A software suite for large-scale video- and image-based analytics

Jasmin Léveillé
jalev51@gmail.com

Isao Hayashi
Faculty of Informatics
Kansai University
Takatsuki, Osaka 569-1095
Japan
ihaya@cbii.kutc.kansai-u.ac.jp

ABSTRACT

The recent proliferation of large video and image databases offers great potential for gaining a deeper understanding about changes in our environment. Although increased use and accessibility of cameras has reduced the difficulty and cost of collecting and storing large video datasets, software tools that support the mining and synthesis of information from these data are largely still lacking. We present a software suite intended to fill that gap; offering a user-friendly scalable database for image and video metadata storage, mining, and synthesis. This demonstration illustrates the user interface and overall workflow when using our software in a typical video analytics scenario.

Keywords

Data science, video mining, video analytics, database

1. INTRODUCTION

Video-based analytics is loosely defined as the application of data mining and machine learning techniques to video data in order to extract useful information. Common examples include the use of anomaly detection or crowd monitoring algorithms on video surveillance data [1, 2, 3]. Although a lot of progress in video analytics is built on methods that work on individual frames, a growing number of these techniques attempt to harness the dynamic information that only multi-frame video data can offer (e.g., gait, etc.; [7, 8]). This progress in video analytics has been made in recent years due in large part to the increased availability of (1) higher performing computers with faster, multithreaded CPUs and GPUs (2) cheaper camera sensors, and (3) large amounts of *labelled* image data [4]. Several image datasets are publicly available (e.g. [5, 6]) that contain sufficient information to train machine learning methods that depend on single-frame input. Extensive video databases are publicly available but often lack the type of information needed to develop multi-frame techniques. This is in part because producing video annotations is a costly process. The software suite we introduce aims to facilitate the creation, maintenance and use of image and video metadata.

2. RELATED WORK

A number of tools exist that can be used to produce video annotations. Anvil [9] is a popular tool used in human communication research that allows users to enter both spatiotemporal and non-spatial data, uses SQL to handle data manipulations while annotations are being performed, and allows some integration with user-defined plugins. OpenSHAPA [10] is intended for a similar use and supports plugin development, but does not offer an easy way to enter spatiotemporal annotations. LabelMe video was essentially a generalization of the LabelMe image annotation tool but doesn't appear to be maintained anymore [11]. VATIC [12] is a software that can harness Amazon Mechanical Turk to crowd-source batch annotation jobs where annotations correspond to tracks of bounding boxes. VATIC uses the Hungarian algorithm in order to aggregate tracks produced by different annotators, but still requires an additional level of quality control to ensure (1) the quality of tracks produced by the multiple annotators and (2) the appropriate matching of tracks as performed by the Hungarian algorithm. ViPER-GT [13] is another tool with similar scope and features.

The above tools suffer from the following shortcomings:

1. **Weak database integration**. This means that there is no easy way to accumulate, maintain, and modify large sets of metadata as the latter often ends up in different locations and formats, sometimes partially duplicated.
2. **Limited extensibility**. Modifying the tools to produce custom spatiotemporal annotations (e.g., key points, polygons) requires a lot of work.

3. DESIGN PRINCIPLES

The main goal of the software suite is to alleviate the above problems while offering additional capabilities to facilitate scaling up to larger datasets and making the process of annotating less difficult. Figure 1 illustrates a typical workflow involving the various components of the system. Accordingly, a user enters metadata through a GUI linked to a database. The stored metadata can then be queried through filters – a light-weight code interface – that link against the same database in order to be used in application-specific code. Although this workflow is rather simple, several key design factors contribute to make our implementation particularly useful for tasks that require video- and image-based analytics:

1. **Scalability**. The implementation relies on a modern document database that supports *horizontal* scaling (i.e. dividing data into shards), facilitates splitting and merging data across databases as well as remote access.

