A Semi-automatic Tool for Detection and Tracking Ground Truth Generation in Videos

I. Kavasidis
Dep. Electrical, Electronics
and Computer Engineering
University of Catania, Italy
kavasidis@dieei.unict.it

S. Palazzo
Dep. Electrical, Electronics
and Computer Engineering
University of Catania, Italy
spalazzo@dieei.unict.it

R. Di Salvo
Dep. Electrical, Electronics
and Computer Engineering
University of Catania, Italy
rdisalvo@dieei.unict.it

D. Giordano
Dep. Electrical, Electronics
and Computer Engineering
University of Catania, Italy
dgiordan@dieei.unict.it

C. Spampinato
Dep. Electrical, Electronics
and Computer Engineering
University of Catania, Italy
cspampin@dieei.unict.it

ABSTRACT

In this paper we present a tool for the generation of ground-truth data for object detection, tracking and recognition applications. Compared to state of the art methods, such as ViPER-GT, our tool improves the user experience by providing edit shortcuts such as hotkeys and drag-and-drop, and by integrating computer vision algorithms to automate, under the supervision of the user, the extraction of contours and the identification of objects across frames. A comparison between our application and ViPER-GT tool was performed, which showed how our tool allows users to label a video in a shorter time, while at the same time providing a higher ground truth quality.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; I.4.9 [**Image Processing and Computer Vision**]: Applications

Keywords

Object Detection, object tracking, ground truth data, video labeling

1. INTRODUCTION

In the last decade, the advancements in camera technology and the reduction of costs have led to a widespread increase in the number of applications for automatic video analysis, such as video surveillance [1, 2], real-life study of animal species behaviour [3]. For all of these purposes, the scientific community has put a lot of effort in the development of algorithms for object detection [4], tracking [5] and recognition [6]. Of course, one of the most important stages in the development of such algorithms is the evaluation of accuracy and performance. Because of the varying nature of the targetted visual environments, it is very difficult – if not impossible – to devise an algorithm which is able to perform very well at all possible scene conditions (i.e. open/closed area, different objects' motion patterns, scene lighting, background activity, etc). For this reason, it is often necessary to establish the suitability of an algorithm to a specific application context by comparing its results to what are expected to be the correct results. The availability and generation of such "correct results", also known as *ground truth*, is therefore an essential aspect in the evaluation process of any low- and high-level computer vision technique.

Unfortunately, the ground-truth generation process presents several difficulties. In the context of object detection, segmentation, tracking and recognition, ground truths typically consist of a list of the objects which appear in every single frame of a video, specifying for each of them information such as the bounding box, the contour, the recognition class and the associations to other appearances of the same object in the previous or following frames. The manual generation of ground truths by a user is therefore a time-consuming, tedious and error-prone task, since it requires the user to be focused on drawing accurate contours and handling tracking information between objects.

In order to support users in tackling this task, several software tools have been developed to provide them with a graphical environment which helps drawing object contours, handling tracking information and specifying object metadata.

One of the most used application for this purpose is ViPER-GT [7], which produces an XML file containing all video metadata information inserted by the user, and provides a user interface with a spreadsheet representation of objects' data, timeline panels to navigate the video and view objects' life span, and metadata propagation features across multiple frames. Although ViPER is widely used, it lacks sup-

VIGTA '12 May 21 2012, Capri, Italy
Copyright 2012 ACM 978-1-4503-1405-3/12/05 ...$10.00.

