ICDAR 2013 Robust Reading Competition

Ground Truthing Procedure for the Video Challenge

Definitions

"***Visible***" is whatever is in the image and "***non visible***" whatever is outside the image.

"***Readable***" is the subjective assessment of whether a word can be understood or not. If part of the word is "Non-Readable" then the whole word is defined as "***Non-Readable***"

Rules

1. Ground truth is created at the word level (i.e. no paragraph, text line or single character regions), with the exception of rule 6.
2. The transcription is assigned based on the part of the word that is visible in the current frame.
    - *E.g. imagine a word that is entering from the right of the scene and exiting to the left, first a few characters appear, the transcription therefore comprises only the few characters that are in the video frame. Subsequently more characters enter the frame. As more characters become visible, the transcription changes to reflect what appears on the video in every frame. The opposite happens as the word goes out of the frame; characters that are not visible any more are removed from the transcription which always reflects what is visible in the frame.*
3. Occluded objects are treated as single objects. In the case of occlusions a single region is defined that includes all parts of the occluded object, and the transcription is done so that it reflects whatever is visible, with a space where the occlusion(s) happen.
    - *E.g. imaging a sign saying "Barcelona" which is occluded by a lamp post so that the "c" and part of the "e" letter is not visible. The whole word would be defined as a single bounding box (that would naturally contain part of the lamp post that occludes the sign) and the transcription should be "Bar lona", where a space has been used in the places where characters are occluded (not visible).*
4. The transcription has to be correct for "medium" and "high" quality regions. For "low" quality (equivalent to "don't care") the transcription does not matter.
5. "Low" quality are equivalent to "don't care" regions. We mark as "low" quality everything that we would not like to cause a penalty if not detected automatically.
6. In the case of "low" quality ("don't care") regions, we don't follow the rule of ground truthing at the word level.
    - *E.g. imagine for example a sign at a large distance, that you know it is text but it is unreadable, the whole sign would be marked as a single "don't care" region as we know it is text, but we cannot make out any words. As words become visible, they would be marked at word level with "medium" or "high" quality and the "don't care" region(s) would change shape to cover the part that is still not readable.*

7. "Medium" and "High" quality are assigned based on subjective evaluation, no particular rules are used. We will not use this for the evaluation process.
8. “Low” quality (= “don’t care” regions) are assessed based on whether the text can be read or not, and not based on the quality of the image.
    - *E.g. In case of text that enters from the bottom or the top, it might happen that in the first frames a small part of all characters is visible, but as it is only a small part the text is not recognizable. This would be a "low" quality area ("don't care"), even if the parts are in high resolution and good image quality.*