Ontology for Multimedia Applications

Hiranmay Ghosh, *Senior Member, IEEE,* Santanu Chaudhury and Anupama Mallik

***Abstract*—This paper provides an overview of the contents of a tutorial on the subject by one of the authors at WI-2013 Conference. The domination of multimedia contents on the web in recent times has motivated research in their semantic analysis. This tutorial aims to provide a critical overview of the technology, and focuses on application of ontologies for multimedia applications. It establishes the need for a fundamentally different approach for a representation and reasoning scheme with ontologies for semantic interpretation of multimedia contents. It introduces a new ontology representation scheme that enables reasoning with uncertain media properties of concepts in a domain context and a language "Multimedia Web Ontology Language" (MOWL) to support the representation scheme. We discuss the approaches to semantic modeling and ontology learning with specific reference to the probabilistic framework of MOWL. We present a couple of illustrative application examples. Further, we discuss the issues of distributed multimedia information systems and how the new ontology representation scheme can create semantic interoperability across heterogeneous multimedia data sources.**

***Index Terms*—Multimedia, Ontology, Learning, Semantic Modeling, MOWL, Abductive Reasoning, Distributed Systems**

I. Introduction

Use of multimedia data on the web has surpassed that of textual data in the recent times. According to a recent survey [1], 300 million photos are uploaded on the *Facebook* every day and 4 billion hours of video have been watched on *Youtube* per month during the year 2012. These numbers do not include the growing volume of media data generated by surveillance cameras, TV broadcasting stations round the world, satellites, medical imaging devices, document scanners and other digitization initiatives, such as cultural heritage preservation.

The phenomenal rise in consumption of audio-visual data has led to research interest in their semantic processing. Some application examples include creation of personal photobooks [2], [3], news aggregation from multiple sources [4], [5] and digital preservation of cultural heritage [6], [7]. This paper intends to present an insight into the challenges in large-scale semantic processing of multimedia data and the approaches to resolve them. As the media content processing technology advances through content-based, concept-based and ontology-based solutions, the specific requirements for knowledge representation scheme for multimedia applications have been discovered. We present a new multimedia ontology representation scheme [8] that addresses these needs. We show that this new scheme can cope up with the challenges of semantic modeling of multimedia data in different contexts. Learning ontology from real-life data is yet another challenge that is dealt with in this paper with a Bayesian learning framework. Further, we illustrate the effectiveness of the new ontology representation scheme with a couple of illustrative application examples. A major motivation for explicit knowledge representation is integration of information from multiple information sources. We discuss how the new ontology representation scheme is more effective in achieving semantic interoperability across heterogeneous multimedia data sources than the existing approaches.

Hiranmay Ghosh is a Principal Scientist with Innovation Labs, TATA Consultancy Services Limited and is an Adjunct Faculty with Electrical Engineering Department, Indian Institute of Technology, Delhi.
E-mail: hiranmay@ieee.org

Santanu Chaudhury is a Professor with Electrical Engineering Department, Indian Institute of Technology, Delhi
E-mail: santanuc@ee.iitd.ac.in

Anupama Mallik is a Project Scientist with Electrical Engineering Department, Indian Institute of Technology, Delhi
E-mail: ansimal@gmail.com

II. Semantic Web and Ontology

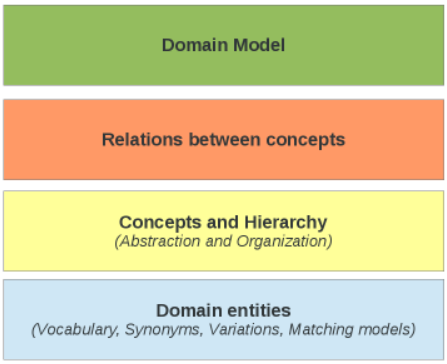


Fig. 1. Layers of abstraction in ontology

The architecture of Semantic Web [9] envisions a world where machines can semantically analyze the data on the web, enhancing the scope of human machine collaboration in specific application contexts. The architecture is based on a *syntactic* layer, where XML is used for describing the data in a uniform way, and a *semantic* layer which relates data items from multiple sources to establish their meanings. An *ontology* that represents an abstract model of a domain, is an essential ingredient of the semantic layer. In context of Information Science, the term "ontology" connotes *formal representation* of knowledge of an *abstraction* of a domain [10]. An ontology defines the "concepts" dealt with in a domain, and establishes their "properties" in context of that domain. Figure 1 depicts the layers of abstraction represented by an ontology. The lowest layer defines the domain entities, i.e. the vocabulary with synonyms, language variations (e.g. "car" or "voiture"), and the matching rules (e.g. use of word-root). The next higher layer brings in abstraction, where concepts are defined and organized into hierarchies. Further up, the properties of the

concepts and their mutual relations are defined and the domain model evolves.

Explicit representation of domain knowledge results in its separation from the program logic. The advantages include generalization and reuse of software agents in multiple domain contexts, convenient knowledge engineering and easier maintenance of knowledge-based applications. The formal specification of domain knowledge enables reasoning with them and discovery of new facts. The relations in an ontology represent rules that can be expressed as First-Order-Logic (FOL). Description Logics (DL) proves to be a convenient tool for logical deductions with such rules. Several techniques for formal knowledge representation had been proposed during the previous decades [11]. W3C has standardized the *Web Ontology Language (OWL)* [12] as the language for ontology representation for semantic interoperability of data on the web in 2004.

In principle, the *concepts* in a domain represent abstract entities and transcended any form of their expressions. But, for any practical use, they need to be represented with some means of communication. Since text is the symbolic representation of human experience and is closest to the abstract model of the world, linguistic constructs (mostly, nouns, verbs and phrases) are used to express the domain model in an ontology. The use of linguistic constructs in representing ontology makes them readily suitable for interpreting text documents in a domain context. Typical uses of ontology in text retrieval and information extraction include query expansion using synonyms, hyponyms (sub-concepts) and hypernyms (super-concepts), creating templates for information extraction, identification of associated concept instances in text documents and resaoning with the discovered facts to find new facts, not explicitly available in the documents.

III. Evolution of multimedia content processing

Multimedia content processing started with content based retrieval systems [13], [14] in early 1990's. These systems provided a *query-by-example* interface and used low level image features, e.g. color and shape, to establish similarity between the query and the database images. It was soon understood that the media features do not represent the semantic contents of the images. The phenomenon is referred to as the *"semantic gap"* in the literature. Several knowledge-based methods have evolved [15] to address this issue. The methods generally involve supervised and unsupervised learning techniques with global or local features. A *bag-of-words* approach [16] creates a "visual vocabulary", when classical information retrieval algorithms can be applied with the "visual words" discovered in a media artifact. Higher level image semantics have been discovered with structural models, e.g. a "beach scene" comprises "sky" at the top and "water" and "sand" below, each of which is characterized by some media features [17]. On the other hand, establishment of the context, e.g. a beach scene, enhances the recognition of constituent objects with similar media features, such as the water and the sky. A part-based human action recognition scheme that exploits context information has been proposed in [18]. Most of the proposed systems attempt to solve domain-specific media interpretation problems with implicit domain knowledge. "Open systems" generally rely on relevance feedback and user profiling data to personalize and to improve on the results.

IV. Ontology for Multimedia Data Interpretation

Incorporation of implicit domain knowledge in multimedia systems and resulting diversity in interpretations hinder semantic integration of information from multiple repositories. With the developments in semantic web technologies, ontology was used to interpret metadata, either manually created or machine generated, in an attempt to achieve semantic interoperability of multimedia artifacts from multiple collections [19]. A logical next step was to extend the ontology with symbolic media properties of the concepts, e.g. a set of color values like "red", "blue", etc. Qualitative relations were established between these media properties, e.g. red is *opposite to* green, but is *close to* brown [20]. Such symbolic property attributions provided limited capability to reason with media properties with concepts. These systems relied on commonality of media annotations, which were available in well-curated media collections in specific domains, e.g. a federation of collaborating museums. Uncontrolled media collections, e.g. those on social networks, do not comply with such requirements. The widespread use of social networks for information sharing has triggered interest in deriving semantics out of crowd-sourced annotations and knowledge organizations [21].

While initial work in creating such ontologies used ad-hoc description schemes, development of MPEG-7 standard [22] provided a mechanism for syntactic compatibility in multimedia content descriptions and motivated creation of ontologies linked to MPEG-7. Since MPEG-7 allows for arbitrary semantic descriptors, a comprehensive visual concept ontology has been proposed in [23] to standardize the vocabulary. To overcome the lack of semantics of XML based MPEG-7 MDS, several research groups created ontologies to formalize the meaning of the multimedia content descriptors. While the different ontologies differed in their coverage and their mode of creation, they can be broadly classified into two classes [24]. Some of the ontologies, e.g. [25], extend themselves to the semantic descriptors of MPEG-7, thereby creating a complete semantic and media based description of multimedia artifacts and collections. This approach poses a challenge for aligning the ontological descriptions for diverse and independently developed repositories. The other MPEG-7 ontologies, e.g. [26], [27], [28], do not include semantic descriptors but focus on media based structural descriptions of the contents. They interoperate with external domain ontologies. This approach has the benefit of using a common domain ontology to interpret media based descriptions of the contents from diverse independent repositories. An architecture for ontology based multimedia data fusion is shown in figure 2.

The approaches for multimedia ontologies described so far create semantic models of repository contents using their MPEG-7 descriptions, but do not attempt to produce a collection independent domain model incorporating multimedia attributes. Another problem with these approaches is the use

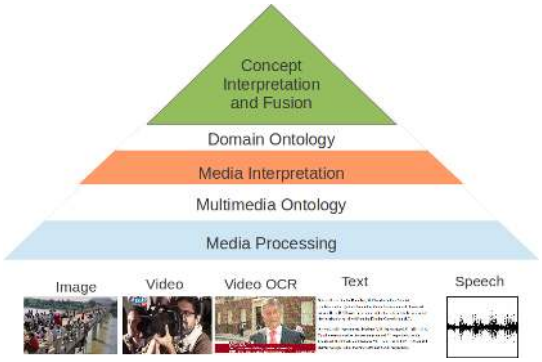


Fig. 2. Ontology based fusion of multimodal data

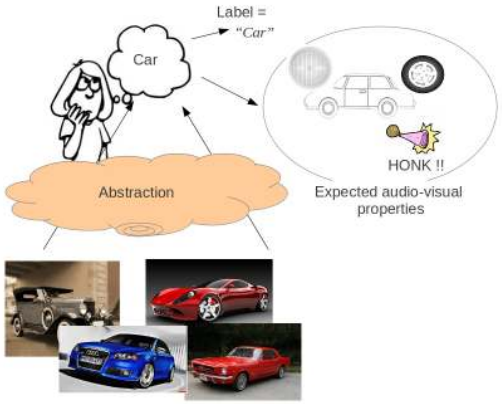


Fig. 3. Perceptual modeling of concepts

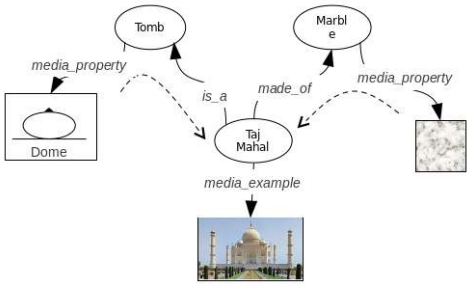
of conventional ontologies that comes with crisp DL based reasoning, which cannot handle the uncertainties associated with the media manifestations of concepts. Following a different approach, domain ontology is extended by [29] with "visual prototypes" or image examples, each of which represent a unique manifestations of a concept. A *query-by-example* search paradigm is used to identify the concepts from the visual contents in a repository. While it is a first step to extending domain ontology to the realm of multimedia, it is quite restrictive in media property specifications. Use of crisp logic for reasoning with interpretation of media contents is another limitation of this approach. Further, an ontology should support reasoning with media properties of concepts, like with the other properties, in a domain context. Media properties of concepts have some special semantics, which are not recognized. We shall shortly discuss the special semantics of media properties.

V. Concept of a "Concept": Perceptual Modeling of Domain

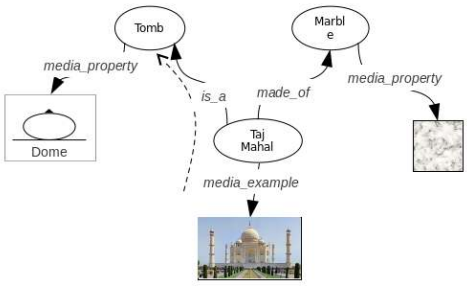
The shortcomings of existing approaches to multimedia ontologies primarily arise from the use of domain description and reasoning techniques that have been developed with text processing applications in view. None of these approaches look into the fundamental needs for knowledge representation in the realm of multimedia data collections. In this context, we note that while text documents are *conceptual* descriptions of human experience, media documents are *perceptual* records of the world, and both are quite dissimilar in nature. The textual descriptions convey the information more crisply than the media instances though they are susceptible to variations in human interpretation and filtering. On the other hand, the media instances are factual records of the world and generally contain a lot more information than text, but they are also likely to contain a lot more noise due to environmental factors. Thus, a conceptual domain model alone cannot cope up with the task of media data interpretation. It needs to be extended to include a *perceptual model*, which may need some different reasoning techniques. The perceptual model of a domain can be the key to bridge the semantic gap between the concepts and their manifestations as media features in multimedia documents.

Though seemingly different, the conceptual model of a domain is not disconnected from the perceptual model, but is derived from the latter [30]. Concepts and concept taxonomies are generated from many observations of the world, mental analysis of their similarities and dissimilarities and the resulting abstractions. An abstract concept is labelled with a natural language construct for the purpose of expression and communication. For example, observation of many cars leads to discovery of some of their common audio-visual properties, which is an abstraction of the concept and which is labeled with a construct, e.g. "car" in a natural language (see figure 3). Further, observation of subtle differences in such audio-visual properties leads to refinement of the concept and formation of concept taxonomy, e.g. "racing car", "vintage car", etc. As a consequence, possibility of manifestation of a concept in a media instance leads to expectation of some common perceptible audio-visual properties. These properties, when observed, leads to a belief in the existence of the concept. For example, a car may be recognized by perceiving one or more of its characteristic audio-visual patterns, e.g. a typical body shape, round wheels and head-lamps, its honk, and so on.

The above observations suggest that the conceptual world is bound to the perceptual world with causal relations. An abstract concept *causes* some perceptible media patterns to appear in multimedia documents. The observation of the media patterns provides evidence towards the concepts in a domain-context. An ontology for multimedia applications needs to encode such causal relations and enable reasoning with them. Further, the media manifestations of concepts are often uncertain and contextual in nature. Thus, it is necessary to incorporate a probabilistic reasoning paradigm with such ontologies. It should also be possible to reason with the media properties in the context of the domain. For example, a monument made of a certain kind of stone is likely to manifest the color and texture properties of the latter. Similarly, the example image of a specific monument is also an example for the generic class to which the monument belongs to (see figure 4). This form of media property inheritance rules are quite distinct from the general property inheritance rules in a concept taxonomy. Moreover, the elementary media properties



(a) Media property propagation



(b) Media example propagation

Fig. 4. Media property and example propagation rules

of a concept often exhibit spatial and temporal relations with each other with some variations in context of the domain. It should be possible to define such spatio-temporal properties in formal yet flexible way in the ontology.

VI. Multimedia Web Ontology Language (MOWL)

A. A conceptual introduction

A new paradigm for perceptual domain modeling with media properties of concepts and for reasoning with the domain model has been proposed in [8] to address the specific needs of knowledge representation for multimedia applications. Since the current ontology languages, e.g. OWL, do not support such model, a new language *Multimedia Web Ontology Language (MOWL)* has been proposed by the authors. The domain model is based on causal relation between the concepts and their media manifestations. Abductive reasoning model with Bayesian network has been proposed for concept recognition to cope up with the uncertainties associated with the causal model.

MOWL supports two types of entities, namely the *concepts* that represent the abstract real world entities and the *media objects* that represent the manifestation of concepts in the media world. For example, while a car can be a concept, its body shape can be a media object. As a special case, visual prototypes as in [29] or example media instances of concepts can also be considered as media objects. Like in other ontology languages, the concepts and the media objects may be organized in a taxonomical hierarchy. The concepts and media objects can have properties. A special class of properties that associates media objects with concepts represent the causal relations in the domain. The uncertainties in such causal relations are captured through a set of conditional probability tables. Another class of properties that relate the concepts signify media property propagation. Such properties can be defined in a domain context. These relations are also probabilistic in nature.

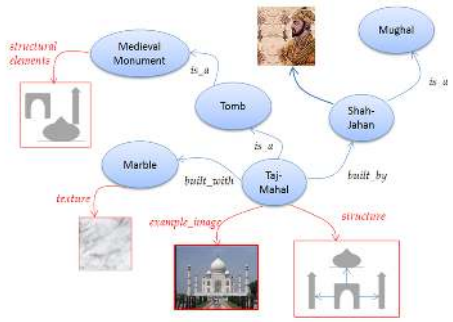
The properties of media objects that represent media manifestation of concepts, can be specified at various levels of complexity. In its simplest form, it can be specified with one of the MPEG-7 elementary audio-visual tools [22]. At the other end of the spectrum, complex media features, e.g. that characterize a dance posture, may need a specially trained classifier. In such cases, a procedural specification or a pointer to an intelligent agent implementing such function may be specified. Another type of complex media property specifications is characterized by spatio-temporal arrangement of simpler media objects. The relative positions of the constituent media objects can have natural variations in different media instances. For example, the relative positions of the dome and the minarets of a monument can be quite different when seen from different perspectives as illustrated in figure 5. MOWL offers constructs to create formal definition of such arrangement with flexibility. The definitions are based on a fuzzy variant of interval algebra, which is consistent with and can be executed with an extended MPEG-7 Query Engine proposed in [31]. Media examples that represent different manifestations of a concept as in [29] can also be associated with media object instances, when an example-based search is used for their detection.

Fig. 5. The Tajmahal seen from two perspectives*(source: http://sv.wikipedia.org/wiki/Taj_Mahal)*

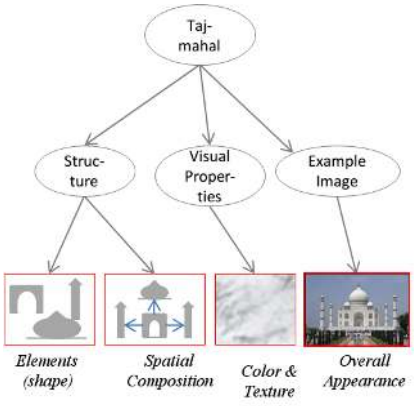
B. Reasoning with MOWL

The causal world model of MOWL prompts an abductive model of reasoning for concept recognition. It is carried out in two steps. In the first step, an *Observation Model (OM)* for a concept is created from the ontology. The OM constitutes the media properties of the concept as well as those of some other related concepts in the domain as determined by the media property propagation rules. The OM is organized as a Bayesian network with the concept at the root node and the expected media properties for that concept at the leaf nodes. Figure 6 shows a possible OM created from a multimedia ontology for the monument “Tajmahal”.

In the second step, each media instance is processed with appropriate feature extraction routines to detect the media properties specified at the leaf nodes of the OM. A leaf node is instantiated when the corresponding media pattern is detected, resulting in a belief revision in the Bayesian network. The posterior probability of the root node as a result of such media property detections signifies the belief in the concept in a multimedia document instance.



(a) Sample domain ontology



(b) A possible Observation Model

Fig. 6. Ontology and Observation Model

C. Discussions

The main difference of MOWL with MPEG-7 based multimedia ontology representation schemes is that the former can be used to model a domain with media properties of concepts, independent of any collection. In this sense, it is similar to the approach presented in [29]. While the latter allows visual prototypes as the only mechanism of media property specification of concepts, MOWL generalizes it to different types of property specifications, including audio-visual examples and MPEG-7 descriptors. Thus, MOWL can be used to interpret MPEG-7 based content descriptions, wherever available. Unlike normative definition of spatio-temporal relations that are used to express the structural composition of events in MPEG-7 informally, MOWL provides for formal yet flexible definition of such relations. Further, the method for media property specification in MOWL can virtually be extended to any type of media properties by using procedural specification. Note that association of different types of media properties with a concept in MOWL provides a natural solution to multi-modal concept recognition and cross-media associations.

Another important attribute of MOWL is reasoning with media properties in a domain context. Media property propagation rules help in creation of Observation Models for concepts incorporating context information. We shall discuss its importance in more details in the next section. While use of Bayesian reasoning is not uncommon for concept recognition in multimedia instances, dynamic creation of the Bayesian network in a domain context is a novelty in MOWL.

VII. Modeling multimedia semantics

Ontological reasoning in the multimedia domain addresses the problem of exploiting information embedded in multimedia assets and making the underlying meaning of the multimedia content explicit. However, the process of attaching meaning to multimedia content is not simple, not even well determined. For example, meaning of an image is not just determined on the basis of image data but also on the situation or context under consideration. Multimedia web ontology language provides a mechanism to attach semantics to the content by specifying possible content-dependent observables of concrete or abstract concepts. For example, we can associate several observable multimodal features, e.g. visual body-shape and typical *huff and puff* audio track with the categorical concept of steam engine. Ontological reasoning scheme of MOWL also facilitate specification of possible contexts for the steam engine, e.g. feature specifications for a pair of railway tracks or human activities in a railway station. Using these specifications, we can search for possible occurrence of steam engine in multimedia assets such as videos, provided that we have appropriate signal analysis algorithms for detection of huff and puff sound and other specified features. Feature detectors essentially embody techniques for distinguishing specific type of signal instances. Machine learning techniques can be used for building such classifiers and detectors. These classifiers and feature detectors provide the initiation point for semantic modeling of multimedia content in the context of ontological reasoning. MPEG-7 standard provides a scheme for specifying such descriptors but does not address the problem of generation of descriptors. These descriptors can encode semantic models at different levels of abstraction. For example, waterfront, as in LSCOM vocabulary [23], can be specified as the corresponding image classifier in the MOWL ontology at the lowest level. This is the key distinguishing feature of MOWL which enables semantic model construction in a hierarchical fashion linking higher level concepts with low level multimedia data.

As an example, we examine the way using which we can represent the concept of human action using the framework described above. We shall use the scheme proposed in [18] for detecting human action in images. Usually verbs indicate human actions; action part is associated with objects related to the action. For example, verb “riding” associated with “bike” indicates human action of riding bike; replacing bike by horse indicates riding horse. In MOWL, the node “riding” can have two specialization nodes bike-riding and horse-riding indicating two different actions. We can associate image based observables to these nodes using the scheme proposed in [18]. Given an image of a human action, many attributes and parts contribute to the recognition of the corresponding action. Actions are characterized by co-occurrence statistics of objects. For example, the “riding attribute is likely to occur together with objects such as “horse and “bike, but not with, say “laptop. Similarly, the “right arm extended upward” is more likely to co-occur with objects such as “volleyball. These interactions of action attributes and parts have been modeled as action bases for expressing human actions in [18]. A particular

action in an image can therefore be represented as a weighted summation of a subset of these bases. The parent node can be represented as weighted summation of union of the subsets of children. In fact, error between reconstruction and test image can be normalized to contribute evidential support. MOWL also provides for specifying observable features for human action in other modalities like text with the same nodes. These features can be used for establishing context with reference to the text associated with an image for a multi-modal multimedia document.

VIII. Learning Multimedia Ontology

An ontology representing concepts and relations of the domain can be hand-coded with inputs from a team of domain experts. Such an ontology may be biased by the opinions of the experts and may not reflect the domain model accurately. This motivates learning of ontology from real-world examples. At another extreme, an ontology learnt from the sample data may not reflect the *human knowledge* of the domain and may be unwieldy. Thus, refinement of a hand-coded ontology with real-world data as an iterative process is considered to be a pragmatic solution to the problem [32].

Machine learning of ontology is essentially a statistical learning process. Probabilistic framework of MOWL is well amenable to it. An Observation Model created from a MOWL ontology models the causal relation between the concept and its possible media manifestations in the form of a Bayesian network. There has been several approaches to ontology learning using Bayesian network. These methods can be used to redefine an Observation Model and in turn, to refine the ontology.

A class of work on Bayesian network learning concentrate on redifining the CPT's in the Bayesian network without changing the network topology. Another class of work, generally referred to as full Bayesian network learning, attempts to discover new relations between concepts (and might drop some existing ones). This approach impacts the network topology. Refining a MOWL ontology can take either of the two forms. A method to update the CPT's in MOWL ontology from implicit user feedback in an retrieval application has been proposed in [33]. In this example, user click-through data has been used to collect implicit user feedback and the ontology is tuned to reflect a specific user's information preferences. A method for full Bayesian network learning in context of a cultural heritage archive has been proposed in [34]. In this example the relations and CPT's of a hand-crafted ontology have been updated using a labelled set of videos depicting classical music and dance.

IX. Application Examples

A. Digital Heritage Preservation

Ontologies have been used in digital museum projects [35], [19] to reason with the domain entities for effective utilization of the digital assets. A shortcoming in these systems is that they cannot reason with the multimedia representations of the artifacts and depend completely on the annotations. In order to deal with this problem, MOWL has been used to model the domain ontology for annotation and semantic navigation in an audio-visual archive *Nrityakosha* of Indian classical dance [7]. Figure 7 depicts an architecture of the system.

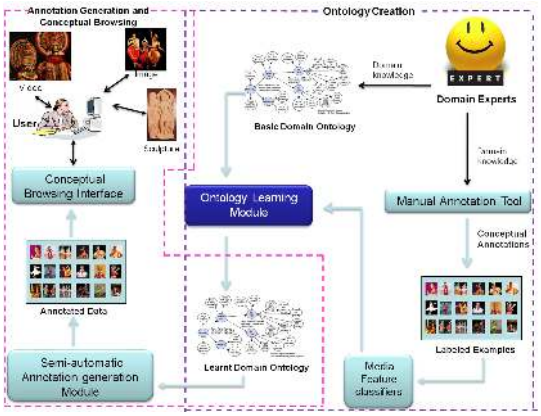


Fig. 7. Architecture of *Nrityakosha*

The domain model of *Nrityakosha* relates various entities, such as dance forms and the accompanying music as well as the myths and the roles that are depicted in those dances. The various concepts *manifest* in some *portrayals*, such as attire of the artistes, dance steps, body postures and musical themes, which are characterized by some audio-visual patterns in the media artifacts. While there is a well-defined grammar for Indian classical dance, individual artistes make their experiments and exercise some freedom resulting in variations to the dance steps. The perceptual and causal model of MOWL has definite advantage over existing ontology languages for such concept recognition tasks. The dance steps are often characterized by a temporal sequence of dance postures with some uncertainties, which can be formally and flexibly expressed with MOWL. Media property propagation rules allow property attributes to "flow" from concepts in mythical stories and roles to the dance steps and postures. While the ontology is initially hand-crafted, it has been refined using the ontology learning method described in the previous section with a corpus of labelled data.

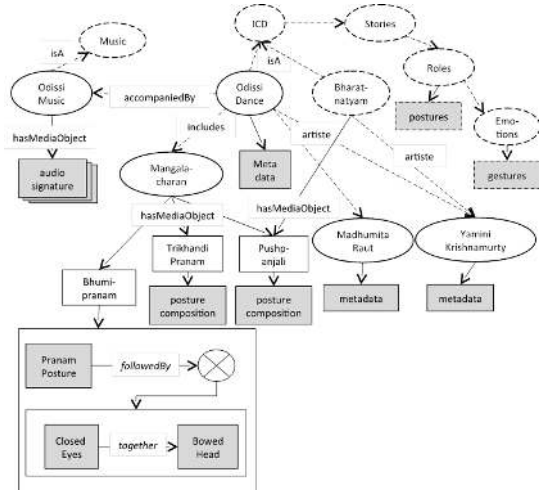


Fig. 8. An ICD Ontology Snippet

To illustrate the use of MOWL in modeling the domain,

let us consider a classical dance form *Odissi* that is typically characterized by an opening act of *Mangalacharan* (invoking the gods). *Mangalacharan* is performed as a combination of three dance steps, each of which manifests in a series of postures. Each of these elementary postures can be detected using a trained set of classifiers. Thus, the dance form *Odissi*, the act *Mangalacharan* and its constituent steps can be modeled as concepts. They are evidenced by the *observable* postures and their sequences, which can be modeled as media objects. The classifiers used to recognize the postures can be expressed as "procedural specification" in MOWL. A few concepts, media objects and their relationships are depicted in figure 8. The edges connecting the concepts with their expected media manifestations are causal and are marked with uncertainties.

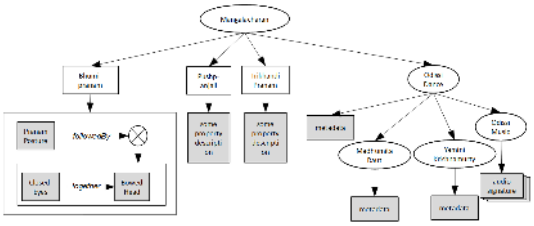


Fig. 9. Observation Model for *Mangalacharan*

Observation Models for concepts like *Mangalacharan* can be constructed from this ontology (see figure 9) and be used for concept recognition. Note that the OM comprises renderings of constituent dance steps (which are temporal sequence of postures) as well as contextual evidences of *Odissi* dance form. A major advantage of this approach is that a concept is recognized with a multitude of evidences, including contextual ones. As a result, failure of a feature detector because of environmental noise has little impact on the overall recognition performance. Further, the elementary postures constituting a higher level concept, e.g. a dance step, have more definitive features than the latter, and it is possible to build more accurate classifiers for them. Deployment of such classifiers and reasoning with their spatio-temporal composition improves the performance of detection of the higher level concepts. Robust concept recognition for audio-visual assets has been used in *Nrityakosha* for their semantic annotation and for establishing their semantic linkages.

B. Product Recommendation for Feature-rich Commodities

Content based filtering technique for product recommendation involves semantic matching of user profile and product features. The semantic associations of features with product categories are quite complex in many domains, such as fashion. Ontology based approaches for apparel recommendation have been presented in [36], [37]. The crisp ontological classification and the first-order reasoning rules deployed in these systems are inflexible to capture the subjectivity and uncertainty associated with choice of apparels. Moreover, they fail to deal with the "look and the feel" (visual and tactile properties), which are important selection parameters for the garments An apparel recommendation system based on perceptual modeling scheme of MOWL is presented in [38].

Fig. 10. Ontology for garment recommendation

Figure 10 shows a high-level view of the fashion ontology that incorporates knowledge about human users, occasion to wear and the garments. Visual attributes have been associated with humans and garments. Garments have been organized in several categories and several visual attributes have been assigned to them. The recommendation rules are based on *Color Season Model* [39] and other information sources.

The recommendation problem is handled in two steps in the system. First, an OM for user visual profile is created and the latter is determined based on observations on user body parameters such as skin color and body shape. Then an OM for the garment (to be recommended) is created by incorporating the discovered user profile. This OM has garment properties, e.g. color, texture, material, etc. as its observable property nodes. The garment catalog is consulted and the garment attributes (both visual and semantic) are analyzed to instantiate the property nodes in the OM. The garments that have highest posterior probability based on analysis of the garment properties qualify for recommendation. Figure 11 show the recommendation results for *Sarees*[1] for an Indian celebrity for different occasions.

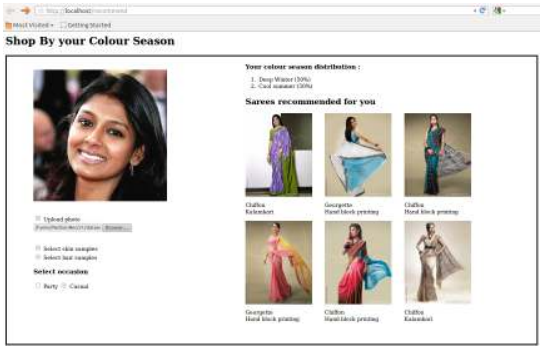


Fig. 11. Results for Apparel Recommendation

This approach provides quite a few benefits as compared to SVM based [36] or SWRL rule-based [37] recommendation. In the first place, the domain rules need not be exhaustively enumerated and it is sufficient to encode the rules connecting the broad classes. MOWL helps in reasoning with the media

[1]An ethnic wear for women popular in South Asia

properties of the concepts like garments, humans and occasions. Further, the abductive reasoning used in MOWL is robust to make recommendations even if all garment properties are not listed in the catalog. Most importantly, the causal probabilistic reasoning enables ranking of the recommendations allowing for user preferences.

X. Distributed Multimedia Applications

Many applications need to integrate information from multiple independent sources, including social media, to meet the user needs. Examples include travel services [40], news aggregation [5], medicine [41] and cultural heritage applications [42]. A multi-agent system [43] is a convenient tool to model such systems. The architecture of typical agent based system used for information gathering from multiple sources is shown in figure 12. The User agent does a pre-processing of the user request before forwarding it to the Broker agent. The Broker agent interprets the request with a background domain knowledge encoded in the form of an ontology and interacts with the Resource agents to retrieve the necessary information, often iteratively. There is usually a good deal of redundancy in Resource agents on the web. Some criteria may be applied to select a limited number of Resource agents to participate in the information gathering process. Finally, the Broker agent semantically integrates the information from multiple sources before reverting to the user.

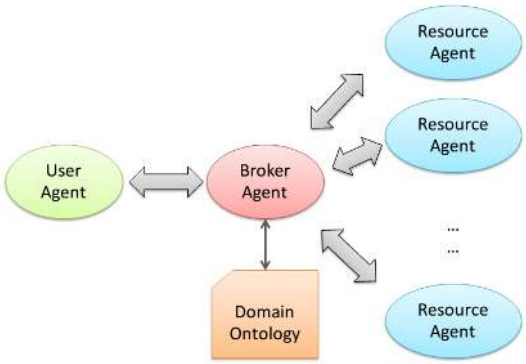


Fig. 12. Architecture on a multi-agent distributed information system

The data to be dealt with in many of the cited domains are often in multimedia format motivating their semantic integration in specific application contexts. MPEG-7 linked ontologies discussed in section IV attempt such integration. As shown in figure 2, semantic integration is effected in these systems at the conceptual level, based on the semantic descriptors for the contents (man-made or machine produced) or other forms of metadata [40] and not based on the information content in the media forms. MOWL provides an opportunity for integration based on analysis of media contents. The domain knowledge available in the Broker agent, when encoded in MOWL, can incorporate a perceptual model of the domain. A user request, when interpreted with such domain knowledge, produces an Observation Model that can be used to interpret media contents by the different Resource agents.

An Observation Model created from a non-trivial domain knowledge generally includes many leaf nodes (observable media properties), signifying many different manifestations of the concept. Generally, it is not necessary to observe all such media properties to have a sufficient belief in the concept. The posterior belief in a concept tends to saturate after a few observations and observation of further media properties does not add significantly to the belief value. Thus, it is desirable to create an *observation plan* by choosing an appropriate set of media properties that can result in sufficient posterior belief in the concept at a minimal computational cost. While the effectiveness of an evidence (media pattern) in identifying a concept depends on the domain knowledge, the computational cost and feasibility for its detection depends on the contents and the data organization in the Resource agents. A method to create resource-specific observation plans, considering both the aspects, using a distributed planning algorithm has been proposed in [44]. The Resource agent that has the potential to produce reliable results within some constraints of computational cost bids for participation. An interesting consequence of such planning is that, while an Observation Model for a concept, say steam locomotive, will contain both audio and visual patterns, the observation plan for an image repository will use some of the visual patterns only, but that for a video repository can use both. While different observation plans are executed by different Resource agents, all of them are derived from the same domain ontology. This facilitates information integration of multi-modal data from multiple sources.

The knowledge about the context is often distributed across multiple agents. For example, while the Broker agent encapsulates the domain ontology, the User agent might model a user profile that incorporates the knowledge about the user's implicit preferences [33]. The Resource agents may include a semantic data model for the contents in their repository [25]. In general, these independently created ontologies employ disparate data models. They need to be aligned to ensure their interoperability. The ontology alignment problem can be stated as discovery of equivalence and subsumption relationship between pairs of entities from two independent ontologies and application of the discovered mapping rules [45]. The equivalence of concepts are generally discovered by establishing context similarity (structure of the ontology graph around the concept), the equivalence of individuals are generally based on commonalty of properties.

An interesting approach to establish relation between entities in different ontologies that has not yet been explored well is by comparing their perceptual properties. Perceptual modeling of domain using MOWL presents such opportunity. While the terminology used to describe a concept can be different in different ontologies and the ontological relations for the concept can be domain dependent, the perceptual properties of a concept are expected to be invariant. Thus, two concepts can be said to be equivalent if the Observation Models for two concepts are similar [46]. Note that the Observation Model of a concept incorporates media properties of related concepts and can thus be used to compare the structural context of the concepts.

XI. Conclusions

Despite significant advances in media content analysis over the last couple of decades, a solution to the problem of "semantic gap" still eludes the researchers. Semantic analysis of media forms is still a subject of vigorous research. Fusion of multi-modal data from heterogeneous and distributed resources poses a much bigger challenge. It appears that an ontology put on top of media analysis services is not a suitable solution. Multimedia Web Ontology Language is a first step towards semantic analysis and integration of multimedia data from information sources in an open Internet environment. While MOWL presently currently deals with audio and visual data, the theoretical framework has the generality to deal with any form of sensor data, thus paving the way for semantic fusion of multi-modal multi-sensor data. The framework further needs to be extended to incorporate other facets of multimedia event models as proposed in contemporary literature [47], [48].

References

[1] Pingdom, "Internet 2012 in numbers," http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/, 2013.

[2] H. Wang, L.-T. Chia, and S. Liu, "Image retrieval ++–web image retrieval with an enhanced multi-modality ontology," *Multimedia Tools Appl.*, vol. 39, no. 2, pp. 189–215, Sep. 2008. [Online]. Available: http://dx.doi.org/10.1007/s11042-008-0202-7

[3] M. Rabbath, P. Sandhaus, and S. Boll, "Automatic creation of photo books from stories in social media," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7S, no. 1, pp. 27:1–27:18, Nov. 2011. [Online]. Available: http://doi.acm.org/10.1145/2037676.2037684

[4] O. Conlan, I. O'Keeffe, and S. Tallon, "Combining adaptive hypermedia techniques and ontology reasoning to produce dynamic personalized news services," in *Proceedings of the 4th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, ser. AH'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 81–90.

[5] R. Bansal, R. Kumaran, D. Mahajan, A. Khurdiya, L. Dey, and H. Ghosh, "Twipix: a web magazine curated from social media," in *Proceedings of the 20th ACM international conference on Multimedia*, ser. MM '12. New York, NY, USA: ACM, 2012, pp. 1355–1356. [Online]. Available: http://doi.acm.org/10.1145/2393347.2396482

[6] M. Hatala, L. Kalantari, R. Wakkary, and K. Newby, "Ontology and rule based retrieval of sound objects in augmented audio reality system for museum visitors," in *Proceedings of the 2004 ACM symposium on Applied computing*, ser. SAC '04. New York, NY, USA: ACM, 2004, pp. 1045–1050.

[7] A. Mallik, S. Chaudhury, and H. Ghosh, "Nrityakosha: Preserving the intangible heritage of indian classical dance," *J. Comput. Cult. Herit.*, vol. 4, no. 3, pp. 11:1–11:25, Dec. 2011. [Online]. Available: http://doi.acm.org/10.1145/2069276.2069280

[8] A. Mallik, H. Ghosh, S. Chaudhury, and G. Harit, "Mowl: An ontology representation language for web based multimedia applications," *ACM Transactions on Multimedia Computing, Communications and Applications*, (In press).

[9] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, May 2001.

[10] T. R. Gruber, "A translation approach to portable ontologies," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[11] F. van Harmelen, V. Lifschitz, and B. Porter, Eds., *Handbook of Knowledge Representation*. Elsevier, 2008.

[12] "Owl web ontology language guide," http://www.w3.org/TR/owl-guide/, 2004.

[13] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The qbic system," *Computer*, vol. 28, no. 9, pp. 23–32, Sep. 1995. [Online]. Available: http://dx.doi.org/10.1109/2.410146

[14] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. fe Shu, "The virage image search engine: An open framework for image management," in *Proceedings of the SPIE Conference on Visual communication and Image Processing*, 1996, pp. 76–87.

[15] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recogn.*, vol. 40, no. 1, pp. 262–282, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2006.04.045

[16] J. Sivic, "Efcient visual search of videos cast as text retrieval," *IEEE Trans. on PAMI*, vol. 31, no. 4, pp. 591–605, April 2009.

[17] J. Fan, Y. Gao, H. Luo, and G. Xu, "Statistical modeling and conceptualization of natural images," *Pattern Recogn.*, vol. 38, no. 6, pp. 865–885, Jun. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2004.07.011

[18] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1331–1338. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2011.6126386

[19] M. Doerr, "The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata," *AI Mag.*, vol. 24, no. 3, pp. 75–92, Sep. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=958671.958678

[20] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, "Sweetening ontologies with dolce," in *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, ser. EKAW '02. London, UK, UK: Springer-Verlag, 2002, pp. 166–181. [Online]. Available: http://dl.acm.org/citation.cfm?id=645362.650863

[21] H. Xu, X. Zhou, M. Wang, Y. Xiang, and B. Shi, "Exploring flickr's related tags for semantic annotation of web images," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '09. New York, NY, USA: ACM, 2009, pp. 46:1–46:8. [Online]. Available: http://doi.acm.org/10.1145/1646396.1646450

[22] F. Nack, J. v. Ossenbruggen, and L. Hardman, "That obscure object of desire: Multimedia metadata on the web, part 2," *IEEE MultiMedia*, vol. 12, no. 1, pp. 54–63, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1109/MMUL.2005.12

[23] J. R. Smith, "Largescale concept ontology for multimedia," *IEEE Multimedia Magazine*, vol. 13, no. 3, pp. 86–91, July–September 2006.

