Using Explicit Discourse Rules to Guide Video Enrichment

Michiel Hildebrand
Centrum voor Wiskunde en Informatica
Science Park 123, 1098 XG Amsterdam
The Netherlands
m.hildebrand@cwi.nl

Lynda Hardman
Centrum voor Wiskunde en Informatica
Science Park 123, 1098 XG Amsterdam
The Netherlands

ABSTRACT

Video content analysis and named entity extraction are increasingly used to automatically generate content annotations for TV programs. A potential use of these annotations is to provide an entry point to background information that users can consume on a second screen. Automatic enrichments are, however, meaningless when it is unclear to the user what they can do with them and why they would. We propose to contextualize the annotations by an explicit representation of discourse in the form of scene templates. Through content rules these templates are populated with the relevant annotations. We illustrate this idea with an example video and annotations generated in the LinkedTV[1] project.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]: Miscellaneous

Keywords

Second screen; Video annotation; Discourse; Scene templates; Rule-based

1. INTRODUCTION

Video content annotations have become a common source of metadata for TV programs. These annotations are provided by a variety of techniques and describe a wide variety of characteristics of the video content. For example, content analysis techniques extract visual concepts, shots and scenes [10, 11], named entities are extracted from the subtitles [5, 9], and crowdsourcing initiatives result in user tags [6].

A potential use case for these annotations is the enrichment of the TV experience by presenting background information on a second screen [2]. Designing second screen interfaces for automatically generated annotations is, however, not straightforward. Currently, second screen applications are carefully designed for the context of a specific TV program, and the presented material is manually curated to ensure it is correct and relevant to the end-users needs. As generated annotations can be incorrect or irrelevant to the user's needs, a straightforward data-driven approach can easily overwhelm the user. Furthermore, the semantics of generated content annotations are typically incomplete and without proper contextualization it might be unclear to the user why certain information is useful. For example, an annotation might indicate that a person occurs in a video, but not what the role of this person is in the video (e.g. the presenter, main character or just someone in the background).

In this short paper we propose to contextualize the annotations by an explicit representation of discourse, comparable to [7]. In a second screen application the discourse for a specific TV program is authored in the form of scene templates. These templates are populated by applying rules to the generated annotations. The authored scene templates provide the context, supporting users to determine relevance of the information. The content rules provide a top-down constraint to select the relevant information by combining annotations and background knowledge. As a proof of concept we describe a scene template for an episode of the TV program *Tussen Kunst en Kitsch* broadcasted by the AVRO, the Dutch version of The Antiques Roadshow, and explain how it can be populated by applying rules on the annotations that are generated for this episode in the LinkedTV project.

2. MOTIVATING EXAMPLE

In the TV program *Tussen Kunst en Kitsch* antique objects, brought in by the public, are assessed by art and historical experts. The experts provide information about the historic or artistic context of the antiques and eventually give an estimate of the object's financial value. In the episode of December 8 2012[2] a guest brings a small drawing. The expert explains that it is made by the Dutch artist Jan Toorop. He continues by contextualizing the painting and describes that the artist made it while he was staying in the Dutch city Domburg. During this period Jan Toorop attracted other famous Dutch artists such as Piet Mondriaan. He also explains how Jan Toorop is famous for using the pointillist painting technique that characterized the neo-impressionist style. By the influence of Vincent van Gogh he moved towards the post-impressionist style in his later work, using stripes instead of points. This is also the technique used for the drawing in this scene.

Viewers interested in the topic might want to explore more information about Jan Toorop, for example from his

[1] http://www.linkedtv.eu/

[2] http://avro.nl/tussenkunstenkitsch/detail/8237850

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

Wikipedia page[3], find out where Domburg[4] is located, or learn more about the impressionist and post-impressionist styles and the typical techniques of these styles.

3. VIDEO CONTENT ANNOTATIONS

Within the LinkedTV project annotations are generated using a variety of techniques. Through content analysis shots, scenes and visual concepts are detected (e.g. person, text, indoor). From the subtitles named entities, such as person and location names, are extracted. The tool chain being developed in the LinkedTV project is described in detail in [1]. The RDF model to represent the generated annotations is described in [4]

The scene of the TV program *Tussen Kunst en Kitsch* about the drawing of Jan Toorop has a duration of about 4 minutes and contains 260 annotations. The reader can access the annotations of this scene at: `http://goo.gl/WIpvP`.

