First International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications

Concetto Spampinato
DIEEI
University of Catania, Viale
Andrea Doria, 6 - 95125
Catania, IT
cspampin@dieei.unict.it

Bas Boom
School of Informatics
University of Edinburgh , 10
Crichton St
Edinburgh, EH8 9AB, UK
bboom@inf.ed.ac.uk

Jiyin He
Centrum Wiskunde &
Informatica
Science Park 123
1098 XG Amsterdam, NL
jiyin.he@cwi.nl

ABSTRACT

The goal of the First International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications is to bring together practitioners and researchers in computer vision and in HCI to share ideas and experiences in designing and implementing visual interfaces for ground truth data generation.

It specifically presents and reports on the construction and analysis of user-oriented tools and interfaces to support automatic or semi-automatic ground truth annotation and labeling in many applications such as object detection, object recognition, scene segmentation and face recognition both in still images and in videos.

Categories and Subject Descriptors

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

General Terms

Image labeling, Ground truth data, Visual Interfaces, Collaborative Interfaces, Object Detection, Pattern Recognition

1. INTRODUCTION

The importance of having image/video database containing high quality ground truth annotations generated by humans for a variety of computer vision applications is recognized by the whole machine vision community. Indeed, one of the most significant efforts during the evaluation process is represented by the development of accurate truth and comparing this truth to the decision of image and video processing applications. For example, datasets with ground truth labels are necessary for supervised learning of object categories. However, the cost of providing labelled data, which implies asking a human to examine images and provide labels, becomes impractical as training sets grow.

The computer vision community has mainly directed its attention to develop methods for collecting large scale datasets by exploiting the collaborative effort of a large population of annotators. Although there exist functional, task-oriented requirements for tools supporting ground truth labeling, the research still lacks in following human computer interaction paradigm to develop user-oriented tools. In fact, annotators, must be at the centre of such tools since the ground truth needs to be established by humans who, for example, on a video sequence of five minutes captured at 30 fps, have to annotate manually a total of 9000 frames.

Having this in mind, the Program Committee (PC) selected 14 papers describing tools, interfaces and methods able to speed-up the process of effective ground truth creation by helping users, through the integration of computer vision and HCI methods, to finish the task with the best accuracy in a reasonable amount of time.

In the next two sections a summary of the accepted papers is presented, whereas in the last section concluding remarks are given.

2. FULL PAPERS

Six papers were accepted as full paper. Two of them specifically address the problem of generation of large scale ground truth starting from small datasets:

- "Efficient Annotation of Image Data Sets for Computer Vision Applications" presents an image annotation approach that aims at supporting users with little-to-know knowledge of computer vision techniques to provide application-specific labels for large data sets. The proposed approach arranges images in clusters using self-organizing map (SOM), and the user interface allows users to select and annotate large number of images at once. This tool improves labeling efficiency compared to a Wizard-like annotation interface.

- "Combinatorial Enlargement of Ground-Truth Datasets and Efficient Evaluation of Segmentation Algorithms" describes a evaluation approach that aims at efficiently comparing different image segmentation algorithms and meanwhile generating ground truth from a small labeled data set. Given an initial set of images with segmentation labels, synthetic images can be generated by combinatorially composing labeled regions from images in the initial set. By applying combinatorial design algorithms, the authors demonstrated that it is

AVI '12 May 21-25, 2012, Capri Island, Italy
Copyright 2012 ACM 978-1-4503-1287-5/12/05 ...$10.00.

possible to create an exponentially larger database of valid images than the starting set.

The other four full papers instead describe user-oriented tools supporting annotators mainly in the task of object detection, recognition and image segmentation in still images and in video streams.

- "Multiscale Annotation of Still Images with GAT" presents a graphical annotation tool for still images that works at both global (entire image) and local (fragment) scales, and the latter both in rough/precise form and with and without semi-automatic segmentation. Moreover, the presented interface assists users in the annotation of images with relation to the semantic classes described in an ontology. The annotation capabilities are also complemented with additional functionalities that allow the creation and evaluation of an image classifier.

- "Synthetic Ground Truth Dataset to Detect Shadows Cast by Static Object in Outdoors" proposes a technique for the generation of ground truth for shadow-detection algorithms, which is based on the use of a rendering software to create a synthetic scene and simulate the flowing of time by accurately computing the sun's position, which is the main source of light. The results shown that the approach drastically reduces the time for the generation of a ground-truth with object shadows, since no manual user intervention is required.

- "An Interactive Tool for Extremely Dense Landmarking of Faces" describes a tool for generating ground truthed landmarks in face images that is able to improve manual landmarking in the case of extremely dense (250+ points per images) annotation of face images with a 2x speedup. A case study suggests also that this sub feature based annotation approach is more efficient than holistic based approach such as the Cootes' tool.

- "A Multi-View Annotation Tool for People Detection Evaluation" introduces a multi-view annotation tool for generating 3D ground truth data of the real location of people in the scene. In order to achieve precise ground truth data the user is also aided by video frames of multiple synchronized and calibrated cameras.

3. POSTER PAPERS AND DEMO

The six papers accepted as poster paper were mainly "work in progress" showing and sharing ideas to build up interfaces for ground truth labeling in several image processing tasks.

- "CoVidA: Pen-Based Collaborative Video Annotation" proposes a pen-based interface which combines pen and touch input to annotate videos. The authors demonstrated also that especially for complex structures the usage of CoVidA device improves the effectiveness of the outlining process.