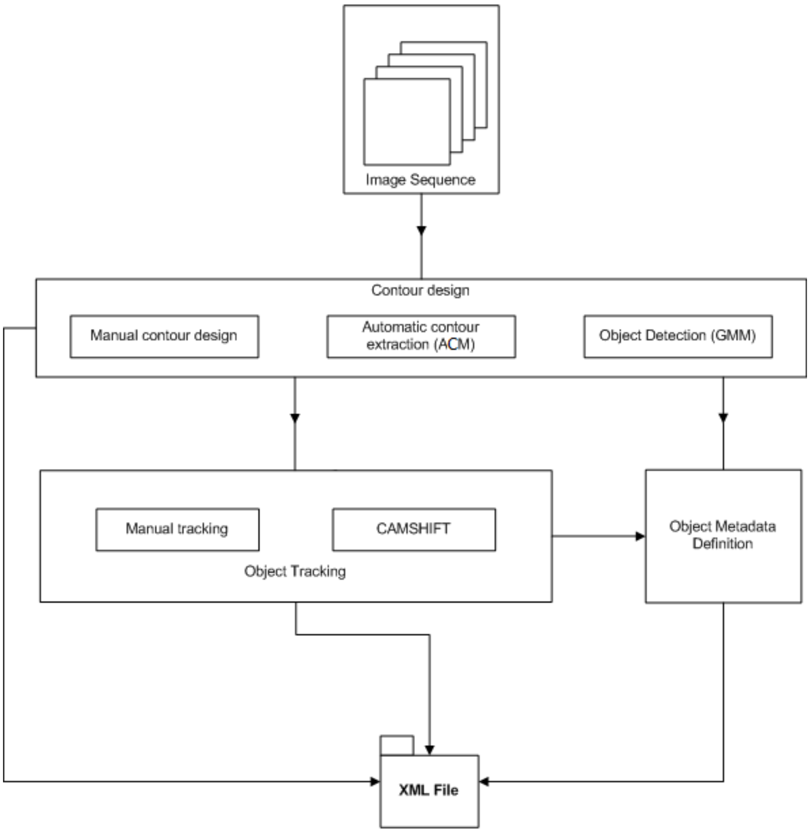


Figure 1: Ground truth generation flowchart

port for automatic or semi-automatic processing, which can be implemented by adding a basic object detection/tracking algorithm to give hints to the user about likely object locations or tracking associations (although, of course, user supervision is still required to guarantee the correctness of the results).

In [8], the authors propose a ground-truth generation tool which employs simple object detection and tracking algorithms to retrieve object's bounding boxes and associate them across frames, however allowing the user to add, delete or resize the bounding boxes and edit the associations.

The *GTVT* tool, described in [9], aims at improving the user experience with respect to ViPER, although it focuses on object detection and classification, rather than segmentation.

In [10], a web-based collaborative annotation tool is described, which is focused on object classification, and integrates a prediction algorithm which tries to infer the expected class of an object by comparing it with previously classified items.

The application described in this paper, called *GTTool*, aims at:

- Providing an easy-to-use interface, specific for generating ground truths for object detection, segmentation, tracking and classification.

- Improving the user experience with respect to ViPER, by showing two panels, each containing a frame at a different time, thus allowing the user to compare a frame's annotations with those from a previous or following frame, and providing quick methods to specify object associations and perform attribute propagation (e.g. hotkeys, drag-and-drop).

- Integrating automatic tools for object segmentation and tracking, effectively reducely the number of objects/frames to be manually analyzed.

- Supporting ViPER XML format, for ground truth importation.

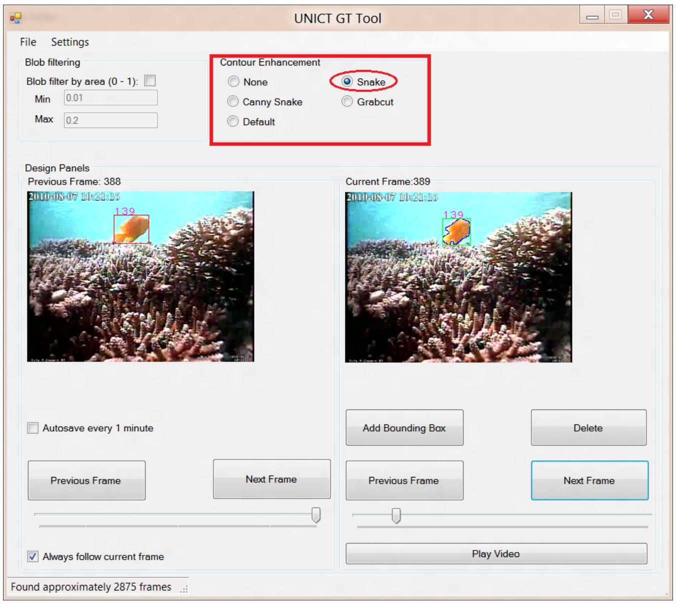
In Section 3, we show the comparison between GTTool and ViPER in the generation of ground truth for a video file. Our evaluation approach is based on a comparison of the time required to label the video with each tool and on an accuracy analysis of the generated contours, compared with those obtained from a higher resolution version of the videos.

