Using Concept Recognition to Annotate a Video Collection

Anupama Mallik and Santanu Chaudhury

Electrical Engineering Department, IIT Delhi
ansimal@gmail.com, schaudhury@gmail.com

Abstract. In this paper, we propose a scheme based on an ontological framework, to recognize concepts in multimedia data, in order to provide effective content-based access to a closed, domain-specific multimedia collection. The ontology for the domain is constructed from high-level knowledge of the domain lying with the domain experts, and further fine-tuned and refined by learning from multimedia data annotated by them. MOWL, a multimedia extension to OWL, is used to encode the concept to media-feature associations in the ontology as well as the uncertainties linked with observation of the perceptual multimedia data. Media feature classifiers help recognize low-level concepts in the videos, but the novelty of our work lies in discovery of high-level concepts in video content using the power of ontological relations between the concepts. This framework is used to provide rich, conceptual annotations to the video database, which can further be used to create hyperlinks in the video collection, to provide an effective video browsing interface to the user.

1 Introduction

Meaningful access to the ever-increasing multimedia data in the pubic domain faces the crunch of available conceptual metadata and annotation text. This textual metadata is helpful in bridging the semantic gap between high-level semantic concepts and the low-level content-based media features. Video annotation is essential for successful content-based video search and retrieval, but done manually it is tedious and prone to inaccuracy. In [1], Zha et al propose to refine video annotation by leveraging the pairwise concurrent relation among video concepts. In [2], the authors have systematically studied the problem of event recognition in unconstrained news video sequences, by adopting the discriminative kernel-based method. Concept Recognition using an ontology for the purpose of enhancing content-based multimedia access as attempted in our work, is a relatively new approach.

In our work, we propose a scheme based on an ontological framework, to recognize concepts in multimedia data, in order to generate rich, conceptual annotations for the data. The annotations generated by this scheme provide associations between the concepts in the domain and the content in the multimedia files, forming a basis for effective content-based access to the multimedia data in a closed, domain-specific collection. The highly specialized knowledge that experts of a scholarly domain have, is encoded into an ontological representation of the domain, and is refined by learning from observables in the multimedia examples of the domain. This approach to concept learning has been detailed in our earlier work [3].

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 507–512, 2009.

