ISSN:2231-0711
www.ijcset.net

A Semantics-Based Approach for Speech Annotation of Images

[1] HarishBabu. Kalidasu, [2]B. Prasanna Kumar, [3] K. PurnaPrakash

Department of Computer Science & Engineering
[1]Priyadarshini Institute of Technology & Science-Tenali, AndhraPradesh
[2]Mandava Institute of Engineering & Technology-Jaggayyapet, Andhra Pradesh
[3]Sir C.R.Reddy College of Engineering, Eluru, Andhra Pradesh

Abstract— The Associating textual annotations/tags with multimedia content is among the most effective approaches to organize and to support search over digital images and multimedia databases. Despite advances in multimedia analysis, effective tagging remains largely a manual process wherein users add descriptive tags by hand, usually when uploading or browsing the collection, much after the pictures have been taken. This approach, however, is not convenient in all situations or for many applications, e.g., when users would like to publish and share pictures with others in real time. An alternate approach is to instead utilize a speech interface using which users may specify image tags that can be transcribed into textual annotations by employing automated speech recognizers. Such a speech-based approach has all the benefits of human tagging without the cumbersomeness and impracticality typically associated with human tagging in real time. The key challenge in such an approach is the potential low recognition quality of the state- of-the-art recognizers, especially, in noisy environments. In this paper, we explore how semantic knowledge in the form of co-occurrence between image tags can be exploited to boost the quality of speech recognition. We postulate the problem of speech annotation as that of disambiguating among multiple alternatives offered by the recognizer. An empirical evaluation has been conducted over both real speech recognizer's output as well as synthetic data sets. The results demonstrate significant advantages of the proposed approach compared to the recognizer's output under varying conditions.

Keywords—Using speech for tagging and annotation, using semantics to improve ASR, maximum entropy approach, correlation- based approach, branch and bound algorithm.

1 INTRODUCTION

Increasing popularity of digital cameras and other multi- media capture devices has resulted in the explosion of the amount of digital multimedia content. Annotating such content with informative tags is important to support effective browsing and search. Several methods could be used for such annotation, as explained below

For image repositories, the first way to annotate pictures is to build a system that relies entirely on visual properties of images. The state-of-the-art image annotation systems of that kind work well in detecting generic object classes: car, horse, Motorcycle, airplane, etc. However, there is limitations associated with considering only image content for annotation. Specifically, certain classes of annotations are more difficult to capture. These include location (Paris, California, San Francisco,etc.), event (birthday, wedding, graduationceremony, etc.), people (John, Jane, brother, etc.),and abstract qualities referring to objects in the image (beautiful, funny, sweet, etc.).

The second and more conventional method of tagging pictures is to rely completely on human input. This approach has several limitations too. For instance, many cameras do not have an interface to enter keywords. Even .if they do, such a tagging process might be cumbersome and inconvenient to do right after pictures are taken. Alternatively, a user could tag images at a later time while either uploading them to a repository or browsing the images. Delay in tagging may result in a loss of context in which the picture was taken (e.g., user may not remember the names of the people/structures in the image).Furthermore, some applications dictate that tags be associated with images right away.

The third possibility for annotating images uses speech as a modality to annotate images and/or other multimedia content. Most cameras have built-in microphone and provide mechanisms to associate images with speech input. In principle, some of the challenges associated with both fully automatic annotation as well as manual tagging can be alleviated if the user uses speech as a medium of annotation. In an ideal setting, the user would take a picture and speak the desired tags into the device microphone.

A speech recognizer would transcribe the audio signal into text. The speech to text transcription could either happen on the device itself or be done on a remote machine. This text can be used in assigning tags to the image. The proposed solution is useful in general scenarios, where users might want to use a convenient speech interface for assigning descriptive textual tags to their images. Such systems also can play a critical role in applications that require real-time triaging of images to a remote site for further analysis, such as reconnaissance and crisis response applications. All three aforementioned tagging approaches are not competing and, in practice, can complement each other.