BICT 2014, December 01-03, Boston, United States

DOI 10.4108/icst.bict.2014.257778

2. **Extensibility**. Two factors contribute toward making the implementation extensible: the use of a schema-less, NoSQL database system, and support for user-defined plugins in the GUI component for spatiotemporal annotations. Basic plugins allow easily entering the typical bounding box annotations available in VATIC software, object contours available in the LabelME video tool, text annotations available in OpenSHAPA and Anvil, as well as object key points – that can be used to train object detectors [15] – and event annotations.
3. **Increased Processing Speed**. The user interface also incorporates features meant to accelerate the production of spatiotemporal annotations. These features include interpolation between frames, use of custom detectors, and fast navigation even in long, high-resolution movies.
4. **Video-based schema**. Although NoSQL databases make few constraints on the data schema, we constrain the schema to make storing and querying operations efficient for video metadata. In particular, we implement a set of requirements identified in [14] as critical for video metadata schemes, and further complement it with one additional requirement, namely, the grouping of *observations* into observation types. Nevertheless, users can choose to exploit the unconstrained nature of NoSQL in order to extend the annotation schema with their own metadata requirements.
5. **Compatibility with image sources.** The software can also be used to handle image (i.e. single-frame) metadata.

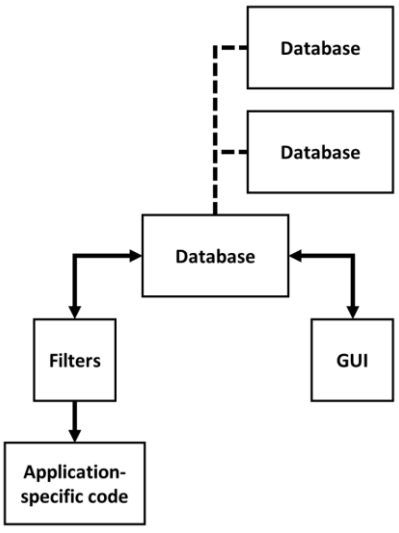

Figure 1. Overall workflow. A user enters metadata into one of several databases and can then use that information through filters in application-specific code.

4. SUMMARY

In this demonstration we present a software suite for large-scale video- and image-based analytics. It is designed to make video and image processing more efficient mine, analyze, and integrate with other types of data.

5. REFERENCES

[1] Lempitsky, V. and Zisserman, A. 2010. Learning to count objects in images. Advances in Neural Information Processing Systems, 23, 1324-1332.

[2] Idrees, M., Saleemi, I., Seibert, C. and Shah, M. 2013. Multi-source, multi-scale counting in extremely dense crowd images. IEEE Conference on Computer Vision and Pattern Recognition, 2547-2554.

[3] Ali, S. and Shah, M. 2008. Floor fields for tracking in high density crowds. Proceedings of the 8th European Conference on Computer Vision, 1-14.

[4] McCool, M., Reinders, J. and Robison, A. 2008. Structured parallel programming: Patterns for efficient computation. MA: Morgan Kaufmann.

[5] ImageNet database, http://www.image-net.org/

[6] Russel, B., Torralba, A., Murphy, K., Freeman, W.T. 2007. LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, 77, 157-173.

[7] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition.

[8] Sadanand, S. and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. IEEE Conference on Computer Vision and Pattern Recognition.

[9] Kipp, M. 2012. Multimedia Annotation, Querying and Analysis in ANVIL. In: M. Maybury (ed.) Multimedia Information Extraction, Chapter 19, Wiley - IEEE Computer Society Press.

[10] OpenShapa, http://openshapa.org/

[11] Yuen, J., Russell, B. C., Liu, C. and Torralba, A. 2009. LabelMe video: building a video database with human annotations. IEEE International Conference on Computer Vision.

[12] Vondrick, C., Patterson, D., Ramanan, D. 2012. Efficiently Scaling Up Crowdsourced Video Annotation. International Journal of Computer Vision, 1-21.

[13] ViPER-GT, http://viper-toolkit.sourceforge.net/products/gt/

[14] Van Rest, J., Grootjen, F.A., Grootjen, M., Wijn, R., Aarts, O., Roelofs, M.L., Burghouts, G.J., Bouma, H., Alic, L. and Kraaij, W. 2013. Requirements for multimedia metadata schemes in surveillance applications for security. Multimedia tools and applications, 70, 573-598.

[15] Bourdev, L. and Malik, J. 2009. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. IEEE International Conference on Computer Vision.