4. SCENE TEMPLATES

Figure 1 shows a mockup of a second screen application for the TV program *Tussen Kunst en Kitsch*. The application makes a distinction between program-specific information and scene-specific information. The mockup shows the page with scene-specific information. The program-specific information is available under a different page, accessible by the links at the top of the page. The program page contains the curated metadata, such as the title, date and a description. It is also contains information about the host, *Nelleke van der Krogt*, and the recording place of the episode, *the Hermitage, Amsterdam*. By presenting this information on the program page it is accessible at any time without interfering with the scene specific information. In addition, the program specific information does not have to be presented again on the scene page, even if it is detected in the visual content or subtitles, saving valuable real estate for the scene specific information.

The scene template is based on the format of the TV program. Each scene is a conversation about an antique object between an expert and a guest. The top left part of the scene template is reserved for this information. Selecting an item from this scene level metadata brings up additional information on the top right. In the mockup the antique object is selected and images of frames in the video depicting the object are shown, allowing the user to inspect it more closely. Other information could be a Wikipedia page, a geographical map, or high resolution images of the object. The scene-level information is fixed and remains available to the user during the time that the scene is active.

The bottom part of the screen shows a browsable carousel with background information about the antique object. Again note, that annotation about the expert, guest and object does not have to be presented anymore. Selecting a background item shows the corresponding information in the detail view on the top right. In addition, the bottom left side shows the object that is currently shown in the video on the main screen, giving quick access to background information of the current topic on the main screen.

[3] `http://en.wikipedia.org/wiki/Jan_Toorop`
[4] `https://maps.google.nl/maps?q=domburg`

5. CONTENT RULES

To fill the scene template two types of annotations are required: the scene metadata and the background information about the object. The scene metadata should contain besides the start and end times, the expert, the antique object and the guest in case it is a known person. The background information can contain among others the persons, locations, art styles and techniques related to the antique object. To fill the scene template we propose a number of rules that define the relevant and contextualized annotations that should be presented to user. These content rules combine annotations generated by different techniques, as well as program metadata and background knowledge.

The rules presented in this paper are used to illustrate what these rules would look like and provide a first indication that a rule based approach is feasible. In addition, we explore what type of generated annotations are required, which are already available and which are still missing. The rules are written in prolog notation. The implementation and analysis of the performance is future work.

To derive who the expert is in a scene, we use the annotations provided by the shot and visual concept detection combined with program metadata. The rule below defines the expert in a particular scene. The head of the rule, `expert(Scene, Name)` requires an identifier of the `Scene` and returns (by Prolog unification) a `Name`. For the body of the rule we use the fact that in each scene the expert is identified by a text label in the video. In addition, we use the program level information that lists all the experts that appear in the program[5]. Specifically, the rule states that there should be a shot in the scene that contains a person as well as a text label. In addition, the text label that is found should match the name of one of the experts from the program metadata. In this case we provide this program metadata by the clause `expert_name`.

```
expert(Scene, Name) :-
    shot(Scene, Shot),
    person(Shot),
    overlay_text(Shot, Name),
    expert_name(Name).

expert_name('Frank Welkenhuysen').
...
```

To satisfy the predicates in the rule body `shot`, `person` and `text` the actual annotations are used. For example, a person can correspond to an annotation with the LSCOM[6] concept *Face*, *Person* or *Commentator_Or_Studio_Expert*. The annotations generated in the LinkedTV project provide the required information to satisfy this rule. The reader can verify this using the annotations listed in the spreadsheet at `http://goo.gl/WIpvP` and the cells A14 (shot), A15 (Text) and A19 (Commentator_Or_Studio_Expert). However, the text label detection does not yet return the actual label. This requires an extension of the LinkedTV tool chain with optical character recognition (OCR).

The rule to derive a *known* guest is similar to the expert rule. A *known* guest occurs in a scene when there is a shot containing a person and a text label. In this case, however,

[5] The website of the TV program lists the names of all the experts `http://avro.nl/tussenkunstenkitsch/Experts/`
[6] `www.lscom.org`