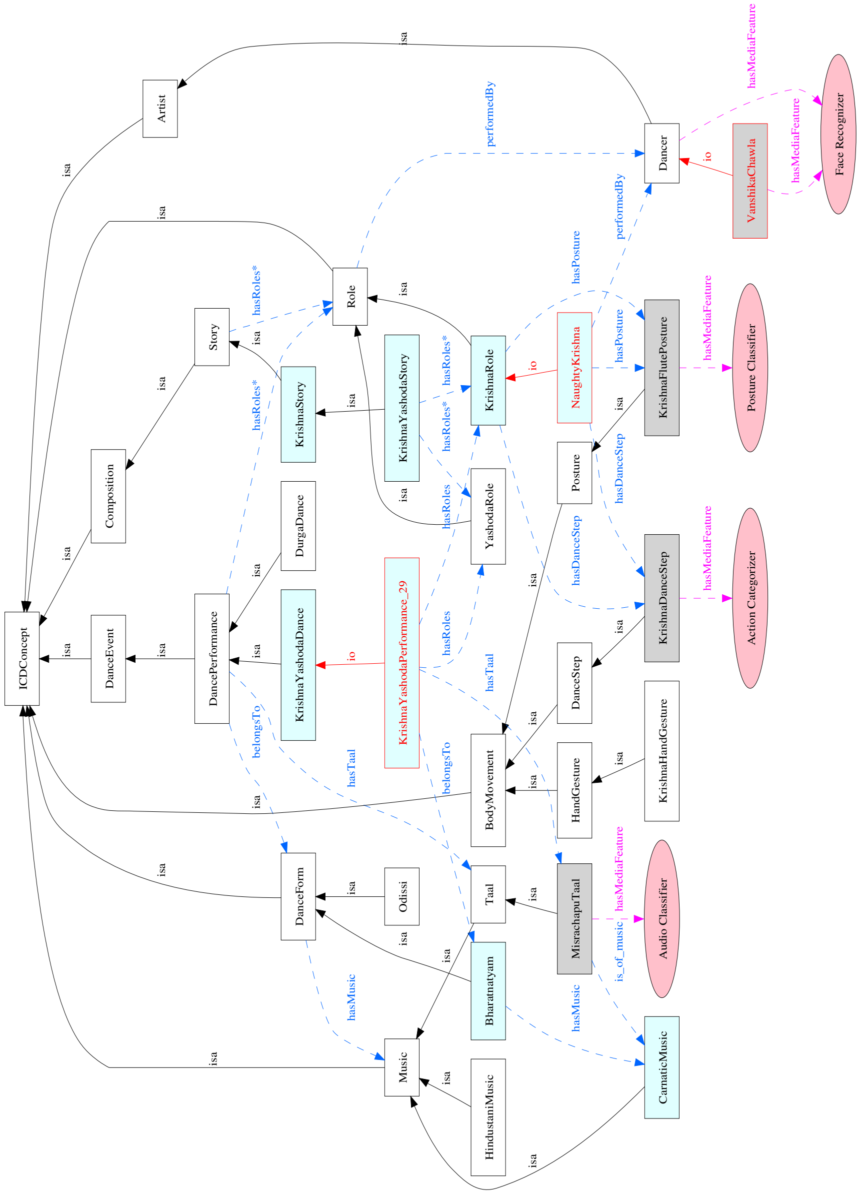

Fig. 1. MOWL Ontology of Indian Classical Dance

The key contribution of our current work is the discovery of high-level concepts in video content using the power of the MOWL encoded ontology to propagate media properties across concepts. We have shown the success of our technique by applying our work to a cultural heritage domain of Indian classical dance. The conceptual annotations generated can be used to create hyperlinks in the video collection, to provide an effective video browsing interface to the user.

2 Concept Recognition Using MOWL Ontology

We have extended the existing ontology representation based on OWL to include a perceptual description of the concepts and formalized it as Multimedia Web Ontology Language(MOWL). MOWL supports probabilistic reasoning with Bayesian Networks in contrast to crisp Description logic based reasoning with traditional ontology languages [4]. In this paper, we have used a robust evidential reasoning framework around MOWL, where a concept can be recognized in a multimedia entity on the evidential strength of physical observations of some its expected media properties.

Figure 1 shows a snippet of the Indian Classical Dance (ICD) ontology represented graphically. Root Node represents an ICD concept, some of which are shown as ’Music’, ’DanceForm’, ’Artist’ and ’Composition’. This snippet focuses on an important mythological figure - an Indian God named Krishna. Stories about Krishna abound in folklore, and all classical dances of India have performances dedicated to events in his life. One of the events depicted here is enactment of a scene between Krishna and his mother, Yashoda through a performance in Bharatnatyam dance form. Linkages and dependencies between a ’Story’, a ’Role’, a ’DancePerformance’, a ’DanceForm’, a ’Dancer’, various ’Body Movements’ are encoded in the MOWL ontology. The leaf nodes in elliptical shape denote ’Media Feature’ nodes and represent various media classifiers like posture recognizer, face recognizer. This graphical representation of the ICD ontology represents a Bayesian Network. The Conditional Probability values are not shown here in order to preserve the visual clarity of the diagram. Evidence is gathered at the leaf nodes, as different media features are recognized or classified by the media classifiers. If evidence is above a threshold, the media feature node is instantiated. These instantiations result in belief propagation in the Bayesian Network, and posterior probability at the associated concept nodes is computed. The algorithm for recognizing the concepts in this BN is as following :

```
Inputs : 1) Video segment $\mathcal{V}$ for which concepts are to be recognized
         2) Bayesian Network $\Omega$ of the relevant MOWL ontolgy segment
Output : Set $\mathcal{C}$ of Recognized Concepts
Algorithm :
1. For each leaf-node Concept $\mathcal{LC}_i$ in $\Omega$,
   i. Run the appropriate Media Feature classifier.
   ii. If Classification evidence for the concept $>$ threshold,
      a. Add $\mathcal{LC}_i$ to the set $\mathcal{I}$ of instantiated nodes
      b. Add $\mathcal{LC}_i$ to the result set $\mathcal{C}$ of recognized nodes.
2. Carry out Belief Propagation in the BN $\Omega$.
3. For each node $\mathcal{IC}_i$ in $\mathcal{I}$
   i. Compute the set $\mathcal{RC}$ of Related concepts at next higher level.
```

```
  ii. For each node $\mathcal{RC}_i$ in $\mathcal{RC}$
      a. Compute the posterior probability P($\mathcal{RC}_i$) at $\mathcal{RC}_i$
      b. If P($\mathcal{RC}_i$) $>$ threshold,
          •Add $\mathcal{RC}_i$ to the set $\mathcal{I}$ of instantiated nodes
          •Add $\mathcal{RC}_i$ to the to the result set $\mathcal{C}$ of recognized nodes.
4. Iterate steps 2 and 3 till Root node is reached.
```

