H.761 Support of a New <concept> Element and a New "recognition" Node-Event to Enable Deep Learning-based Analyses for Media-Nodes

Antonio Busson
TeleMídia - PUC-Rio
busson@telemidia.puc-rio.br

Alan L. V. Guedes
TeleMídia - PUC-Rio
alan@telemidia.puc-rio.br

Sergio Colcher
Informatics Departament - PUC-Rio
colcher@inf.puc-rio.br

ABSTRACT

Machine Learning field, methods based on Deep Learning (e.g. CNN, RNN) becomes the state-of-the-art in several problems of the multimedia domain, especially in audio-visual tasks. Typically, the training of Deep Learning Methods is done in a supervised manner, and it is trained on datasets containing thousands/millions of media examples and several related concepts/classes. During training, the Deep Learning Methods learn a hierarchy of filters that are applied to input data to classify/recognize the media content. In computer vision scenario, for example, given image pixels, the series of layers of the network can learn to extract visual features from it, the shallow layers can extract lower-level features (e.g. edges, corner, contours), while the deeper combine these features to produce higher-level features (e.g. textures, part of objects). These representative features can be clustered into groups, each one representing a specific concept. H.761 NCL currently lacks support for Deep Learning Methods inside their application specification. Because those languages still focus on presentations tasks such as capture, streaming, and presentation. They do not consider programmers to describe the semantic understanding of the used media and handle recognition of such under-standing. In this proposal, we aim at extending NCL to provide such support. More precisely, our proposal able NCL application support: (1) describe learning-based on structured multimedia datasets; (2) recognize content semantics of the media elements in presentation time. To achieve such goals, we propose, an extension that includes: (a) the new <knowledge> element describe concepts based on multimedia datasets; (b) <area> anchor with an associated "recognition" event that describes when a concept occurrences in multimedia content.

KEYWORDS

NCL, Ginga

1 BACKGROUND

Machine Learning field, methods based on Deep Learning (e.g. CNN, RNN) becomes the state-of-the-art in several problems of the multimedia domain, especially in audio-visual tasks. Typically, the training of Deep Learning Methods is done in a supervised manner, and it is trained on datasets containing thousands/millions of media examples and several related concepts/classes. During training, the Deep Learning Methods learn a hierarchy of filters that are applied to input data to classify/recognize the media content. In computer vision scenario, for example, given image pixels, the series of layers of the network can learn to extract visual features from it, the shallow layers can extract lower-level features (e.g. edges, corner, contours), while the deeper combine these features to produce higher-level features (e.g. textures, part of objects). These representative features can be clustered into groups, each one representing a specific concept.

H.761 NCL [2] currently lacks support for Deep Learning Methods inside their application specification. Because those languages still focus on presentations tasks such as capture, streaming, and presentation. They do not consider programmers to describe the semantic understanding of the used media and handle recognition of such under-standing.

2 PROPOSAL

- (1) describe learning-based concepts on structured multimedia datasets using the new <concept> element;
- (2) recognize content semantics of the media elements in presentation time using a virtual anchor, called SemanticAnchor, with an associated new "recognition" event.

The overview of the proposal is illustrated in the next Figure

The <concept> element aiming at group and associate media datasets. It consists of an NCM Composite Node. The Composite Node elements (e.g. <body> and <context>) are useful to define compositions of multimedia data. That way, we <concept> to represent whole media datasets for specific concepts, as well to specify the associations between them. All <concept>s elements in a document are grouped in the <knowledged> at <head>. We define two types of associations among <concept>s:

- The **hierarchy association** defines a parenthood relation where is applied the rule: $(c_1 \Rightarrow c_2)$, the concept node c2 has media features from concept node c1, where c1 and c2 are called parent and child concepts, respectively.
- The **mereology association** defines a parthood relation, where is applied the rule: $(c_1 \vdash c_2)$, the concept node c_2 is part of concept node c_1, this association indicates that media from c_2 are parts of media of c_1.

The following <knowledge> code illustrate the definition of the <concept>s elements and their media sets. In particular, the "tony_face" and "jony_face" <concept>s have hierarchy association "face" <concept>, consequently, "person" <concept>. Moreover, "face_p" <port>" define a mereology. In other words, "face" is a part of "person".

In: Future of Interactive Television Workshop (V WTVDI), Rio de Janeiro, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2019.
ISSN 2596-1683

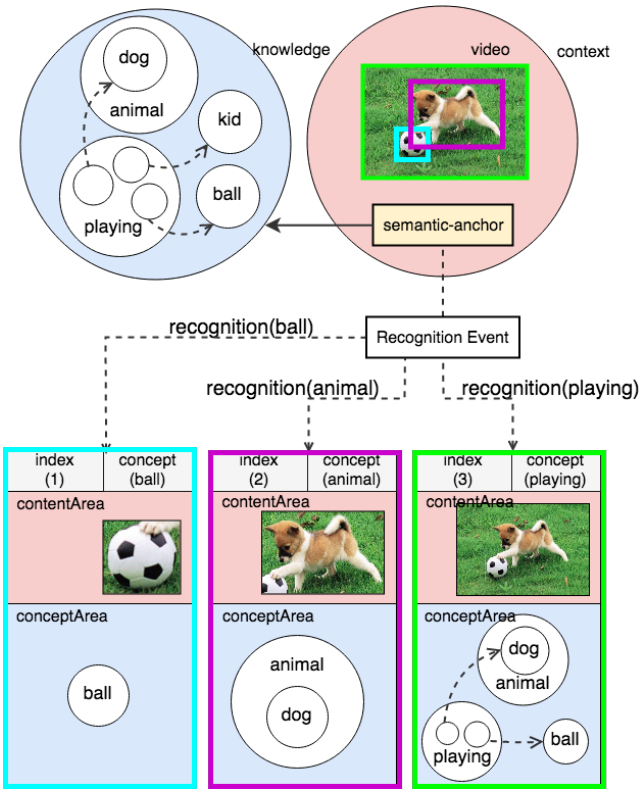

Figure 1: Example of three recognition events.

```
<knowledge>
 <concept id="person">
    <!--any part of any person's media set-->
    <port id="face_p" interface="face">
    <concept id="face">
      <!-- any person face media set-->
      <concept id="tony_face">
        <!-- tony face media set-->
      </concept>
      <concept id="jony_face">
        <!-- jony face media set-->
      </concept>
    </concept>
 </concept>
 </concept>
</knowledge>
```

Listing 1: NCL code for knowledge definition.

Current NCL <media> may define bounding boxes through coords <area> attribute. Additionally, NCL <media> may define time intervals where concepts occurrences are detected through are begin <area> attribute. However, those <area> usually are defined in the authoring phase and just occur in the execution phase. Our virtual <area> anchor with an associated "recognition" event aiming at recognizing content semantics of the media elements in presentation time. This proposal is based on the paper of Busson. al. "Embedding Deep Learning Models into Hypermedia Applications" (accepted to be published)[1].

3 USE CASE

The following code illustrate the usage of the <concept> element. The <head> presents of <concept>s elements and their media sets. in particular, the "dog" <concept> has hierarchy association with "animal" <concept>, whereas "head" <concept>" has mereology association with "dog" <concept>. The NCL link at lines 28-33 will be trigger when a dog appears in the video.

```
<head>
  <knowledge>
   <concept id="animal">
     <!--any animal media set-->
     <concept id="dog">
       <port id="head_p" interface="head" />
       <media src="dog1.png" />
       <media src="dog2.png" />
       <concept id="head">
         <media src="dog_head1.png" />
         <media src="dog_head2.png" />
       </concept>
     </concept>
   </concept>
  <knowledge>
  <connectorBase>
    <causalConnector id="onRecognizeStart">
      <simpleCondition role="onRecognize"/>
      <simpleAction role="start"/>
    </causalConnector>
</connectorBase>
</head>
<body>
 <port id="start" component="video1">
 <media id="textDog" src="textDog.png">
 <media id="video1" src="videos/dog.mp4">
   <area id="s_anchor1" knowledge="animals" score="0.9" filter="
     typeof=dog" />
 </media>
 <link xconnector="RecognizeStart">
   <bind role="recognize" component="video1"
     interface="s_anchor1" />
   <bind role="start" component="textDog" />
 </link>
</body>
```

Listing 2: NCL code fragment using the proposed approach.

REFERENCES

[1] Antonio Busson, Álan Lívio Guedes, and Sergio Cocher. 2019. Embedding Deep Learning Models into Hypermedia Applications. In *Lecture Notes in Computer Science (LNCS).* (accepted to be published).

[2] ITU. 2009. *H.761: Nested Context Language (NCL) and Ginga-NCL for IPTV Services.* Technical Report. ITU, Geneva, Switzerland. https://www.itu.int/rec/T-REC-H.761