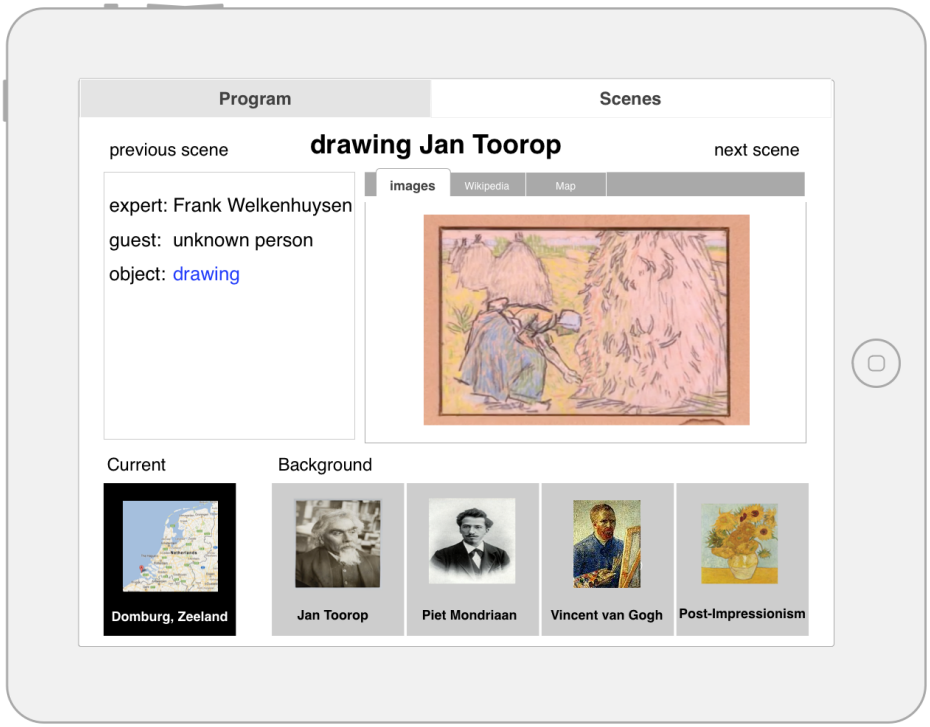

Figure 1: Mockup of a second screen application for the TV program *Tussen Kunst en Kitsch*.

the label should not be one of the experts and also not the name of the presenter.

To detect the antique object that is discussed in a scene we make use of additional information provided by the broadcaster in the form of photographs of the objects. Using these photographs as input an object re-detection algorithm can generate all the shots in which these objects occur. The rule `antique_object` derives the object given an identifier of the scene. The predicate `object_redetection` returns the object corresponding to the photograph that is most frequently detected in the scene. In addition, it returns all the shots in which the object occurred. Currently, the broadcaster's photographs were not used in the LinkedTV workflow. Instead the object re-detection was run using a manually selected region in the video as the input. Supporting this rule requires that additional information, in this case photographs, are used in the video content analysis process.

```
antique_object(Scene, Object, Shots, Type) :-
    object_redetection(Scene, Object, Shots),
    shot_annotations(Shots, Annotations),
    exclude(visual_work, Annotations, Types),
    most_frequent(Types, Type).
```

The rule also derives the type of the object, e.g. a drawing. In this case both the visual concepts and the concepts detected in the subtitles can provide information about this type. The rule selects all the annotations that co-occur in the shots depicting the object, and filters these using background information. In this case we include only the annotations that match with concepts contained in the *visual works* facet of Getty's Art and Architecture Thesaurus (AAT). The AAT concept that is most frequently found is selected.

The named entities extracted from the subtitles contain useful background information about the antique object. For example, it contains the name of the artist *Jan Toorop*, the name of related artists, *Piet Mondriaan* and *Vincent van Gogh*, the name of the place the object was created in, *Domburg*, and the Dutch province this is located in, *Zeeland*. Simply presenting all the recognized named entities can, however, lead to confusing information. For example, the Dutch place *Haarlem* is also detected, but this is mentioned by the guest as the location where she found the object in her grandmother's house. A first step to eliminate irrelevant information is to define a rule that selects only the entities in utterances of the expert. After all, the expert provides all the relevant background information. The rule below defines the entities uttered by the expert. It uses the expert rule we defined before, takes the entities found in the scene and checks if these are uttered by the expert.

```
expert_entity(Scene, Entity) :-
    expert(Scene, Expert),
    named_entity(Scene, Entity),
    in_utterance(Scene, Expert, Entity).
```

To contextualize the named entities we would like to derive how they are related to the antique object that is discussed in the scene. There are two sources that could provide information about this relationship. The subtitles contain besides the named entities (groups of) words that can indicate the relation, e.g. Jan Toorop *was influenced by* Vincent van Gogh. However, relation extraction is a difficult task. Another approach is to use relations from background knowledge. For example, in DBPedia both Jan Toorop and