The rest of the paper is organized as follows: Section 2 describes in detail our application's features and user interface; Section 3 shows a performance and accuracy comparison between our GTTool and ViPER; finally, Section 4 draws some conclusion remarks on our tool and its possible future developments.

2. GTTOOL

2.1 General description

The proposed tool relies on a modular architecture (Fig. 1) which allows users to define the ground truth by using an



Blob filter by area (0 - 1): □
Min 0.01
Max 0.2

○ None ● Snake
○ Canny Snake ○ Grabcut
○ Default

Previous Frame: 388
Current Frame:389

□ Autosave every 1 minute

☑ Always follow current frame

Found approximately 2875 frames

Figure 2: GUI for automatic contour extraction

easy graphical user interface (GUI). The developed application integrates a number of computer vision techniques, with the purpose of enhancing the ground-truth generation process in terms of both accuracy and human effort. In particular, Active Contour Models (ACM) are integrated to automatically extract objects' contours; and object detection algorithms and state-of-the-art edge detection techniques are employed in order to suggest to the user the most interesting shapes in the frame. Moreover, by using a two-window GUI layout, the application enables the user to generate tracking ground truth through straightforward drag-and-drop and context-menu operations. The user can also open previous ground-truth XML files in order to add new objects or edit the existing ones and save the performed improvements to the same or a new file.

2.2 Automatic contour extraction

In order to make ground truth generation faster, automatic contour extraction techniques have been integrated. In particular, when the object's boundaries can be clearly identified (i.e. the object's border colors differ substantially from the background in its vicinity), the application is able to automatically extract the object's contour by using one of the following methods:

- Snakes [11];
- GrabCut [12];
- Snakes with Canny contour enhancement.

To accomplish this, the user has to draw just a bounding box containing the whole object and choose one of the available techniques for automatic contour extraction from the corresponding panel (Fig. 2).

2.3 Manual contour extraction

As in nearly every common ground-truth generation application, the developed tool allows the user to draw ground truths manually by using the pencil tool or the polygon tool to trace the contour of an object of interest. Though slow and tedious to the user, the usage of these tools is often necessary, because the automatic contour extraction methods may fail to segment correctly the objects of interest.

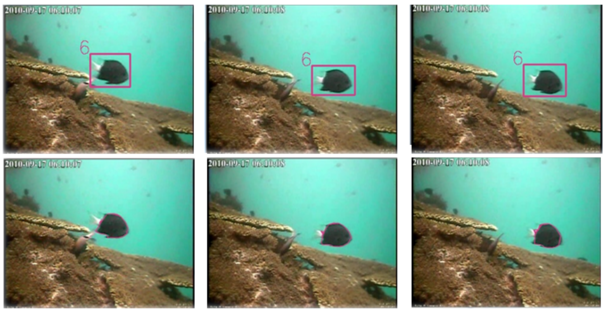


Figure 3: Automatic Object Tracking and detection: In the top row the output of the tracker is shown, while in the bottom row the output of the automatic detection module is shown.

2.4 Automatic object detection

While automatic contour extraction allows the user to extract object contours in an automatic and easy way, object detection aims at identifying possible interesting objects, and to do so the Gaussian Mixture Model algorithm (GMM) [13] is employed. At each new frame, the GMM algorithm detects moving objects and allows the user to automaticaly add the detected objects' contour to the generated ground truth by using the object's context menu (Fig. 3). Because the GMM algorithm needs to be initialized with an adequate number of frames, this method performs progressively better in later stages of long video sequences.