3 Annotation Generation

The input to our concept-recognition scheme is an initial multimedia ontology of the domain constructed with the help of domain knowledge provided by a group of domain experts, and fine-tuned by learning from the training set of annotated videos [3]. A semi-automated annotation generation module provides an interface where domain concepts present in the video content are recognized automatically by the system, and presented to the annotator to verify and confirm their existence. The module consists of 5 functional components :

- **Object/Feature Extractor:** This module extracts the features/objects from the multimedia data. The extracted features are given to the XML generator, to store them in XML format.
- **MOWL Parser:** This module is responsible for generating the Bayesian network from the given MOWL ontology.
- **Concept Recognizer:** The task here is to recognize the high-level semantic concepts in multimedia data with the help of low-level media-based features. This module gets the feature values either by invoking the feature extractor or from the feature database. The concept recognizer either highlights or prompts the concept to the annotator, resulting in a kind of supervised annotation. It can also directly convey the concept/s recognized to the XML generator.
- **Classifiers:** Media Feature classifiers are trained with feature vectors extracted from the multimedia data. These are detailed in section 4.
- **XML based Annotation generator:** This module is responsible for generating the XML. The inputs to this module are the manual annotations, conceptual annotations and features of the multimedia data (in MPEG-7 format) and output is the XML file as per MPEG-7 standard, containing the video annotation as well as media based feature-vectors.

4 Experimental Results

We tested our ontology based annotation scheme on a captive collection of videos which belong to the scholarly domain of Indian Classical Dance(ICD). We compiled a heritage collection by gathering dance videos from different sources. We started work with a data set of approximately 200 videos of duration 10 to 15 minutes. These consist of dance performances of different Indian classical dance forms - Bharatnatyam, Odissi, Kuchipudi and Kathak; plus music performances of Indian classical music forms - Carnatic Music and Hindustani Music.

With reference to the snippet in Fig. 1, concept-recognition occurs with belief propagation in BN. The concept nodes highlighted in gray color are the low-level concepts which are recognized due to presence of the media features in data. These are ’Misrchapu Taal’(a taal/beat in Carnatic music), ’KrishnaDanceStep’, ’KrishnaPosture’ and ’VanshikaChawla’ (a dancer). Due to further belief propagation in the BN, higher level concept nodes (in cyan color) are recognized to be present in the video. Conceptual annotations are generated and attached to the video through the Annotation Generation module detailed in section 3. Videos are hyperlinked if they are annotated with common concepts, or ontologically related concepts. This hyperlinking formed the basis of an ontology-based Browsing application for the video database(not detailed here due to space constraint). Some of the Media feature classifiers used by our concept-recognition scheme are detailed below :

4.1 Human Action Categorization Using pLSA

Our framework for detecting human action categories includes the following steps :
- Spatio-temporal interest points are extracted for frames of a video.
- The extracted spatio-temporal interest points are used in the bag of words approach to summarize the videos in the form of spatio-temporal words.
- The process automatically learns the probability distributions of the spatio-temporal words and intermediate topics for detecting action categories using pLSA technique [5].
- The topic-to-video probability distributions we get from pLSA training and testing, are fed to an SVM classification scheme for categorisation of actions.

