序时，将计算并保存螺线上不同位置的所有关键帧区域的模板。再次运行时，则直接读取文本文件的内容，将模板数据写入内存，直接根据关键帧位置套用对应的模板数据绘制螺旋。实验证明，采用模板加载摘要的算法大大加快了绘制螺旋摘要的时间，实现了平滑的动态缩放过程。

### 3.2　草图注释

螺旋视频摘要是由关键帧按照时序排布而成，但关键帧所能直接传递的信息有限，帧与帧的内在联系无法体现。为了弥补这一缺陷，可在螺旋摘要中额外拓展了草图注释的功能。

所谓草图注释，是指以草图的形式对视频中某一关键帧进行注释。在使用螺旋摘要时，若发现某一关键帧有额外信息需进行补充，可以绘制草图注释并建立起该关键帧和注释的联系，还可以让一个草图注释与多个关键帧相关联，体现不同的关键帧的内在联系。例如，图 4 中标注了多个含有卡车的关键帧，当播放到具有草图注释的关键帧时，草图会自动显现，提示用户相应信息(此处有卡车)。直到播放到下一关键帧时，草图和连线才会消失。

草图注释并非一次性使用。每一次在螺旋摘要上构建好相应的草图注释后，草图的墨迹文件和与之关联的关键帧序号均会被自动保存到本地。当重新打开螺旋摘要时，系统会自动读取本地文件的内容，将之前画好的草图注释及其关联信息写入内存，在播放时予以显示。

### 3.3　目标分布螺旋

螺旋摘要中每一个关键帧的呈现空间是有限的，为了让语义信息能够更加明确，本文提出用目标分布螺旋来反映目标的类型和密集程度。

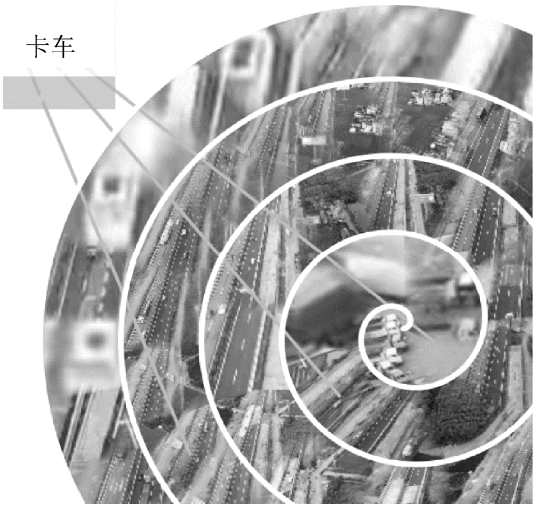



**图 4**　草图注释卡车出现的关键帧

所谓目标分布螺旋，是指在和螺旋摘要的每一关键帧区域一一对应的另一个螺旋上，用绘制实心圆的方式来反映每一关键帧中的目标分布。其中，圆的颜色表示目标的类型，圆的半径表示目标的数量，半径越大，数量越多，如图 5 所示。

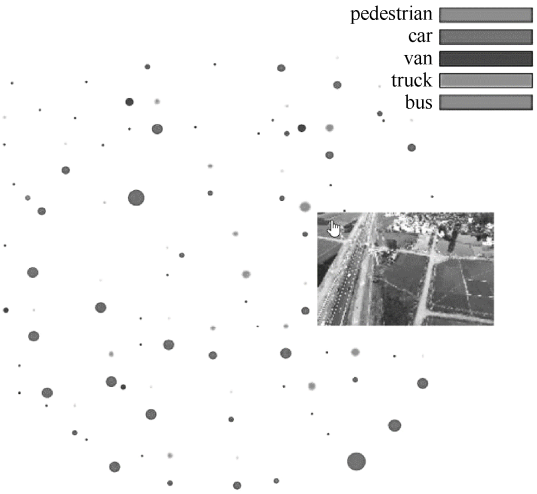



**图 5**　目标分布螺旋

目标分布螺旋呈现的不再是关键帧，而是与关键帧对应的目标检测结果。在绘制实心圆时，本文规定不同目标的圆为不同的颜色，如图例所示。点击任意一种目标的颜色条，即显示出该类别的目标分布。借助目标分布螺旋，用户能够看出哪种目标的数量最多且集中在哪个时间段。和螺旋摘要一样，目标分布螺旋也具备关键帧预览和视频定位的功能。当螺旋摘要的尺度发生变化后，目标分布螺旋的尺度也会有相应的改变。