2.5 Automatic Object Tracking

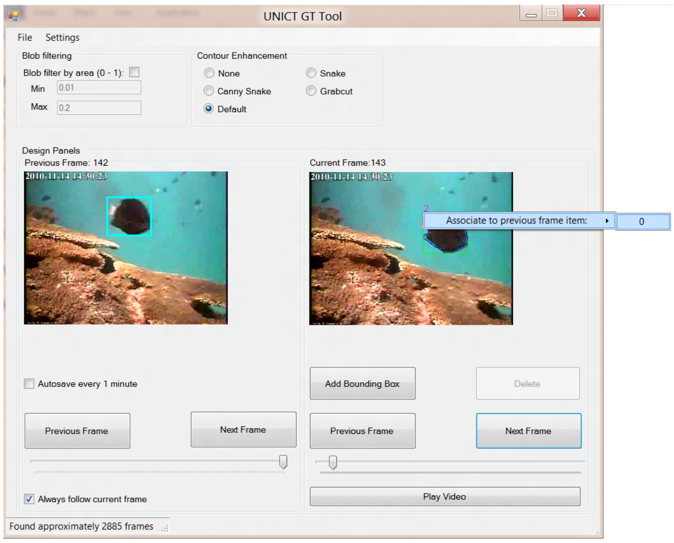
In conjuction with the GMM algorithm, CAMSHIFT [14] is used to generate automatic object tracking ground-truth data. The algorithm takes as input the objects identified in the previous frames and suggests associations with the objects localized (either manually or automatically) in the current frame (Fig. 3). As in the case of automatic object detection, the user is always given the choice to accept or refuse the suggested associations.

2.6 Manual Object Tracking

As aforementioned, the two-window GUI layout makes the task of creating tracking ground truth easier to the user. The right window always shows the current frame, while in the left window the user can select one of the previously labelled image. By using the right window's objects' context menus, the user can specify the associations to the objects in the left window (Fig. 4).

2.7 Metadata Definition

Besides object segmentation and tracking, it is possible to add arbitrary metadata, such as for classification purposes, by defining labels and assigning values to each object. When used in conjunction with tracking, these metadata are automatically propagated across all instances of an object, thus requiring the user to specify them only once.



Blob filtering

Blob filter by area (0 - 1): □

Min 0.01

Max 0.2

Contour Enhancement

○ None ○ Snake
○ Canny Snake ○ Grabcut
● Default

Previous Frame: 142

Current Frame:143

□ Autosave every 1 minute

☑ Always follow current frame

Found approximately 2885 frames

Figure 4: Manual Object Tracking

```
<?xml version="1.0" ?>
<video name="test_video">
    <frame id="0">
        <object trackingId="0" instanceId="0">
            <bounding_box>
                <point x="170" y="189" />
                <point x="170" y="230" />
                <point x="206" y="230" />
                <point x="206" y="189" />
                <point x="170" y="189" />
            </bbox>
            <contour>
                <point x="17" y="17" />
                ...
            </contour>
            <metadata>
                <label name="class">pedestrian</label>
                <label name="action">walk</label>
                ...
            </metadata>
        </object>
        ...
```

Figure 5: Example of GTTool's output XML file.

2.8 XML output and ViPER file importation

The set of annotations added to a video can be exported to an XML file, for example to simply store it or to share it with others. An example of the XML format we use is shown in Fig. 5. In order to make the adoption of GTTool easier to ViPER users, the application allows to import and convert ViPER files to GTTool's schema, so no loss of previous work occurs when switching from the former to the latter.

3. EXPERIMENTAL RESULTS

In order to assess the performance of the proposed tool in terms of time and accuracy, we asked 20 users to annotate fish in 100 consecutive frames of 10 different videos taken from underwater cameras (resulting in 20000 annotated frames), with both GTTool and ViPER. The users were asked not only to draw the boundaries of the objects, but also to create tracking ground truth by using the tools offered by the two applications. The achieved results in terms of efficiency and accuracy are shown in Table 1. The accuracy of the segmented objects was computed by evaluating the overlap ratio with ground-truth data drawn by experts on higher-resolution versions of the same videos.

Method	GTTool	ViPER
Total drawn objects	16347	13315
Manually drawn objects	3114	13315
Automatically drawn objects (GMM)	8101	-
Automatically drawn objects (ACM)	5132	-
Average time per object	4,8 seconds	13,7 seconds
Accuracy	91%	76%
Learnability	8.4	3.2
Satisfaction	7	5.1

Table 1: Comparison between the proposed tool and ViPER.

As can be seen from the results, the time required to analyze manually the videos with GTTool is about one third of the time needed to perform the labeling task by using ViPER. This was mainly due to the markedly smaller number of objects which had to be drawn manually by the users (about 3 objects out of 4 are automatically segmented by our tool).