For performing pLSA categorization, some of the recognizable dance actions selected from Bharatnatyam dance were - **Sitting and getting up**, **Side-stepping**, **Taking a circle**, **Krishna Step**, **Teermanam Step**. Approximately 30 video shots of each action were submitted to the pLSA for training. We performed 6-fold cross-validation tests on 77 videos to test the classification of the various dance actions by pLSA technique. The accuracy of classification was found to be approx. 76.8% on an average.

4.2 Dance Posture Recognition Using SIFT

We have used the SIFT approach [6] to recognize dance postures in still images taken from dance videos. Steps of our computation are :
- Collect labelled examples of Dance Posture images from different dance videos.
- Extract SIFT descriptors for all the images, and quantize the SIFT descriptors by K-means clustering algorithm to obtain a discrete set of local N_s SIFT words.
- A posture image P_i is represented by an Indicator vector $iv(P_i)$, which is a histogram of its constituent SIFT words,

$$iv(P_i) = \{n(P_i, s_1), ..., n(P_i, s_j), ..., n(P_i, s_(N_s))\} \tag{1}$$

where $n(P_i, s_j)$ is the number of local descriptors in image P_i, quantized into SIFT word s_j.
- Train an SVM classifier with the indicator vectors to classify the postures.

We extracted about 288 images of various Dance postures from our set of ICD videos. These were classified into 7 broadly similar dance postures, as shown in table 1. An average of 232 detected points (depending on image content) and K-means clustering with 50 cluster centers yielded indicator vectors for all the 288 images. A 10-fold

Table 1. SVM Classification Results for Dance Postures

$Classes$	$TPRate$	$FPRate$	$Precision$	$Recall$	$F-Measure$	$ROCArea$
RightAnjali	0.867	0	1	0.867	0.929	0.933
LeftAnjali	0.966	0	1	0.966	0.982	0.983
FrontPranam	0.931	0.004	0.964	0.931	0.947	0.964
ArmsUp	0.731	0.068	0.704	0.731	0.717	0.831
Squat	1	0	1	1	1	1
KrishnaPose	0.936	0.057	0.889	0.936	0.912	0.94
HandsOnWaist	0.552	0.039	0.615	0.552	0.582	0.757

cross-validation using SVM classifer on Weka machine learning framework, yielded an accuracy of 87.8%. The detailed results are shown in table 1.

5 Conclusion

In this paper, we have outlined a novel approach to recognize concepts in a closed collection of videos belonging to a scholarly domain, and use it to generate conceptual annotations for videos at different levels of granularity. The multimedia ontology is capable of incorporating uncertainties attached to media observables, and thus offers a probabilistic framework, which can be enhanced using ontology learning from annotated data. Thus the whole system is self-enhancing where ontology is refined from annotated data, and data annotation is improved based on fresh, refined knowledge from the ontology. This ontological framework can offer a robust ground for several multimedia search, retrieval and browsing applications.

References

1. Zha, Z.J., Mei, T., Hua, X.S., Qi, G.J., Wang, Z.: Refining video annotation by exploiting pairwise concurrent relation. In: MULTIMEDIA 2007: Proceedings of the 15th international conference on Multimedia, pp. 345–348. ACM, New York (2007)
2. Xu, D., Chang, S.F.: Video event recognition using kernel methods with multilevel temporal alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1985–1997 (2008)
3. Mallik, A., Pasumarthi, P., Chaudhury, S.: Multimedia ontology learning for automatic annotation and video browsing. In: MIR 2008: Proceeding of the 1st ACM international conference on Multimedia information retrieval, pp. 387–394. ACM, New York (2008)
4. Ghosh, H., Chaudhury, S., Kashyap, K., Maiti, B.: Ontology specification and integration for multimedia applications. Springer, Heidelberg (2006)
5. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, UAI 1999, pp. 289–296 (1999)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 20, 91–110 (2003)