Selection Measure of Stable Lighting Video for Region-Based Video Object Annotation

Xiaomeng Wu[1], Wenli Zhang[2], Tamon Sadasue[1], Shunsuke Kamijo[1], and Masao Sakauchi[1]

1. Introduction

In our previous work, we proposed a semi-automatic object extraction and annotation approach based on the region-based model matching method that did not consider the impact of the lighting changes and the typical feature representation impreciseness. In this paper, we propose a selection measure of stable lighting video for the proposed approach, which quantitatively defines the instability of the lighting condition of the video. For videos with relatively stable lighting conditions, the proposed annotation approach showed a good performance with a 74.4% precision and a 94.0% recall.

2. Region-Based Object Extraction

By consulting a semantic object model database, the same semantic objects such as video characters in key-frames of each video shots can be queried and semi-automatically annotated based on the similarity of low-level features of each region. In general, an object model is a template describing a specific object. During the matching process, each template is inspected to find the closest match. We proposed a semantic object model, which has a hierarchy structure, from object, salient regions to low-level features. The low-level features, such as the color, texture, area size, and position of each region, are extracted automatically from segments obtained from a key-frame.

The key-frames extracted from the video are all segmented so that they contain a set of regions in a 12-dimensional feature space. During the matching process, all these key-frames are retrieved to find the "closest" match corresponding to each semantic object model. The key-frame is considered a final candidate result of the semantic object if it contains the regions that are reasonably close to the model regions in terms of the color distance, the spatial relationship, the area ratio, etc.

Using the above model matching method, video shots containing the query object of varying size and location can be retrieved. The user needs only to select those frames that contain the true query object, ignoring any false ones, using the supplied interface. This is the only manual operation required during the object annotation. Subsequently, all other annotation requirements are completed automatically. Thus, the whole annotation procedure is called *semi-automatic object annotation*.

3. Entropy Error Rate

In this section, an information-theoretic measure, termed Entropy Error Rate (EER), which is used for single-image-enhancement, is proposed as a quantitative measure of the information distribution within an image.

In the intensity histogram, pixels are classified as darker pixels, where the intensity values of pixels are below the mean intensity value, and lighter pixels otherwise. A simple statistic S, called the singularity, can be introduced to estimate the relative position of the mean within the intensity histogram.

$$S = 4\left(\frac{I_{\max} - I_{mean}}{I_{\max} - I_{\min}}\right)\left(\frac{I_{mean} - I_{\min}}{I_{\max} - I_{\min}}\right)$$

Inside an image, information of the darker pixels H_D and the lighter pixels H_B can be measured respectively. Moreover, the average entropy of each side, which measures the amount of information contained in only one unit of intensity level, can be calculated.

$$\overline{H}_D = \frac{H_D}{I_{mean} - I_{\min} + 1} \qquad H_D = \sum_{K=I_{\min}}^{I_{mean}} -P(K)\log P(K)$$

$$\overline{H}_B = \frac{H_B}{I_{\max} - I_{mean} + 1} \qquad H_B = \sum_{K=I_{mean}}^{I_{\max}} -P(K)\log P(K)$$

The asymmetry of the image information distribution between these two sides can be measured using the simple statistic in the following equation, which is termed the Entropy Error Rate (EER).

$$EER = \frac{\overline{H}_D - \overline{H}_B}{S}$$

With the EER, the tendency of the information distribution within an image can be easily inferred. When the EER is positive and its absolute value is relatively large, the information lies mainly in the darker pixels, which means that the image appears darker; when the EER is negative and its absolute value is relatively large as well, information is principally distributed in the lighter pixels, which means that the image appears lighter. A high-quality image should be neither too dark nor too light, so its EER should be within a preset acceptance range.

4. Instability of Lighting Conditions

In this paper, we propose a selection measure of stable lighting video to quantitatively define the instability of the lighting condition of the video. Based on this selection measure, video content providers can quantitatively realize the lighting condition of the video and choose the object videos that adapt to the needs of different applications.

Consider the mean intensity values of a set of frames extracted from a video in time order, $\{I_{mean1}, I_{mean2}, \ldots I_{meanN}\}$. The average mean intensity values of these frames, i.e. the mean intensity value of the video, can be computed.

$$E(I_{mean}) = \frac{1}{N}\sum I_{mean}$$

We use this value as the average lighting condition of the video to divide each frame into darker pixels, where the intensity values of pixels are below $E(I_{mean})$, and lighter pixels, where the intensity values of pixels are above $E(I_{mean})$. Furthermore, EER, which is used in Section 3.1 for adaptive frame enhancement, is computed to represent the lighting condition of each frame in the video. If the EER values of the frames are far different from each other, it means that the lighting condition of the video varies frequently in time order; if the EER values of the frames are almost the same, it means that the light condition of the video is stable and only vary within a narrow scope. Therefore, the variance of the EER values is finally computed to quantitatively represent the instability of the lighting condition of the video (*IL*). a and b are two parameters that control the value range of *IL*.

$$S' = 4\left(\frac{I_{\max} - E(I_{mean})}{I_{\max} - I_{\min}}\right)\left(\frac{E(I_{mean}) - I_{\min}}{I_{\max} - I_{\min}}\right)$$

$$EER' = \frac{\overline{H}_D - \overline{H}_B}{1 + a^{S'}}$$

$$IL = bV(EER')$$

To examine the wellness that the proposed *IL* characterizes the instability of the lighting condition, the relationship between *IL* and the performance of the proposed approach is statistically simulated. 36 data sets are randomly extracted from the 15 videos used above, the size of which is 100 images/set. From these data sets, 50 models are constructed and retrieved using the proposed object extraction approach. Fig. 1 shows the experimental result (a=100, b=1000).

From Fig. 1, we can see that the proposed *IL* could almost exactly reflect the instability of the lighting condition, and with *IL* increasing, the performance of the object extraction approach decrease. For videos with relatively stable lighting conditions (*IL*<10), the proposed object extraction approach showed a good performance with a 74.4% precision and a 94.0% recall averagely. Based on this selection measure, video content providers can quantitatively realize the lighting condition of the video and choose the object videos that adapt to the needs of different applications.

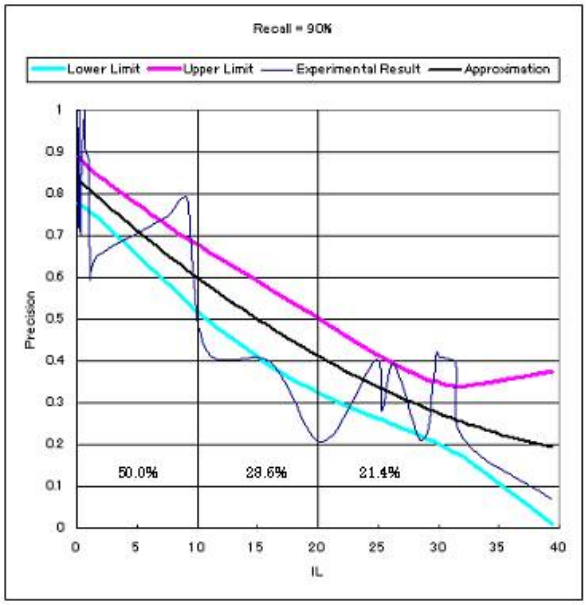

Figure 1. *IL* vs. Precision (Recall=90%)