We also asked users to fill in a usability questionnaire [15], in order to get their feedback on how they felt using the two tools. In particular, we asked the participants to grade both tools in terms of learnability and satisfaction. Learnability represents the ease of learning the usage the tools, while satisfaction represents the subjective feelings of the users about their experience with each tool; both values range from 1(worst) to 10 (best). The results show that their experience with GTTool was more satisfactory than with ViPER, mainly, according to most comments, because of the two-window layout (which avoids having to go back and forth through the video to check one's previous annotations) and of the integrated algorithms (which drastically reduced the number of frames and objects which had to be manually analyzed).

4. CONCLUDING REMARKS

In this paper a novel tool for ground truth generation is presented. The main contribution of the proposed application is the improvement of the user's experience during the extraction of contours by means of a simple and intuitive graphic interface and the use of automatic techniques for the detection of objects across frame sequences. A modular architecture has been developed in order to enhance ground truth generation in terms of both accuracy and human efforts. Several techniques for automatic contour extraction (Active Contour Models and the Gaussian Mixture Model motion detection algorithm) and object tracking (CAMSHIFT) have been integrated, while still allowing the user to define ground-truth data manually if the automatic methods fail to identify and track correctly the objects of interest. XML support allows to both save the inserted ground truth to file (to share it with others or to be modified at a later time) and to import ViPER files, thus supporting the migration process to GTTool. The experimental results show that the proposed solution outperformed ViPER in every test we ran, reducing the time needed to label an entire video by a factor of 3.

Some suggestions for future developments would be the integration of crowdsourcing and collaborative capabilities in order to permit to different users to collaborate in the ground truth generation process. This will be achieved by providing a web interface that will implement the same functionalities of GTTool, adding multi-user capabilities and video library management. Moreover, clustering techniques could be applied to the automatic object detection and tracking results in order to automatically insert metadata information for the detected objects.

5. ACKNOWLEDGEMENTS

This research was funded by European Commission FP7 grant 257024, in the Fish4Knowledge project[1].

6. REFERENCES

[1] M.-Y. Liao, D.-Y. Chen, C.-W. Sua, and H.-R. Tyan, "Real-time event detection and its application to surveillance systems," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 2006.

[2] A. Faro, D. Giordano, and C. Spampinato, "Soft-computing agents processing webcam images to optimize metropolitan traffic systems," in *Computer Vision and Graphics*, ser. Computational Imaging and Vision, K. Wojciechowski, B. Smolka, H. Palus, R. Kozera, W. Skarbek, and L. Noakes, Eds. Springer Netherlands, 2006, vol. 32, pp. 968–974.

[3] C. Spampinato, J. Chen-Burger, G. Nadarajan, and R. Fisher, "Detecting ,tracking and counting fish in low quality unconstrained underwater videos," in *Proc. 3rd Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, 2008, pp. 514–520.

[4] A. Faro, D. Giordano, and C. Spampinato, "Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 4, pp. 1398 –1412, dec. 2011.

[5] C. Spampinato, S. Palazzo, D. Giordano, I. Kavasidis, F. Lin, and Y. Lin, "Covariance based fish tracking in real-life underwater environment," in *International Conference on Computer Vision Theory and Applications, VISAPP 2012*, 2012, pp. 409–414.

[6] C. Spampinato, D. Giordano, R. D. Salvo, Y. J. Chen-Burger, R. B. Fisher, and G. Nadarajan, "Automatic fish classification for underwater species behavior understanding," in *First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*, 2010, pp. 45–50.

[7] D. Doerman and D. Mihalcik, "Tools and techniques for video performance evaluation," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 4, 2000, pp. 167–170.

[8] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo, "A semi-automatic system for ground truth generation of soccer video sequences," in *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, Genova, 2009, pp. 559–564.

[9] A. Ambardekar and M. Nicolescu, "Ground Truth Verification Tool (GTVT) for Video Surveillance Systems," in *Advances in Computer-Human Interactions, 2009. ACHI '09. Second International Conferences on*, Cancun, 2009, pp. 354–359.

[10] C. Lin and B. Tseng, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets," *Proceedings of the TRECVID 2003*, 2003.

[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[12] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[13] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 246–252, 1999.

[14] G. R. Bradski, "Computer Vision Face Tracking For Use in a Perceptual User Interface," *Intel Technology Journal*, pp. 1–15, 1998.

[15] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '88. New York, NY, USA: ACM, 1988, pp. 213–218.

[1] www.fish4knowledge.eu