- Two papers instead propose tools to deal with the annotations of medical images: "An annotation tool for dermoscopic image segmentation" and "Manual labeling strategy for Ground Truth estimation in MRI Glial Tumor Segmentation". Both works share the idea to involve domain specialists during the development stages of the tools.

- "Robust Interactive Segmentation via Coloring" introduces an original idea to support annotation for image segmentation on mobile devices. In this setting, requirements are to have lightweight segmentation methods and simple touch or gesture based input methods. Given these constraints the authors use existing methods to perform the segmentation, but introduce a novel error correction method for this setting and dynamic segment indication (labeling) methodology, i.e., the image is segmented while a user "colors" (draws in) the image. A user based evaluation, carried out on 15 subjects with capacitive mobile touch screen, showed the proposed method outperforms two existing methods (graphcut and intelligent scissors) in both segmentation performance and easiness and entertainment.

- In "Manually fitting a 3D skeleton to multi-view video" the authors propose a interactive interface to label the skeleton on the videos, where they allow users to manipulate the skeleton from different angles. The accuracy of the annotations is evaluated against the HumanEva dataset showing that their tool does better than the baseline line method of the dataset.

- "A Multi-Tool for Ground-Truth Stereo Correspondence, Object Outlining and Points-of-Interest Selection" describes a tool that supports three types of manual annotation tasks in creating ground truth for computer vision applications: (1) matching the corresponding points in multi-view/stereo processing, (2) outlining objects in images, and (3) collecting sample points from regions of interest. The annotations collected using this tool can be useful in various types of computer vision tasks.

The two demo papers present tools to support annotation for object detection and tracking in videostreams by integrating instruments and methods, such as a jog shuttle wheel or a simple detection algorithm, to simplify and speed up the labeling tasks to the end users.

4. CONCLUDING REMARKS

The First International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications represents the first attempt to bring together researchers in computer vision and in HCI to share ideas in designing and implementing user-oriented interfaces for ground truth data generation. However, this aim was not fully reached since most papers addressed mainly computer vision needs, although a first endeavour of integration between computer vision and human computer interaction research is evident. Therefore, there is still room for improvement and we hope that the selected papers will serve as valuable reference for the future research on user-oriented ground truth data generation interfaces in computer vision tasks.

5. ACKNOWLEDGEMENTS

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

[1] www.fish4knowledge.eu

6. REFERENCES

[1] Ambardekar, A., Nicolescu, M., and Dascalu, S. Ground truth verification tool (gtvt) for video surveillance systems. In *Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions* (Washington, DC, USA, 2009), ACHI '09, IEEE Computer Society, pp. 354–359.

[2] Bashir, F., and Porikli, F. Performance evaluation of object detection and tracking systems. In *In PETS* (2006).

[3] Bertini, M., D'Amico, G., Ferracani, A., Meoni, M., and Serra, G. Sirio, orione and pan: an integrated web system for ontology-based video search and annotation. In *Proceedings of the international conference on Multimedia* (New York, NY, USA, 2010), MM '10, ACM, pp. 1625–1628.

[4] Canini, M., Li, W., Moore, A., and Bolla, R. Gtvs: Boosting the collection of application traffic ground truth. In *Traffic Monitoring and Analysis*, M. Papadopouli, P. Owezarski, and A. Pras, Eds., vol. 5537 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 54–63.

[5] Constantine, L. L., and Lockwood, L. A. D. *Software for use: a practical guide to the models and methods of usage-centered design*, vol. 32. Addison-Wesley, 1999.

[6] Doermann, D., and Mihalcik, D. Tools and Techniques for Video Performances Evaluation. In *ICPR* (2000), vol. 4, pp. 167–170.

[7] D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., and Mazzeo, P. L. A semi-automatic system for ground truth generation of soccer video sequences. In *Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (Washington, DC, USA, 2009), AVSS '09, IEEE Computer Society, pp. 559–564.

[8] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision 88*, 2 (June 2010), 303–338.

[9] Lazarevicmcmanus, N., Renno, J., Makris, D., and Jones, G. An object-based comparative methodology for motion detection based on the F-Measure. *Computer Vision and Image Understanding 111*, 1 (July 2008), 74–85.

[10] Li, X., Aldridge, B., Fisher, R. B., and Rees, J. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In *ISBI* (2011), pp. 1438–1441.

[11] Mohd, M., Crestani, F., and Ruthven, I. Design of an interface for interactive topic detection and tracking. In *Flexible Query Answering Systems*, T. Andreasen, R. Yager, H. Bulskov, H. Christiansen, and H. Larsen, Eds., vol. 5822 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 227–238.

[12] Ralph, S. K., Irvine, J., Stevens, M. R., Snorrason, M., and Gwilt, D. Assessing the performance of an automated video ground truthing application. In *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 202–207.

[13] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision 77*, 1-3 (Oct. 2007), 157–173.

[14] Saund, E., Lin, J., and Sarkar, P. Pixlabeler: User interface for pixel-level labeling of elements in document images. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition* (Washington, DC, USA, 2009), ICDAR '09, IEEE Computer Society, pp. 646–650.

[15] Vittayakorn, S., and Hays, J. Quality assessment for crowdsourced object annotations. In *Proceedings of the British Machine Vision Conference* (2011), BMVA Press, pp. 109.1–109.11.

[16] Yao, B., Yang, X., and Zhu, S.-C. Introduction to a Large-Scale General Purpose Ground Truth Database: Methodology, Annotation Tool and Benchmarks. 2007, pp. 169–183.

[17] yung Lin, C., Tseng, B. L., and Smith, J. R. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *In Proceedings of the TRECVID 2003 Workshop* (2003).