### 3.4　双螺旋播放

螺旋摘要虽然能最大化利用空间呈现摘要，

但摘要的关键帧时序是由内旋转到外，和人直观的从左到右认知信息的方式存在一定的差异，需要用户对此进行适应。为了保留螺旋摘要的优势，需进一步省去这种认知差异带来的适应时间，本文提出了双螺旋播放的交互形式。

双螺旋是由 2 个镜面对称的单螺旋和直线桥梁搭接而成，如图 6 所示。其整体绘制过程与 3.1.2 节的步骤一致。可将双螺旋看作是一个磁带，直线桥梁的正中央是当前正在播放的关键帧，用一个箭头指示。左螺旋显示的是尚未播放的关键帧，右螺旋显示的是已播放了的关键帧，前后溢出的关键帧会隐藏在左右螺旋的中心处。当视频播放到下一关键帧的时候，双螺旋上所有的关键帧会集体向右移动一个关键帧的距离，实现类似于磁带播放的动态滚动效果，与 3.1.3 节实现螺旋摘要的动态尺度缩放效果的算法一致。

(a) 双螺旋播放-视频开始阶段　　(b) 双螺旋播放-视频结束阶段

**图 6** 双螺旋播放

总体而言，双螺旋和单螺旋都可以视作一种以关键帧为单位的时间轴的变体，都支持基于关键帧的预览和视频定位。但是，双螺旋对称的形式更符合用户的认知习惯，降低了用户的理解和使用门槛。此外，双螺旋正在播放的关键帧始终显示于箭头指示的正中央，也保证了用户能注意到当前播放位置附近的关键内容。

## 4 验　证

对于时长 16 分 3 秒、帧率为 50 帧/秒、分辨率为 1920×1080 的无人机视频，采用基于给定最小/最大间隔帧数的关键帧提取算法，在 i5-8300H CPU 的电脑上进行测试，结果见表 1。可以看出，MinFrames/MaxFrames 保持不变，将 Step 从 3 增加为 30 时，计算耗时整体减少了 47.4%，而最终得到的关键帧数量只减少了 1.41%；Step 保持 30 不变，MinFrames/MaxFrames 从 100/500 降为 100/300 时，计算耗时整体减少了 11.9%，而最终得到的关键帧数目却增加了 20.0%。

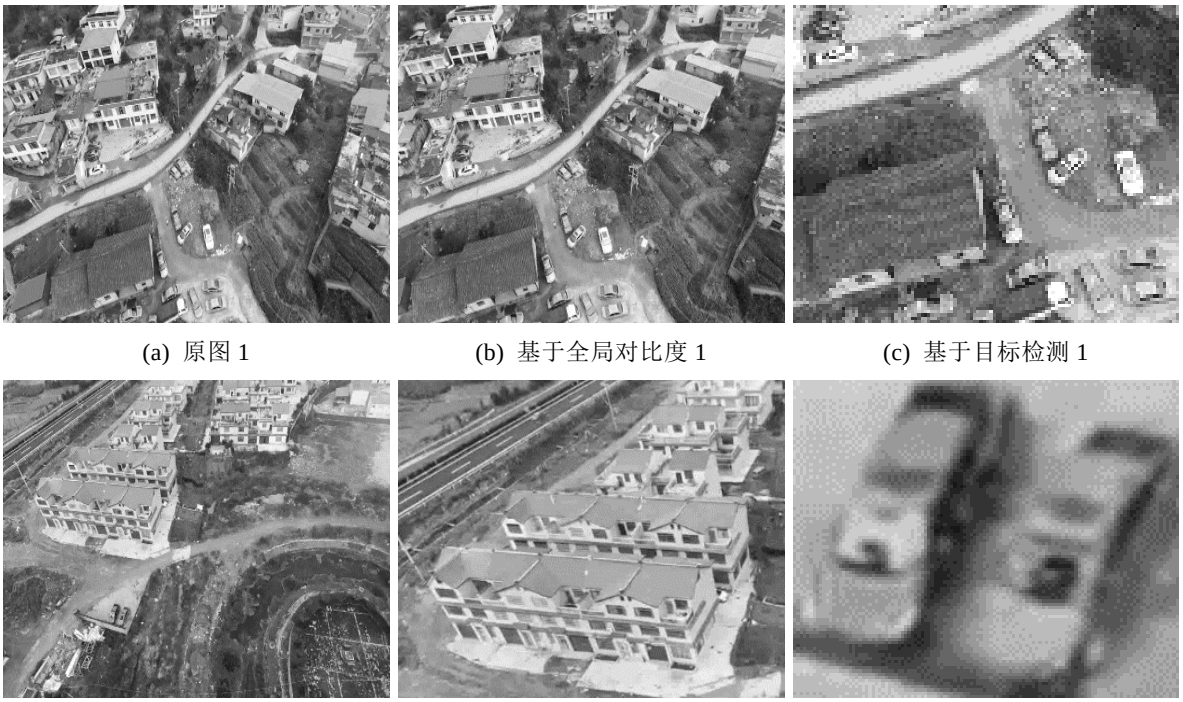
由此得出结论，Step 参数的引入可在对关键帧数目影响不大的情况下，大幅加快计算效率；MinFrames/MaxFrames 参数则能够非常直观而有效地控制关键帧密度，且并不对计算耗时产生较大影响。

