Contextual Augmentation of Ontology for Recognizing Sub-Events

Setareh Rafatirad
Information and Computer Science Department
University of California Irvine
Irvine, USA
Email: srafatir@ics.uci.edu

Ramesh Jain
Information and Computer Science Department
University of California Irvine
Irvine, USA
Email: jain@ics.uci.edu

***Abstract*—With the advances in technology and proliferation of cheap storage, high rate of digital multimedia interaction signifies the increasing need of computer users for a decent application to organize personal media in a meaningful way. In this paper, we want to organize personal media in terms of the sub-events they cover. A semantic gap exists between media, and people's perception of the events and memories associated with this media. A framework is needed to address such gap. This paper describes a novel model-based approach for partitioning and organizing personal photo archive in terms of high-level subevents that capture and represent human experience. Since photos are the most ubiquitous and prolific form of user generated content, we focus on the automatic annotation of personal photo collection in this paper. We introduce R-Ontology (Recognition-Ontology) that is a context-aware model with concrete contextual information for subevent recognition. Currently our approach utilizes the mereological, spatial and temporal properties of modeled-events in R-Ontology. Personal media will then populate R-Ontology. We tested this approach using our personal photo archive describing two different scenarios: Trip and Indianwedding.**

***Keywords*-Ontology Instance Augmentation; R-Ontology; Filtering; Organizing; Contextual Information; Flexible Modeling;**

I. Introduction

A user study [22] shows that events are important means to recall photographs. We want to organize personal media in terms of the events they cover. In recent years, we have faced the explosion of user-generated content (UGC): 3.7B photos on Facebook and Flickr [16], and 60M of uploaded photos to Facebook per week [17] are the indications of large scales of UGC. This trend signifies the need for a personal media management system that can organize the UGC to bridge the semantic gap. Current media management applications such as Picasa, iPhoto, and MyLifeBits[1] manage the stored personal media to some extent. There is no doubt that such systems have had significant impact on the media management trend but there is still a lot to go from here because of the following reasons:

- Information sources have become available extensively on the web, however there has not been a significant effort on utilizing them in photo organizing applications.
- Current applications can not organize photos using the subevent structure of a life chronicle event a person participates; there is still a gap between visual data and richness of human semantics.

This problem cannot be solved only with machine learning techniques because they only work well for low-level semantics, whereas this problem addresses high-level semantics. According to a survey on recent technical achievements in high-level semantic-based image retrieval [20], so far no generic approach is proposed for such retrieval. To tackle this challenge, we introduce a novel approach that focuses on the usage of ontology as a contextual-model (not a content-model), and uses multi-modal context; also when needed, photo content features are used. The multi-modal context is fed into our model through web-based sources. This approach becomes more relevant to devices like mobile phones where lot more contextual information is available. In fact, this shows the beauty of this approach. Ontologies became important because they describe high-level semantics to recreate, invent, or visually present a person, place, event, or action so that the reader may picture that, which is being described. Ontologies are nothing but formal conceptual models at the "semantic" level that are independent from lower level data models [13]. Ontology is traditionally defined in philosophy as "the study of the kinds of things that exist" [4]. The state of art shows that ontology has been studied beyond philosophy, in AI (NLU) and general software systems, however most of such ontological models are only used for description. We want to shift from using ontology just as a descriptive language, and pick up the pace towards using Ontology for recognition of subevents related to visual data. Subevents are linguistic descriptions and a linguistic description is almost always contextual[15]. Ontology allows explicit specification of models that could be modified using context information to provide very flexible models for recognition of such high-level semantics. Ontologies have been proposed in two main types according to the levels of abstraction: 1) domain, and 2) core/upper ontologies. The former describes relevant concepts of a certain domain e.g., marriage for Wedding domain. In contrast, the latter can be applied to a variety of

[1] http://research.microsoft.com/en-us/projects/mylifebits/

knowledge domains; e.g., DOLCE-Lite describes universal concepts which are applicable to all domains. In recent years, upper ontological frameworks have been proposed as modeling infrastructures in event-based information systems; e.g. VERL [7] for activity recognition in video data, Event Calculus[8] for knowledge representation, and Event Model-F[9] for describing events in event-based multimedia applications. Domain ontologies have also been adopted alongside with core/upper ontologies: a) some are created during ontology acquisition process using machine learning techniques while experiencing the problem of quality assurance that arises from the data-dependency nature of this operation: the given corpus may not be always sufficient; and b) some are created manually. On one hand, these models are introduced; on the other hand they have not become explicitly available to other systems to actually start using them. Some domain ontologies are small variations of concepts in more generic form of domain ontologies; flat architecture for small variations of the same domain ontology results in replicating common concepts. The shaded area of Fig1-(a) reflects this architecture. On the other hand, sticking with only one model for all variations produces unsatisfactory results because it cannot model them all, or even if it does, it will be too complex to process; this is equivalent to say that one generic domain model cannot be used for every variation. We avoid this by proposing a multi-layered (hierarchical) architecture in the shaded area of Fig1-(b). It shows a multi-layered framework to avoid replicating knowledge regarding conceptual variations of a domain. Domain ontology by itself is not enough for recognition since it does not carry the required contextual information (CI) for a particular subevent. An actual event is represented by both visual and contextual semantics. Visual semantics specify the visual concepts used to describe the event. CI are textual data that evaluate the non-visual properties of events such as absolute/relative time, location/place, participants,structure, etc. Domain Ontology is a DAG by design. Upper and domain ontological models mostly support the types and relationships. Notice that NOT all contextual data is hardcoded in these models. We divide CI into two groups: 1) constant-CI (CI_c), 2) variable-CI (CI_v). CI_c are non-changing textual properties that describe knowledge in a logically consistent and constant manner that remain the same through all instances of the same model (e.g. subevent-structure, relative time between subevent classes); therefore they exist in the domain ontology. In contrast, CI_v are changing textual data that vary through instantiation of the same model (e.g. time-interval and location of "visiting Forbidden city in China" is different from "visiting Great wall of China", although they are instances of the same class "visiting"). These concepts are formally defined in the following way:

$$\begin{aligned} &Ontology\ Instance:\ O_i = (I, R) \\ &R - Ontology:\ O_r = (I \times_A CI_v, R), \\ &I \times V = \{(i_1, c_1), ..., (i_n, c_n)\}, i_k \in I, c_k \in CI_v \end{aligned} \quad (1)$$

In (1) we formally define ontology instance and R-Ontology. I is a set of instances of the classes in the underlying domain ontology (O_d); R is the set of relationships between the instances in I. In (1), O_r is generated by augmenting O_i using the CI_v extracted from identified data sources. O_r is a DAG in which I is the set of nodes and R is the set of edges between nodes. Augmentation is shown by the $\times_A$ operator whose job is evaluating the properties of members of I. Later in this paper, we will explain how variable contextual information such as absolute time (interval),location (GPS/boundingbox/named-place),and participants is evaluated for ontology augmentation and utilized in addition to the constant context information, in describing and recognizing real world actual events and their subevent.

$$\begin{aligned} &\forall O_i \in O_d : \exists CI_c (CI_c \sqsubseteq O_i) \wedge (CI_c \sqsubseteq O_d) \\ &\forall O_r \exists CI_v\ E(CI_v)\ \wedge (CI_v \not\sqsubseteq O_d)\ \wedge (CI_v \sqsubseteq O_r) \end{aligned} \quad (2)$$

In (2), E(a) is the predicate "a is extracted information". O_r is formed using both constant and variable contextual information. If CI_v is given in the domain ontology, the model is not flexible to describe another instance of the same class. For subevent recognition, we need flexible models that carry CI_v; to generate such models, we need to augment instance of the employed domain model using variable contextual information; we refer to the augmented result as R-Ontology. Fig1-(b) shows a particular context is used to generate O_r for a particular personal wedding, another context is used to generate O_r for another wedding. With augmentation, different contexts in the same domain of discourse can be embraced. The basic analogy of this framework is the notion of polymorphism in the context of object-oriented programming (OOP) because semantics of different augmented instances are handled using a uniform domain model. This framework enables reusing the knowledge in the domain ontology. Our contribution in this paper is proposing a framework that a) creates O_r, and b) populates it with the experiential evidences (within a personal media archive) automatically. Our approach considers available event models, the mereological, relative and/or absolute temporal and spatial properties of modeled-events, data sources containing CI, available multimedia content-models as well as the contextual semantics of photos such as date, time, geo-tag, and imaging parameters using the EXIF metadata. We tested our approach on two photosets. The experimental results looked promising. This paper is organized as follows: Related works are reviewed in section 2. Section 3 introduces the architecture for our proposed framework. Section 4 presents our experimental results on two domain scenarios

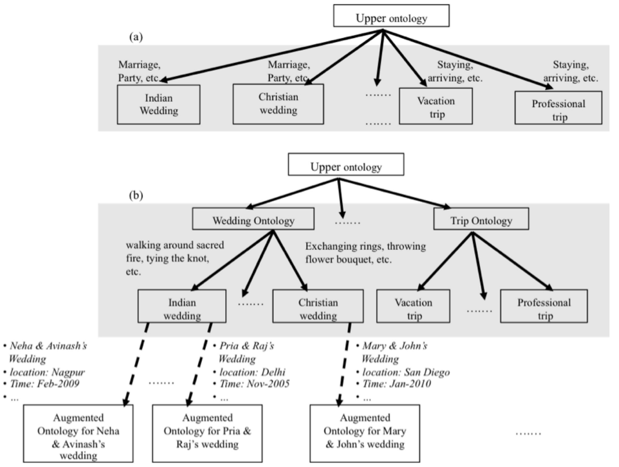

Figure 1. Domain Ontology Architectures and R-Ontology (a) Flat: concept-replication (b) Layered with abstraction levels: no concept-replication, R-Ontology at the bottom level.

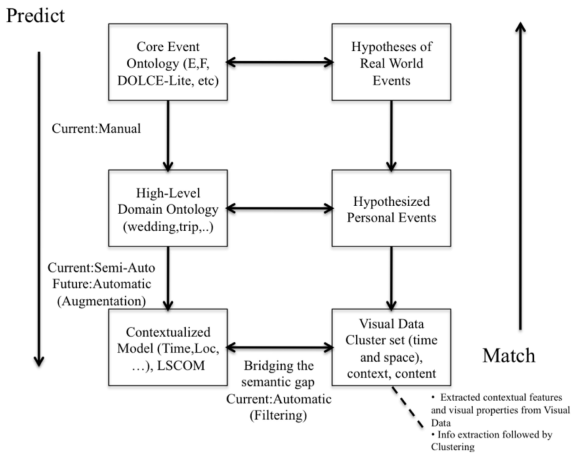
Predict
Core Event Ontology (E,F, DOLCE-Lite, etc)
Hypotheses of Real World Events
Current:Manual
High-Level Domain Ontology (wedding,trip,..)
Hypothesized Personal Events
Current:Semi-Auto
Future:Automatic
(Augmentation)
Contextualized Model (Time,Loc, …), LSCOM
Visual Data Cluster set (time and space), context, content
Match
Bridging the semantic gap
Current:Automatic
(Filtering)
• Extracted contextual features and visual properties from Visual Data
• Info extraction followed by Clustering

Figure 2. Recognition Framework.

”Vacation-Trip” and ”Indian-Wedding”, followed by user-based evaluation in section 5, and finalized by conclusions and future work in section 6.

II. RELATED WORK

This section surveys state of art in ontological recognition and classification. The main theme of this section is to address the lack of actual usage of ontology in event-based multimedia systems. Ontologies are considered in AI statistical applications of NLP. The work has been related to named entity extraction and recognizing events from textual data. [3] proposes an ontology-driven scheme that is used to for describing events on news stories. In [11], existing ontologies are combined with existing textual databases for ontology acquisition, and event structure is employed as the index unit for information retrieval (IR). Such merits are only considered for textual data, not for visual data. Ontology has not been used for visual recognition properly. An application of ontology is geometric object recognition; [10] uses ontology for describing geometric objects for Automatic Target Recognition; the approach uses a very tedious annotation. [19] proposes a pixel-based ontology for recognition of pixel-based objects in images. Except for NLP, machine learning, and knowledge acquisition disciplines, ontologies are studied for description rather than recognition in content-centric multimedia systems e.g., COMM[6] and ABC [5] are for describing information objects (i.e., digital data entities) with multimedia content-centric view that include the concepts defined in MPEG7. In event-based multimedia systems the progress is limited to the ontology creation e.g. F-Model and E-Model [2]. Some progress is made towards activity recognition from surveillance video [18]. In contrast to these merits, we promote the usage of ontology from description to sub-event recognition for visual data. An important issue that makes our approach novel is ontology instance augmentation using available models and contextual data to make the recognition context-aware.

III. SYSTEM OVERVIEW

A. Problem Formulation

”The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation” [15]. Binford hierarchical geometric model [14] is one of the early works in model-based image understanding that is used to create flexible models for object recognition at different feature levels (low-level). This motivates and elaborates the flow and components of our approach in Fig2 by replacing 3D Object Models with Core Ontology, 2D Projections of 3D Models with Domain Ontology, and the next two levels with available content models (e.g., LSCOM[21]) and context-aware recognition models to bridge the semantic gap.

Our problem is formulated as follows: Given photos $P :< p_1, \ldots, p_n >$ with EXIF metadata for an event E , we partition them into its subevents $< se_1, \ldots, se_m >$. We use O_d corresponding to the type of the event, instantiate that using information available for the event (i.e. time, location, participating people), and augment O_i with all available information related to the context of the subevents i.e. CI_v using operator $\times_A$. Finally, O_r is used to partition P. We will answer to the following questions in this work:

- How does operator $\times_A$ create O_r?
- How can O_r be employed for partitioning P?

According to the flow of our proposed framework in Fig2, O_r is the output of augmentation operation. We need this model for sub-event recognition. To apply this model on a large photo collection, it is necessary to extract recognizable features from photos. To decrease the computation time, we

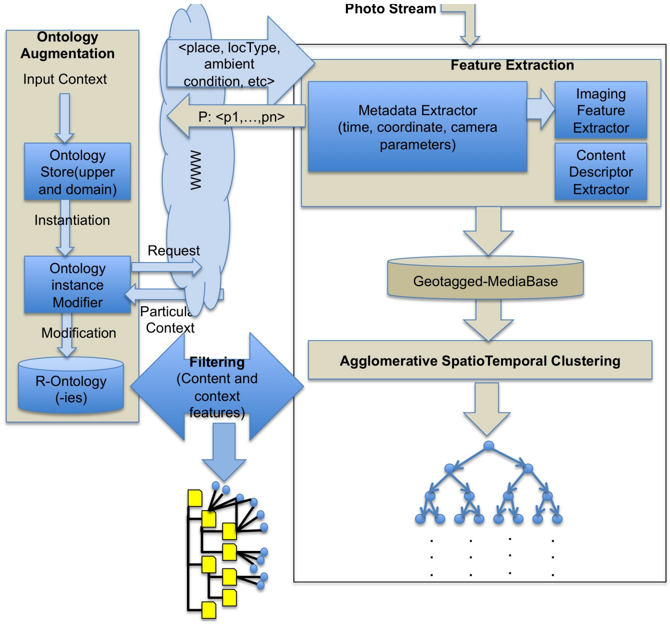
Photo Stream
Ontology Augmentation
Input Context
Ontology Store(upper and domain)
Instantiation
Ontology instance Modifier
Modification
R-Ontology (-ies)
<place, locType, ambient condition, etc>
P: <p1,…,pn>
WWW
Request
Particular Context
Filtering (Content and context features)
Feature Extraction
Metadata Extractor (time, coordinate, camera parameters)
Imaging Feature Extractor
Content Descriptor Extractor
Geotagged-MediaBase
Agglomerative SpatioTemporal Clustering

Figure 3. General Framework.

run agglomerative clustering on some of the extracted features with O(log n) computation time; it creates a cluster-tree at the end. Operation Filtering assigns the proper clusters to the events in O_r and filters the irrelevant photos in each assigned cluster according to the described event quality in O_r. The final result is a hierarchical event topology in which nodes are associated with the relevant visual data. These operations constitute the building blocks of our proposed framework indicated in Fig3.

B. Upper and Domain Ontology

We used a basic derivation of E^* [23] as our upper ontology that contains the fundamental relationships associated with Perdurants : occurs-during, occurs-at, has-subevent, has-Quality, has-participants ,has-objects, has-processingUnit. The temporal model for absolute and relative time-interval is derived from E^*; absolute interval is represented by standard attribute-pair *start-time* and *end-time* through *occurs-during* relationship, and the basic form of relative interval is expressed by standard temporal predicates *after* and *before*. Using the relative time and the reference interval, a new temporal entity is created and associated to one or more existing entities. Relative interval is associated with a reference time interval. In fact a relative interval is initially implicitly expressed until the referencing interval is evaluated e.g. if *A* and *B* are both Perdurants, and *B* is associated with a relative interval that references *A.I* (absolute interval for A), then the following describes the above inference rule:

$$\begin{aligned}&occurs-during(A,I)\wedge before(B,A)\rightarrow\\&occurs-during(B,I')\wedge\\&(I'.end-time\leq I.start-time)\end{aligned}\tag{3}$$

Both the upper and domain ontology are represented with OWL Web Ontology Language. Currently the spatial model supports absolute and named locations i.e. GPS coordinate (i.e. lat-lng), bounding-box (i.e. a pair of GPS coordinates) with real values, and place name (i.e. a string e.g. "Disney Land").The following RDF N-triple describes the association of absolute locations to events using relationship *occurs-at*:(Perdurant occurs-at Place),(Place has-boundary BoundingBox), (BoundingBox s-contains Coordinates).
The relationship has-subevent is an irreflexive, asymmetric and transitive relation by definition and design between a pair of events. According to the entailment rule for relationship *has-subevent*, if *A,B* are both Perdurants and *has-subevent(A,B)*, then the spatiotemporal extent of B is covered by that of A's, i.e.:
(A has-subevent B) → $\exists I_A, Pl_A, I_B, Pl_B$ (A occurs-during I_A) ∧ (A occurs-at Pl_A) ∧ (B occurs-during I_B) ∧ (B occurs-at Pl_B) ∧ ($I_B \sqsubseteq_T I_A$) ∧ (Pl_A has-boundary Box_A) ∧ (Pl_B has-boundary Box_B) ∧ ($Box_A \sqsubseteq_S Box_B$)).
Relationship *has-Quality* associates an ambient quality e.g. weather,scene to an event. The RDF triple (A has-Quality outdoor) for *outdoor* event A. Participants are people attending an event, indicated by relationship *has-participants*. Relationship *has-object* associates the key visual objects that help in subevent recognition, e.g. for event *taking portrait*, face is an important visual object (designed as enumerated type). Each event class is associated to class *Process* through relationship has-processingUnit. This is used in augmentation. The following RDF N-triple describes a process:

```
(Perdurant has-processingUnit Process)
(Process has-name literal:String)
(Process has-source literal:String)
(Process has-inputPath literal:String)
```

Each process has a unique name e.g. "LandmarkFinder" (2^{nd} line), a source-path e.g. "www.apix.com" (3^{rd} line) , and input-path (4^{th} line). Input-path is the source for Input Context in Fig3 (e.g. calendar, user manual input, etc). In designing the Domain Ontology, each type of event class is described via CI_c such that CI_c is a subset of spatial, temporal, quality, structural (subevent), and conceptual (objects) information. For instance, event class "having lunch" has place category "restaurant" ($CI_c^{spatial}$) and all its instances occur during time interval [11am - 3pm]

($CI_c^{temporal}$). However, this class may not make sense to be described in terms of the ambient qualitysize(,); (see equation4).

$$\forall Perdurant \in O_d \exists CI_c \subseteq < CI_c^{spatial} \cup CI_c^{temporal} \cup CI_c^{quality} \cup CI_c^{structural} \cup CI_c^{conceptual} > \tag{4}$$

In Fig3 the ontology store maintains the upper and domain ontology(-ies).

1) Ontology Augmentation: Augmentation is an operation to create R-Ontology. Fig1 shows how upper ontology is used as a universal model. Next, our domain representation model can be multi-layered, i.e. the domain ontology gets conceptually extended at each layer. Fig4-Left shows a general domain ontology at the top layer of our domain representation framework, and Fig4-Right shows the next layer that is extended from the previous layer. Domain ontology generalizes an event-model for ALL of the actual events that fall inside that domain; therefore contextual information of a specific situation cannot be part of domain ontology. By the same token, domain ontology by itself cannot sufficiently analyse a personal photo stream. CI_v denotes the set of circumstances that surround an actual event, and will be only available in R-Ontology. An important aspect during augmentation is to identify the required data sources that provide CI_v to augment ontology instance, hence it is intuitive to understand that CI_v varies for different events. In selection of data sources we considered two factors: 1) CI_v should include the required contextual information for events specifically (e.g., time interval, location, participants), and 2) CI_v is better to be structured to avoid exhaustive operation. The following algorithm explain operation $\times_A$.

Algorithm 1 Ontology Augmentation

```
INPUT: inputContext, source, $O_r$ ,class, relationship
OUTPUT: $O_r$
$CI_v = wrapper(source, inputContext)$
for each context $\in$ $CI_v$ do
    $actualEvent = newclass()$
    $actualEvent.AugmentVia(relationship)$
    $Augment(actualEvent, context)$
    $Append(O_r, actualEvent)$
end for
return $O_r$
```

Fig5 shows an R-Ontology (instantiated from the model in Fig4-Right) that can be used for analysing a photo stream. Ontology augmentation is demonstrated as a building block of system architecture in Fig3.

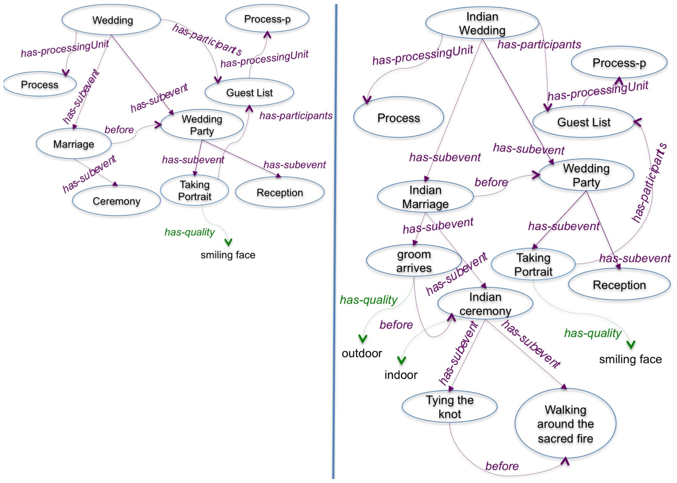

Figure 4. Left:Wedding Ontology; Right:Indian Wedding Ontology

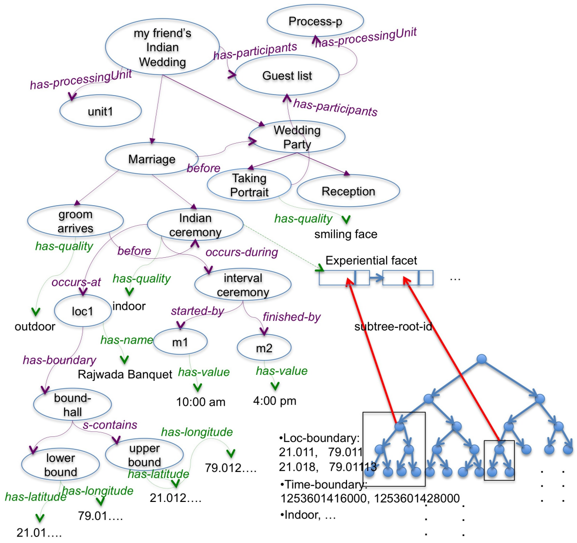

Figure 5. R-Ontology for Recognition.

2) Information Extraction: Extracting features and metadata over a large collection is necessary since it can be leveraged during filtering to organize the collection. Photo stream is imported to our framework from smart-phones and GPS-enabled camera devices. We extracted the EXIF (such as location, time, lat-lng, and imaging parameters like focal length and aperture) from our geotagged photoset using ExifTool[2]. This is followed by calling reverse-geocoding service SimpleGeo[3] to find the closest category to each extracted geo-tag (e.g., the closest venue/point of interest to <lat0,lng0> is a 'hotel', so the category is 'hotel'). It is important to automatically retrieve the venue name correspondent to a geo-coordinate when APIs such as Foursquare[4], and Gowalla[5] are already using reverse geocoding services. Another class of features is extracted from the content of media objects such as faces, etc. The *Feature Extraction* building block in Fig3 corresponds to

[2] http://owl.phy.queensu.ca/ phil/exiftool/
[3] http://simplegeo.com/
[4] http://foursquare.com/
[5] http://gowalla.com/

this operation.

3) Agglomerative Spatiotemporal Clustering: Information extraction is followed by an agglomerative /hierarchical clustering on time and space attributes of the media objects. The resulting cluster-tree is indicated as a blue tree at Fig3 and Fig5. Events by definition happen in time and space; therefore time and space are important attributes of events in R-Ontology. On the other hand, photos are associated with time and space (if GPS is used). This motivate us to apply spatiotemporal clustering. The attributes of each leaf cluster is: cluster-id, Timestamp(ms) t, and coordinate (x, y). The attributes of each nonleaf cluster is: cluster-id, children-ids, temporal interval (timestamp(ms) t1, timestamp(ms) t2), spatial boundary(minX,minY,maxX,maxY).

4) Filtering: Filtering has two components: a) A function that filters redundant clusters based on the absolute/relative spatiotemporal extent of an event in R-Ontology by pruning the branches of the cluster-tree that lead to irrelevant clusters by using the following conditional expression:
$Inside_{ST}(cluster, event) \rightarrow Populate(event, cluster)$;
Many-to-many cardinality exists between events in R-Ontology and clusters in cluster tree, this is followed by b) A function that filters redundant images in each assigned cluster for an event based on the described quality of the event in R-Ontology using the image content and context features (see equation 5) e.g., for an outdoor event, imaging parameters taken from EXIF such as focal-length, aperture and time are employed to recognize whether the photos in the assigned cluster(s) represent indoor or outdoor scene.We used an existing API in our lab [12]. Such descriptions can play a crucial role once extensive types of visual content and context features are taken into account e.g., image histogram to find distance measure based on the tonal distribution. We will bring in such features extensively in our future work.

$$\begin{aligned}&\forall m \in subevent.Media\ \forall f \in subevent.getFeatures() \wedge \\ &!Consistent(Extract(m, f.type).val, f.val) \longrightarrow \\ &Remove(subevent.Media, m)\end{aligned} \tag{5}$$

Our filtering approach eliminates the need for optimization. The filtering is deeply guided by R-Ontology (see Fig5). During this operation, mereological/subevent relations are used for navigating the R-Ontology. Filtering based on absolute properties is centralized around only one event in R-Ontology, whereas filtering based on relative properties is distributed over multiple events: we give higher priority to analyse the events with absolute spatiotemporal extent to compute the relative properties of other events (that have necessary dependence towards counterpart absolute properties); our algorithm relatively translates relative properties to absolute ones (see equation 3). In order to track distinctive characteristics of each event in R-Ontology, we implemented a graph navigator using OWL API [1]. R-Ontology is written in OWL.

IV. EXPERIMENTS

We applied our proposed framework on two scenarios: 1) Trip, and 2) Wedding. Our motivation is their planned nature so that their schedule becomes available to our framework. These are types for events. We used our own personal photo archive since we could not find enough number of photos in photo sharing websites like Flickr, Picasa and Photobucket[6] regarding a particular trip or wedding event. We created a Trip, and wedding ontology to describe the abstract skeleton of general trip, and wedding events respectively. We used Single-Linkage clustering that creates binary tree over the underlying dataset, however any other agglomerative spatiotemporal clustering that can form reasonable clusters can replace this. Time and location attributes of each photo (extracted by operation extraction) are used during this operation. The right/left child of each non-leaf cluster node may either be a non-leaf or leaf level node. The leaf level nodes represent atomic clusters that do not carry any children. Each node has one and only one parent.

We had to manually create a database of points of interest (POI) to find the closest landmark category to a geo-tag since our photoset was captured inside China and India that are not covered by SimpleGeo. CANADA and US are the only countries covered by SimpleGeo. We used Wordnet API to match these categories with those of event locations. In the following subsections we will show the visual results for each case.

A. Trip Scenario

During clustering, the contents of clusters were reasonably meaningful i.e. pictures of the same sub-event appeared in the same cluster. We compared our clustering to Kmeans. Bottom of Fig6 shows two relevant photos (in terms of the event they cover) appear in two separate clusters in Kmeans clustering while they should have been located in the same cluster. It also shows the result generated by single linkage clustering (top of the figure); the photos appeared in the same cluster, which is the satisfying result. Due to the nature of our example photoset, we found the single-linkage algorithm more effective in the initial stage of our framework. The reason for this observation is relevant to the cluster model of the two methods. This may be worthwhile to be investigated more; however that is beyond the scope of this paper. Our Trip dataset was collected during a Trip to Beijing. Given the type (trip), and location (Beijing) of this planned event, operation augmentation automatically creates R-Ontology, using web services Trip Advisor, and Yahoo Travel that find the top ten Landmarks for famous cities in the world. Each landmark from these services has attribute ”category” e.g., historical place, mall. The category helps in associating

[6] http://photobucket.com/

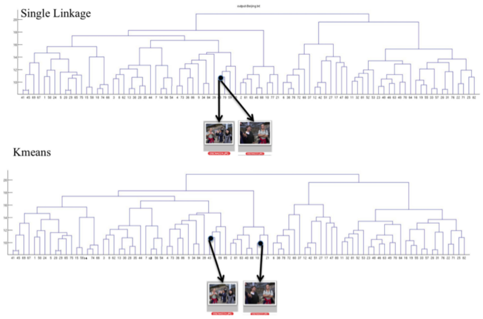

Figure 6. Linkage Clustering vs. Kmeans Clustering.

the landmark to the appropriate events in R-Ontology. Given the landmark name (e.g., forbidden city), operation augmentation uses geocoding service Yahoo PlaceFinder to find the corresponding geographic coordinates and boundingbox to make our framework location-aware. In respect to time, some trip planner web sites provide the working hours of the underlying named place. This could be automatically crawled, however it was done manually in our work since we encountered the absence of such information source in respect to ”Beijing, China”. Further extension to the ontology instance for impromptu events is handled manually via Protege ontology editor[7].

Our wedding dataset belongs to an Indian wedding. We manually created the wedding ontology. The time and location of the schedule of the wedding is extracted from an invitation eCard. This data source is chosen because it contains the information values for the parameters that well describe this event. It is reasonable to assume that the schedule of the wedding in the invitation eCard has a structured format with all the information such as the location/address and temporal schedule of the wedding, start and end of ceremony and reception, hosts. Also guest list i.e. participants is easily accessible. Services like Evite/facebook-event API are free and ready-to-use data sources that maintain such data. Fig7(a) shows the matched photos for event ”groom arrives”. We used face.com API to detect faces during operation extraction. The detected faces are provided with interesting features such as smiling, male, female, right-eye, left-eye, and mouth. We investigated that in wedding family pictures, people present themselves with smiling gesture in most cases. Also family photos contain guests’ faces. We considered this as one of characteristics for event *Taking Portrait* in our wedding model. Fig7(b) shows the final result for this event.

V. Evaluation

Because of the novelty of this work, we could not find a standard way, but a qualitative user study to evaluate our framework. The test data consisted of two sets, one with 86 (for Trip) and the other with 87 (for Wedding) photos. We asked the owner of these photo collections to manually organize those photos in terms of our modeled sub-events since the person who took the photos, actually participated in the corresponding events, and that the way these photos are organized by him can serve as a ground truth for our results to some extent. We compared the results obtained from manual annotation with the outcome of our approach. TableI reflects the result of this evaluation. COV indicate the coverage of photos for each event, and COR indicates the correctness of the assigned photos for that event. These values show that for each event, the relevant visual data is covered, however due to the lack of features in the second step of filtering, COR does not look as good as COV. We are currently in the process of collecting more data to evaluate our work in a larger extent. We plan to use the ground truth and evaluation introduced by MediaEval[8] that is a benchmark for multimedia evaluation.

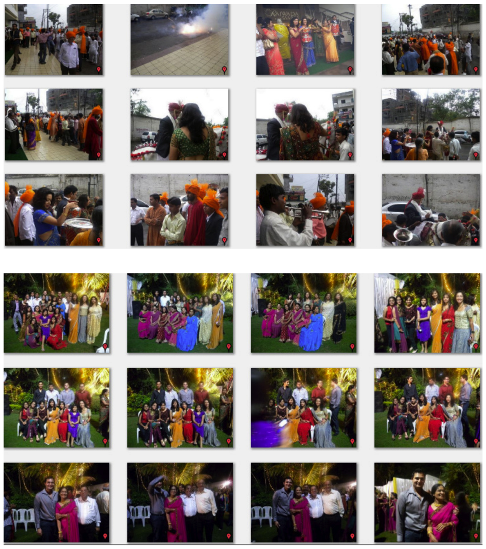

Figure 7. (a) Top 3 row: groom arrives; (b) Bottom 3 rows:Taking Portrait.

Table I
Evaluation

Album	Events	COV	COR
Trip to Beijing	having dinner	1	0.83
	ordering food	1	1
	serving food	1	0.75
	visit Forbidden city	1	0.9
	ACM conference	1	1
Neha and Avinash’s Wedding	groom arrives	1	1
	religious ceremony	1	0.75
	taking portrait	0.95	1

[7] http://protege.stanford.edu/

[8] http://www.multimediaeval.org/mediaeval2011/SED2011/

VI. Conclusion

Our approach is used and successfully tested to analyse the personal photos of high-level domain-specific sub-events. The domain ontology may be constructed automatically using machine learning and statistical methods in the future, however so far no machine learning technique has been qualified enough to create such structure from the data by itself. On the other hand, we can not ignore the fact that many (if not all) of our daily activities are executed based upon a pattern and agenda kept and maintained through generations up to now such as the type of agenda one may follow for a wedding event. Hence it is a great privilege to leverage such existing documented/nondocumented knowledge as heterogeneous sources of information for media organization. Upon the availability, we used righteous available sources of information that serve as contextual data during augmentation. Such sources are identified based on how well they describe the underlying situation. In this work we encountered some limitations according to the lack of discovering such sources, so some of contextual data had to be created manually, however this can be tackled in the future as more services and databases become publicly available. We also aim to extend on employing content and context features of visual data towards event recognition. This can happen with the presence of large number of personal sensors in smart-phones and can improve the evaluation of our framework on the correctness (COR) metric. We consider these extensions in our future work.

Acknowledgment

The authors would like to thank Hamed Pirsiavash for helping in clustering our photo collection.

References

[1] *The OWL API: A Java API for Working with OWL 2 Ontologies*, OWLED , 6th OWL Experienced and Directions Workshop, Chantilly, Virginia,2009.

[2] U. Westermann, R. Jain, *Toward a Common Event Model for Multimedia Applications*, IEEE MultiMedia , 2007.

[3] M. Vargas-Vera, and D. Celjuska,Ontology-driven Event Recognition on Stories, KMi Tech. Report KMI-TR-135, 2003.

[4] B. Chandrasekaran,J.R. Josephson, V.R. Benjamins, *What Are Ontologies, and Why Do We Need Them?*, Ohio State University, University of Amsterdam.

[5] C. Lagoze, J. Hunter, *The ABC Ontology and Model*, Dublin Core Conference, pp. 160-176, 2001.

[6] R. Arndt, R. Troncy, S. Staab, L. Hardman,M. Vacura, *COMM: Designing a Well-Founded Multimedia Ontology for the Web*, Proc. Int. Semantic Web Conf./Asian Semantic Web Conf., pp. 30-43, 2007.

[7] A.R.J. Francois, R. Nevatia, J. Hobbs, and R.C. Bolles, *VERL: An ontology framework for representing and annotating video events*, IEEE MultiMedia, 2005.

[8] E.T. Mueller, *Handbook of Knowledge Representation*, chapter Event Calculus, Elsevier, 2008.

[9] A. Scherp, T. Franz, C. Saathoff, Staab, *F-A Model of Events based on the Foundational Ontology DOLCE+ Ultra Light*, In: 5th International Conference on Knowledge Capture, Redondo Beach, California, USA, 2009.

[10] M.M. Kokarand, J., Wang, *Using ontologies for recognition: An example*, In Proceedings of the Sixth International Conference on Information Fusion, 2002.

[11] S.H. Wu, T.H. Tsai,W.L. Hsu, *Domain event extraction and representation with domain ontology*, Proceedings of the IJCAI- 03 Workshop on Information Integration on the Web, Acapulco, Mexico,2003.

[12] P. Sinha,R. Jain, *Semantics in Digital Photos: a Contenxtual Analysis*, Int. J. Semantic Computing 2(3): 311-325 (2008).

[13] T. Gruber, L. Liu, M. Tamer zsu , *In the Encyclopedia of Database Systems*, Springer-Verlag, 2009.

[14] T.O. Binford, *Image understanding: intelligent systems*, In Image Understanding Workshop Proceedings, volume 1, pages 18-31, Los Altos, California, 1987.

[15] A. Smeulders, et. al., *Content-Based Image Retrieval at the End of the Early Years*,IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.

[16] G. Oates, *Holy moly!*,The Flickr Blog, http://blog.flickr.net/en/2007/11/13/holy- moly/, 2007.

[17] D. Beaver, *Facebook photos infrastructure*, The Facebook Blog, http: //blog.facebook.com/blog.php?post= 2406207130, 2007.

[18] U. Akdemir, P.Turaga, R. Chellappa, *An ontology based approach for activity recognition from video*, Proceeding of the 16th ACM international conference on Multimedia, NY, 2008.

[19] B. Zheng,L. Huang, X. Lu, *Ontology for the Object Recognition in Images*, International Conference on Multimedia Information Networking and Security, 2009.

[20] Y. Liu, D. Zhang, G. Lu, W. Ma, *A Survey of content-based image retrieval with high-level semantics*, Pattern Recognition, 2006.

[21] J.R. Smith, S.F. Chang, *Large-scale concept ontology for multi- media*, IEEE Multimedia, 2006.

[22] M. Naaman, S. Harada, Q. Wangy, H. Garcia-Molina, A. Paepcke, *Context data in georeferenced digital photo collections*, in: 12th International Conference on Multimedia, New York, New York,2004.

[23] A. Gupta, R. Jain, *Event Management-Managing Event Information* , Morgan Claypool,2011.