[24] S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, and M. G. Strintzis, "Enquiring mpeg-7 based multimedia ontologies," *Multimedia Tools Appl.*, vol. 46, no. 2-3, pp. 331–370, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1007/s11042-009-0387-4

[25] C. Tsinaraki, P. Polydoros, and S. Christodoulakis, "Interoperability support between mpeg-7/21 and owl in ds-mirf," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 2, pp. 219–232, Feb. 2007. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.33

[26] J. Hunter, "Enhancing the semantic interoperability of multimedia through a core ontology," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 13, no. 1, pp. 49–58, Jan. 2003. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2002.808088

[27] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura, "Comm: designing a well-founded multimedia ontology for the web," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ser. ISWC'07/ASWC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 30–43. [Online]. Available: http://dl.acm.org/citation.cfm?id=1785162.1785166

[28] K. Dalakleidi, S. Dasiopoulou, G. Stoilos, V. Tzouvaras, G. Stamou, and Y. Kompatsiaris, "Knowledge-driven multimedia information extraction and ontology evolution," in *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, G. Paliouras, C. D. Spyropoulos, and G. Tsatsaronis, Eds. Berlin, Heidelberg: Springer-Verlag, 2011, ch. Semantic representation of multimedia content, pp. 18–49. [Online]. Available: http://dl.acm.org/citation.cfm?id=2001069.2001071

[29] M. Bertini, A. D. Bimbo, G. Serra, C. Torniai, R. Cucchiara, C. Grana, and R. Vezzani, "Dynamic pictorially enriched ontologies for digital video libraries," *IEEE MultiMedia*, vol. 16, no. 2, pp. 42–51, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1109/MMUL.2009.25

[30] H. Kangassalo, "Conceptual level user interfaces to data bases and information systems," in *Advances in information modelling and knowledge bases*, H. Jaakkola, H. Kangassalo, and S. Ohsuga, Eds. IOS Press, 1991, pp. 66–90.

[31] S. S. Wattamwar and H. Ghosh, "Spatio-temporal query for multimedia databases," in *Proceedings of the 2nd ACM workshop on Multimedia semantics*, ser. MS '08. New York, NY, USA: ACM, 2008, pp. 48–55. [Online]. Available: http://doi.acm.org/10.1145/1460676.1460686

[32] F. Wu and D. S. Weld, "Automatically refining the wikipedia infobox ontology," in *Proceedings of the 17th international conference on World*

Wide Web, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 635–644. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367583

[33] H. Ghosh, P. Poornachander, A. Mallik, and S. Chaudhury, "Learning ontology for personalized video retrieval," in *Workshop on multimedia information retrieval on The many faces of multimedia semantics*, ser. MS '07. New York, NY, USA: ACM, 2007, pp. 39–46. [Online]. Available: http://doi.acm.org/10.1145/1290067.1290075

[34] A. Mallik and S. Chaudhury, "Acquisition of multimedia ontology: an application in preservation of cultural heritage," *IJMIR*, vol. 1, no. 4, pp. 249–262, 2012.

[35] L. Schneider, "Designing foundational ontologies - the object-centered high-level reference ontology ochre as a case study," in *in Conceptual Modeling 2003, 22nd International Conference on Conceptual Modeling, I.-Y.* Springer, 2003, pp. 91–104.

[36] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan, "Hi, magic closet, tell me what to wear!" in *Proceedings of the 20th ACM international conference on Multimedia*, ser. MM '12. New York, NY, USA: ACM, 2012, pp. 619–628.

[37] D. Vogiatzis, D. Pierrakos, G. Paliouras, S. Jenkyn-Jones, and B. J. H. H. A. Possen, "Expert and community based style advice," *Expert Systems with Applications: An International Journal*, vol. 39, no. 12, pp. 10 647–10 655, September 2012.

[38] S. Ajmani, H. Ghosh, A. Mallik, and S. Chaudhury, "An ontology based personalized garment recommendation system," in *Workshop on Personalization, Recommender Systems and Social Media*, November 2013.

[39] B. Kentner, *Color me a season: A complete guide to finding your best colors and how to use them*. Ken Kra Publishers, 1979.

[40] Y. T. F. Yueh, D. K. W. Chiu, H.-f. Leung, and P. C. K. Hung, "A virtual travel agent system for m-tourism with semantic web service based design and implementation," in *Proceedings of the 21st International Conference on Advanced Networking and Applications*, ser. AINA '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 142–149. [Online]. Available: http://dx.doi.org/10.1109/AINA.2007.25

[41] H. Kosch, R. Slota, L. Böszörményi, J. Kitowski, J. Otfinowski, and P. Wójcik, "A distributed medical information system for multimedia data - the first year's experience of the parmed project," in *Proceedings of the 8th International Conference on High-Performance Computing and Networking*, ser. HPCN Europe 2000. London, UK, UK: Springer-Verlag, 2000, pp. 543–546. [Online]. Available: http://dl.acm.org/citation.cfm?id=645564.658465

[42] X. Pan, T. Schiffer, M. Schröttner, R. Berndt, M. Hecher, S. Havemann, and D. W. Fellner, "An enhanced distributed repository for working with 3d assets in cultural heritage," in *Proceedings of the 4th international conference on Progress in Cultural Heritage Preservation*, ser. EuroMed'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 349–358.

[43] E. Shakshuki, M. Kamel, and H. Ghenniwa, "A multi-agent system architecture for information gathering," in *Proceedings of the 11th International Workshop on Database and Expert Systems Applications*, ser. DEXA '00. Washington, DC, USA: IEEE Computer Society, 2000, pp. 732–. [Online]. Available: http://dl.acm.org/citation.cfm?id=789091.790428

[44] H. Ghosh and S. Chaudhury, "Distributed and reactive query planning in r-magic: An agent-based multimedia retrieval system," *IEEE Trans. on Knowl. and Data Eng.*, vol. 16, no. 9, pp. 1082–1095, Sep. 2004. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2004.40

[45] V. Ermolayev and M. Davidovsky, "Agent-based ontology alignment: basics, applications, theoretical foundations, and demonstration," in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, ser. WIMS '12. New York, NY, USA: ACM, 2012, pp. 3:1–3:12. [Online]. Available: http://doi.acm.org/10.1145/2254129.2254136

[46] B. Maiti, H. Ghosh, and S. Chaudhury, "A framework for ontology specification and integration for multimedia applications," in *Proceedings of Knowledge Based Computer System (KBCS 2004)*, December 2004.

[47] V. Mezaris, A. Scherp, R. Jain, M. Kankanhalli, H. Zhou, J. Zhang, L. Wang, and Z. Zhang, "Modeling and representing events in multimedia," in *Proceedings of the 19th ACM international conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 613–614. [Online]. Available: http://doi.acm.org/10.1145/2072298.2072391

[48] P. Appan and H. Sundaram, "Networked multimedia event exploration," in *Proceedings of the 12th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '04. New York, NY, USA: ACM, 2004, pp. 40–47. [Online]. Available: http://doi.acm.org/10.1145/1027527.1027536