本文可以计算基于关键帧的视频目标检测算法相较于对逐帧检测视频的时间优势。该无人机视频总共有 48 150 帧，以提取出 251 张关键帧为例(表 1)。在 Titan X 上，根据输入图片大小的不同，YOLOv3 检测单张图片的耗时在 20~50 ms 之间[2]。假设耗时为 20 ms，在目标检测环节中，基于关键帧的视频目标检测算法节省了 957.98 s；假设耗时为 50 ms，则在目标检测环节节省了 2 394.95 s。从表 1 可以看出，只要控制好 Step 参数，提取关键帧所需的额外耗时与目标检测节省的时间相比，是微不足道的。在实际应用中，GPU 配置往往达不到 Titan X 那样的水平。本文在配置差一些的 GTX 1050Ti 电脑上进行测试，YOLOv3 检测一张 416×416 的图耗时约 60 ms，检测一张 608×608 的图耗时约 120 ms。显然，目标检测计算耗时的差距会更大。而随着无人机视频的时长增加，差距则会被进一步拉大。可见，本文提出的算法对于视频目标检测的效率而言是有着实际意义的。

**表 1　不同参数提取关键帧的数量与耗时对比**

| 给定参数 | 基于颜色筛选的关键帧数量 | 耗时(s) | 基于目标检测筛选的关键帧数量 |
|---|---|---|---|
| Step=3<br>MinFrames=100<br>MaxFrames=500 | 218 | 455.92 | 71 |
| Step=30<br>MinFrames=100<br>MaxFrames=500 | 205 | 239.89 | 70 |
| Step=30<br>MinFrames=100<br>MaxFrames=300 | 251 | 211.38 | 84 |

对于无人机视频中提取的关键帧，分别基于全局对比度的显著性检测算法和 YOLOv3 目标检测算法进行关键区域的提取。从图 7(a)~(c)的对比可以看出，由于原图内容复杂，基于全局对比度的算法几乎判定了整张图均为显著性区域，提取了过多的关键区域，以致于重点不突出；而 YOLOv3 算法则成功检测了图中车辆目标的位置，关键信息一目了然。在图 7(d)~(f)中，基于全局对比度的算法错误地将颜色鲜艳的房屋判定为关键区域；YOLOv3 算法却成功检测到了人的肉眼很难注意到的车辆目标，锁定了图中真正的关键信息。可见，对于无人机视频的关键帧，通过无人机视角的小目标检测引入语义信息，运用基于目标检测的关键区域提取算法，能够取得比传统算法更好的结果。



(a) 原图 1　(b) 基于全局对比度 1　(c) 基于目标检测 1

(d) 原图 2　(e) 基于全局对比度 2　(f) 基于目标检测 2

**图 7**　基于全局对比度和基于目标检测的无人机视频关键帧的关键区域提取对比

## 5　结 束 语

本文以无人机视频的目标检测为中心，设计并实现了一种面向无人机视频的多尺度螺旋摘要。首先，为了获取无人机视频的语义信息，本文基于 VisDrone 数据集和 YOLOv3 目标检测算法，训练了能够检测无人机视角下的行人、车辆等小目标的神经网络模型，具有良好的泛化性。然后，考虑对视频每一帧进行目标检测会耗时较长且检测结果大量重复，本文提出了一种基于关键帧的视频目标检测算法，以关键帧的目标检测结果代表整个视频的检测结果，大幅提高效率。而为了提取视频的关键帧，本文在传统的基于颜色特征的关键帧提取算法的基础上作出改进，提出了一种给定最小/最大间隔帧数的自适应阈值的关键帧提取算法，该算法能够直观地控制关键帧的提取密度和计算效率。然后，将训练好的网络模型应用于筛选出的关键帧上，就得到了整个视频的目标检测结果。以此为基础，为了实现对无人机视频的可视分析，本文根据关键帧的目标检测结果，提取出其中的关键区域作为视频摘要的呈现单元，并以螺旋的形式从内到外地将其呈现，辅以基于关键帧的视频定位和尺度缩放功能。最后，本文开发了基于螺旋摘要的草图注释、目标分布螺旋、双螺旋播放等新颖的交互工具，拓展螺旋摘要的潜力。

整个面向无人机视频的多尺度摘要的设计，是基于无人机视频的目标检测结果的可视交互，是对视频关键信息的提炼和呈现，使得用户能够在短时间内从无人机长视频中高效获取自己感兴趣的信息，因此在无人机地面监控领域具有良好的应用前景。而在视频交互中引入目标检测这样的语义信息，使计算机能够更准确地理解视频画面的含义，同样也是将来人机交互领域的研究方向。

## 参 考 文 献


[1] BARNES C, GOLDMAN D B, SHECHTMAN E, et al. Video tapestries with continuous temporal zoom[J]. ACM Transactions on Graphics, 2010, 29: 4.

[2] LIU Y J, MA C, ZHAO G, et al. An interactive spiraltape video summarization[J]. IEEE Transactions on Multimedia, 2016, 18(7): 1269-1282.

[3] REDMON J, FARHADI A. Yolov3: an incremental improvement[EB/OL]. [2019-12-10]. http://xueshu.baidu.com/usercenter/paper/show?paperid=e02671f7b0527c6ecee43ce8bd7918b6&site=xueshu_se&hitarticle=1.

[4] WOLF W. Key frame selection by motion analysis[C]//1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. New York: IEEE Press, 1996: 1228-1231.

[5] ZHANG H J, WU J, ZHONG D, et al. An integrated system for content-based video retrieval and browsing[J]. Pattern Recognition, 1997, 30(4): 643-658.

[6] ZHUANG Y, RUI Y, HUANG T S, et al. Adaptive key frame extraction using unsupervised clustering[C]// Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269). New York: IEEE Press, 1998: 866-870.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems. 2012, 25(2): 1097-1105.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2019-12-10]. http://xueshu.baidu.com/usercenter/paper/show?paperid=2801f41808e377a1897a3887b6758c59&site=xueshu_se.

[9] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 770-778.

[10] REN S, HE K, GIRSHICK R, et al. Fasterr-CNN: towards real-time object detection with region proposal networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. New York: IEEE Press, 2017: 1137-1149.

[11] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 779-788.

[12] XU H J, DAS A, SAENKO K. R-C3D: region convolutional 3D network for temporal activity detection[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2017: 5783-5792.

[13] CHAO Y W, VIJAYANARASIMHAN S, SEYBOLD B, et al. Rethinking the faster R-CNN architecture for temporal action localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 1130-1139.

[14] DRAGICEVIC P, RAMOS G, BIBLIOWITCZ J, et al. Video browsing by direct manipulation[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2008: 237-246.

[15] GOLDMAN D B, GONTERMAN C, CURLESS B, et al. Video object annotation, navigation, and composition[C]//Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology. New York: ACM Press, 2008: 3-12.

[16] ZHU P F, WEN L Y, BIAN X, et al. Vision meets drones: a challenge[EB/OL]. [2019-12-10]. https://arxiv.org/abs/1804.07437.

[17] CHENG M M, ZHANG G X, MITRA N J, et al. Global contrast based salient region detection[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2011: 409-416.