Vincent van Gogh are known as post-impressionist painters[7]. Semantic patterns are an active research topic in the Semantic Web community [3]. In the art domain patterns to define relationships between persons could also be formulated as rules. For example, that persons are related if they share an art style.

```
related_person(Scene, Person1, Person2, Style) :-
    person(Person1),
    person(Person2),
    has_style(Person1, Style),
    has_style(Person2, Style).
```

6. DISCUSSION

Within the LinkedTV project we aim to further explore the approach presented above. Initially we will focus on episodes of the TV program *Tussen Kunst en Kitsch*, as discussed in this paper. The content rules we drafted make us believe that deriving and contextualizing relevant information for this TV program is feasible. This is partly due to the strict format of the program and the scope of the content that is discussed, focussing on artistic or historic information. It needs to be investigated how the approach generalizes to other types of TV programs. The scenes in a news program (or items) typically follows a strict format too, containing the news reader that discusses a *current* topic, possibly assisted by a correspondent. The content will contain persons (possibly being interviewed), locations and events. In addition, it needs to be investigated how the approach could work for TV programs that require multiple scene templates.

From the example discussed in this paper we derived three requirements on content annotations. To derive the names of the experts and *known* guests, the labels from overlay texts should be made available through character recognition. The current state of the art should be sufficient for this use case [8]. To detect the antique object in a scene photographs of the objects should be used in the object re-detection algorithm. To detect more relevant background information, domain specific knowledge should be used in the named entity extraction process. For the TV program *Tussen Kunst en Kitsch* thesauri from the art domain should be used, such as Getty's Art and Architecture Thesaurus (AAT) and the United List of Artist Names (ULAN)[8].

In this paper we focussed on the annotations itself and only briefly touched upon the information that end-users can explore via these annotations, e.g. Wikipedia pages, geographical maps. It requires further user studies to explore which information users want access to and how they want to access these.

7. ACKNOWLEDGMENTS

This research was partially supported by the LinkedTV project, funded by the European Commission through the 7th Framework Programme (FP7-287911).

[7] `http://dbpedia.org/class/yago/Post-impressionistPainters`

[8] `http://www.getty.edu/research/tools/vocabularies/`

8. REFERENCES

[1] E. Apostolidis, M. Dimopoulos, V. Mezaris, D. Stein, J. Blom, I. Lasek, M. Sahuguet, B. Huet, N. de Abreu Pereira, and J. Muller. D1.1. State of the art and requirements analysis for hypervideo. LinkedTV project deliverable, September 2012.

[2] L. Baltussen and J. Oomen. Antiques interactive. In *PATCH'12: Workshop on Personalized Access to Cultural Heritage*, Nara, Japan, November 2012.

[3] A. Gangemi and V. Presutti. Towards a pattern science for the semantic web. *Semantic Web Journal*, 1(1-2):61–68, 2010.

[4] J. L. R. Garcia, R. Troncy, and M. Vacura. D2.2. Specification of lightweight metadata models for multimedia annotation. LinkedTV project deliverable, October 2012.

[5] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW'12, Workshop on Linked Data on the Web*, Lyon, France, April 2012.

[6] R. Gligorov, M. Hildebrand, J. Van Ossenbruggen, G. Schreiber, and L. Aroyo. On the role of user-generated metadata in audio visual collections. In *Proceedings of the International Conference on Knowledge Capture (K-CAP)*, pages 145-151, Banff, Canada, June 2011.

[7] S. Kim, H. Alani, W. Hall, P. H. Lewis, D. E. Millard, N. R. Shadbolt, and M. J. Weal. Artequakt: Generating tailored biographies with automatically annotated fragments from the web. In *Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02), the 15th European Conference on Artificial Intelligence, (ECAI'02)*, pages 1-6, Lyon, France, July 2002.

[8] R. Lienhart. Automatic text recognition for video indexing. In *Proceedings of the fourth ACM international conference on Multimedia*, MULTIMEDIA '96, pages 11-20, New York, NY, USA, 1996.

[9] D. Odijk, M. de Rijke, and E. Meij. Feeding the second screen: Semantic linking based on subtitles. In *Open research Areas in Information Retrieval (OAIR 2013)*, Lisbon, Portugal, May, 2013.

[10] C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25:5–35, 2005.

[11] D. Stein, E. Apostolidis, V. Mezaris, N. de Abreu Pereira, J. Müller, M. Sahuguet, B. Huet, and I. Lasek. Enrichment of news show videos with multimodal semi-automatic analysis. In *Proceeding of the NEM-Summit*, Istanbul, Turkey, October 2012.