Motivating Application Domain

While STI technology is of value in a variety of application domains, our work is motivated by the emergency response domain. In particular, we have explored STI in the context of the Situational Awareness for Firefighters (SAFIRE) project wherein our goal is to enhance the safety of the public and Fire fighters from fire and related hazards. We are developing situational awareness technologies that provide firefighters with synchronized real-time information. These tools are expected to enhance safety and facilitate decision making for firefighters and other first responders. The ultimate goal is to develop an information-and-control-panel prototype called the Fire Incident Command Board. This device will combine new and existing hardware and software components that can be customized to meet the needs of field incident commanders. FICB tools will allocate resources, monitor status and locale of Personnel, and record and interpret site information.

The FICBs will integrate and synchronize sensors and other information flows from the site and provide customized views to individual users while seamlessly interacting with each other. While STI technology is of value in a variety of application domains, our work is motivated by the emergency response domain. In particular, we have explored STI in the context of the Situational Awareness for Firefighters (SAFIRE) project wherein our goal is to enhance the safety of the public and firefighters from fire and related hazards. We are developing situational awareness technologies that provide firefighters with synchronized real-time information. These tools are expected to enhance safety and facilitate decision making for firefighters and other first responders. The ultimate goal is to develop an information-and-control-panel prototype called the Fire Incident Command Board. This device will combine new and existing hardware and software components that canbe customized to meet the needs of field incident commanders. FICB tools will allocate resources, monitor status and locale of personnel, and record and interpret site information.

The FICBs will integrate and synchronize sensors and other information flows from the site and provide customized views to individual users while seamlessly interacting with
each other.

1.2 Our Approach

Our work addresses the poor quality of annotations By incorporating outside semantic knowledge to improve interpretation of the recognizer's output, as opposed to blindly believing what the recognizer suggests. Most speech recognizers provide alternate hypotheses for each speech utterance of a word, known as the N-best list for the utterance.

2 RELATED WORK

In this section, we start by discussing work related to other speech-based annotation systems. We then cover some of closely related solutions that do not deal directly with speech. Finally, we highlight the contribution of this paper compared to its preliminary version.

2.1 Speech Annotation Systems

Several speech annotation systems have been proposed that utilize speech for annotation and retrieval of different kinds of media [3], [19], [20], [21], [31], [32]. In [20], [21], the authors propose to investigate a simple and natural extension of the way people record video. It allows people to speak out annotations during recording. Spoken annotations are then transcribed to text by a speech recognizer.

however, requires certain syntax of annotations, and specifically that each content-descriptive free speech annotation is preceded by a keyword specifying the kind of annotation and its associated temporal validity. Our approach does not require any particular syntax of annotations. The system does not utilize any outside knowledge to improve recognition accuracy.

The authors propose a multimedia system for semi automated image annotation. It combines advances in speech recognition, natural language processing, and image understanding. It uses speech to describe objects or regions in images. However, to resolve the limitation of speech recognizer, it requires several additional constraints and tools:

a) Constraining the vocabulary and syntax of the utterances to ensure robust speech recognition. The active vocabulary is limited to 2,000 words.
b) Avoiding starting utterances with such words as "this" or "the." These words might promote ambiguities.
c) Providing an editing tool to permit correction of speech transcription errors.

2.2 Nonspeech Image Annotation

Due to practical significance of the problem, many different types of image tagging techniques have been developed. In the previous section, we have already reviewed techniques that utilize speech for annotation. In this section, we will overview those that do not employ speech for that purpose. Observe that while the goal of our problem is to derive image tags from the corresponding speech utterances, the goal of the techniques discussed in this section is naturally different, since they do not use speech. Typically, their goal is to derive tags automatically from image features or to assign them manually by the user.

Because of the difference of the goals, the techniques mentioned in this section are not competing to our approach. They are rather complementary as they can be leveraged further to better interpret utterances of spoken keywords, but developing techniques that can achieve this is beyond the scope of this paper. Many content-based annotation algorithms have been proposed to annotate images automatically, based on the content of images and without using speech. Such algorithms usually generate several candidate annotations by learning the correlation between keywords and visual content of images. Given a set of images annotated with a set of keywords that describe the image content, a statistical model is trained to determine the correlation between keywords and visual content. This correlation can be used to annotate images that do not have annotations. Candidate annotations are then refined using semantics.

The correlation between keywords and image features can be also captured by learning a statistical model, including Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) Annotations for unlabeled images are generated by employing these models. The authors encode image features as visual keywords. