Interactive Video Annotation Tool for Generating Ground Truth Information

Sungjoo Park
Smart Media Research Center
Korea Electronics Technology Institute (KETI)
Seoul, Korea
bpark@keti.re.kr

Chang Mo Yang
Smart Media Research Center
Korea Electronics Technology Institute (KETI)
Seoul, Korea
cmyang@keti.re.kr

***Abstract*— In this paper, we propose an interactive video and image annotation tool to generate ground truth information, that is essential information for training deep neural network. The proposed annotation tool not only generates various ground truth (GT) information such as object, motion, and event information, but also supports a semi-automatic video and image annotation method for fast generation of ground truth. The ground truth generated in the proposed tool is stored in the metadata database as a form of XML. The implementation results show that the proposed annotation tool provides faster and more detailed ground truth information compared to the existing methods.**

Keywords— interactive annotation; ground truth; video and image tagging; metadata of GT

I. INTRODUCTION (*HEADING 1*)

As interest in researches on understanding and analyzing videos and images and images based on deep neural networks, the importance of data for learning and evaluating deep neural networks has been emphasized. These dataset for deep neural networks should contain accurate, detailed and sufficient GT information for the services to be applied [1-2]. In building dataset, conventional manual annotation methods are time consuming, have low accuracy, and have a difficulties in generating consistent GT information [3-5]. Therefore, conventional manual method is not suitable to be applied to a typical deep neural network-based video analysis technology such as an intelligent surveillance system [6].

In this paper, we propose an interactive video and image annotation tool to generate GT information more efficiently compared with the conventional methods. The proposed annotation tool not only provides automatic analyzing method of videos/images for fast generation of GT, but also provides various GT such as object information, motion information, and event information. The GT generated in the proposed annotation tool is automatically converted into metadata and stored in the database

II. ANNOTATION SYSTEM

A. System Structure

Fig. 1 shows the structure of the interactive annotation tool proposed in the paper. In order to generate the GT information for the video input, the analysis for the fixed object, the moving object and the event is performed respectively. Object detection is a process for checking whether an object exists in video input, and object segmentation is a process for detecting object position information (x-y coordinate information). Metadata tagging is a process for analyzing attribute information of a detected object in detail. And metadata building process is for encoding object information and attribute information into the predefined metadata.

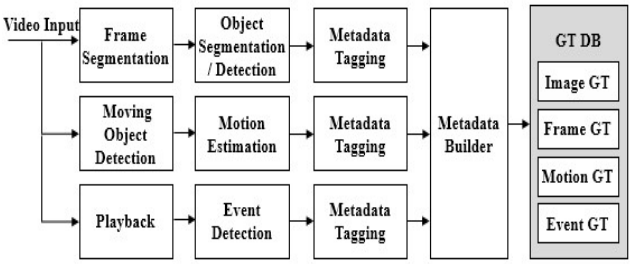


Fig. 1. Structure of interactive annotation tool

These process of generating GT information is performed on a frame-frame basis. Objects located in each video frame are independently described through frame segmentation, object segmentation, metadata tagging, and metadata building. Process for generating GT information about motion objects and video event is executed in parallel. GT information generated through the metadata builder is stored in each GT database and stored GT in databases is accessible through the open APIs.

B. Metadata Structure of Ground Truth Information

In the proposed interactive annotation tool, the GT information is converted into metadata and it is generated based on XML(eXtensible Markup Language). The generated metadata has components such as annotation, source, objects, motions, and events specifically, as shown in Fig. 2.

Fig. 3 shows the structure of annotation and source. “annotation” is a component that is commonly used in the metadata of videos and images. It is consists of information describing tagger name, tagging time, approval information, approval time, and other description information as shown in Fig. 3. “source” describes information of input file, such as input database name, input types, input URL, file name, file format, codec information, frame rate, and resolution.

```
<?xml version="1.0" encoding="UTF-8"?>
<gt>
     <annotation> </annotation>
    <source> </source>
    <objects> </objects>
    <motions> </motions>
    <events> </events>
</gt>
```

Fig. 2. Structure of metadata

```
<?xml version="1.0" encoding="UTF-8"?>
<gt>
     <annotation>
       <tagger> </tagger>
       <taggingtime> </taggingtime>
       <validated> </validated>
       <validationtime> <validationtime>
       <description> </description>
    </annotation>
    <source>
       <database> </database>
       <type> </type>
       <camera> </camera>
         <gps></gps>
         <url> </url>
         <filename> </filename>
         <format> </format>
         <codec>
                  <image> </image>
                  <video> </video>
                  <audio> </audio>
         </codec>
         <framerate> </framerate>
         <size>
                  <width> </width>
                  <height> </height>
                  <depth> </depth>
         </size>
         <segmented> </segmented>
         <normalized> </normalized>
    </source>
</gt>
```

Fig. 3. Structure of annotation and source

```
<?xml version="1.0" encoding="UTF-8"?>
<gt>
    <objects>
        <nframe> <inframe>
        <nobject> </nobject>
        <id><iid>
        <otype> <iotype>
        <pose><ipose>
        <truncated> <truncated>
        <validity></validity>
        <bndbox>
                 <xmin></xmin>
                 <ymin></ymin>
                 <xmax></xmax>
                 <ymax> </ymax>
        </bndbox>
        <attribute>
                 <race> </race>
                 <gender </gender>
                 <age> </age>
                 <name> <iname>
                 <height>  </height>
                 <topcolor> </topcolor>
                 <pantscolor></pantscolor>
                 <glasses> </glasses>
        </attribute>
    </objects>
        <licenseplate>
                 <bounding>
                   <xmin></xmin>
                   <ymin>)</ymin>
                  <xmax></xmax>
                   <ymax> </ymax>
                 </bounding>
                 <character></character>
        </licenseplate>
    </objects>
    <objects> ... </objects>
</gt>
```

Fig. 4. Structure of objects

Fig. 4 shows the structure of objects. “objects” describes information about an individual object of video frames. “objects” includes information such as the frame number, the number of objects, the object ID, the object type, the posture of the object, and the coordinates of the object

```
<?xml version="1.0" encoding="UTF-8"?>
<gt>
     <motions>    !motion object
         <mobject>
                  <mid> </mid>
                  <nframe> </nframe>
                  <bndbox>
                     <xmin> </xmin>
                     <ymin> </ymin>
                     <xmax> </xmax>
                     <ymax>  </ymax>
         </bndbox>
         <nframe> </nframe>
         </mobject>
       </motions>
       <events>
         <startframe  </startframe>
         <endframe> </endframe>
         <event> )</event>
         <validity>  </validity>
       </events>
 </gt>
```

Fig. 5. Structure of *motions* and *events*

Fig. 5 shows the structure of “motions” and “events”. They are valid only for video input. “motions” describe the location information of the motion objects in list form based on the motion object ID. “events” describe event information detected in videos based on playback time information.

C. *Frame-based Motion Interpolation*

For more efficient generation of GT information for motion objects, frame-based motion interpolation techniques is applied in this paper. If an arbitrary section of video is set, a GT for the motion object is generated on a frame-by-frame basis and is stored as metadata. If the GT for a motion object is modified in a specific frame, the GT information of the front and rear frames on the time axis is automatically corrected. Fig. 6 shows the frame-by-frame motion interpolation scheme

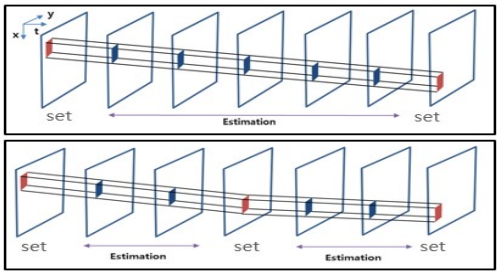


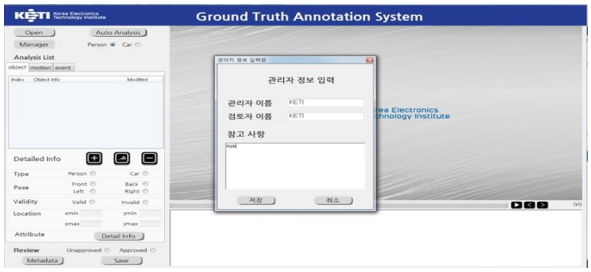
Fig. 6. Scheme of frame-baed *motions interpolation*

D. *Automatic Image and Video Analysis*

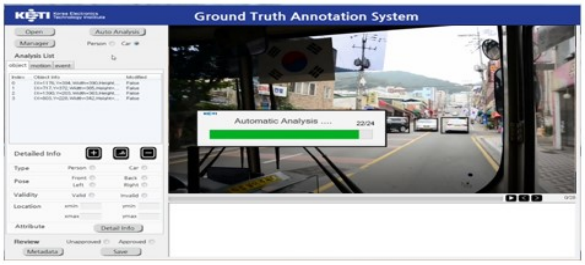
The conventional methods use the manual annotation to generate GT information. However, it take a lot of effort and time in order to generate accurate, diverse GT information and maintain the consistency. In order to overcome the drawbacks of the conventional methods, we propose a semiautomatic annotation tool that accelerates processing speed by automating

the object extraction and analysis process, which requires a lot of time and effort in generating GT information.

To automate the object extraction and analysis process, the proposed annotation tool uses the cloud APIs of Sighthound [7]. Sighthound's cloud API uses square coordinates to provide object type and object location information, as well as object property information. The property information provided by Sighthound's cloud API includes various detailed information about person and vehicles object types.

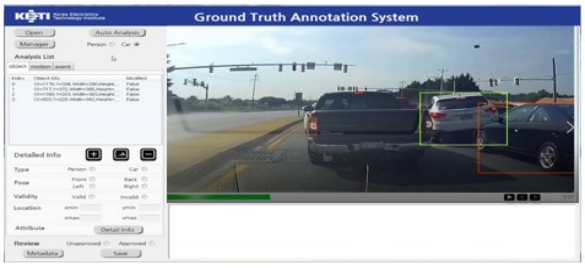


(a) Main GUI of annotation tool with user registration



(b)Automatic GT generation (frame-by-frame)

(c) Modification of detailed GT information



(d) Metadata generation

Fig. 7. Implementation results of the proposed annotation tool

III. Implementation Results

Fig. 7 shows the implementation results of the interactive annotation tool proposed in this paper. Fig. 7(a)-(d) show the result of executing the proposed annotation tool, automatic analysis of GT, modification of GT information, and metadata encoding respectively. As shown in Fig. 7(a), the implemented annotation tool consists of an information input part, an automatic analysis execution part, an analysis list presentation part, a detailed information presentation part, an image and video playback part, and a metadata generation part. The proposed annotation tool performs object extraction and analysis using the cloud APIs of Sighthound. After the automatic analysis, the generated analysis results can be modified by users, converted into metadata and stored in the database.

IV. Conclusions

In this paper, we propose a new interactive video and image annotation tool to generate GT information that is essential for the deep neural networks. By using the cloud APIs of Sighthound to perform automatic objects analysis, the proposed annotation tool accelerates the generation of GT information. Implementation results show that the proposed annotation tool not only can perform faster annotation by applying automatic analysis of objects, but also generate various GT information consistently.

Acknowledgment

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2018-0-00348, Development of Intelligent Video Surveillance Technology to Solve Problem of Deteriorating Arrest Rate by Improving CCTV Constraint).

References

[1] J. Park, and S. Yi, "Development of video database and a video annotation tool for evaluation of smart CCTV system," *Journal of Korea Institute of Electronic Communication Science*, vol. 9, no. 7, pp. 739-745, July 2014.

[2] J. S. Lopez-Villa, H. D. Insuasti-Ceballos, S. Molina-Giraldo, A. Alvarez-Meza, G. Castellanos-Dominguez, "A novel tool for ground truth data generation for video-based object classification," *Symposium on Signal Processing, Images and Computer Vision*, pp. 1-6, September 2015.

[3] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," *Asian Conference on Computer Vision*, pp. 31-44, November 2012.

[4] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing PASCAL VOC," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41-48, September 2014.

[5] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1-8, June 2005.

[6] L. Li, W. Huang, I. Y. Gu, R. Luo, and Q. Tian, "An Efficient Sequential Approach to Tracking Multiple Objects Through Crowds for Real-Time Intelligent CCTV Systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 38, no.5, pp. 1254 – 1269, October 2008.

[7] https://www.sighthound.com/products/cloud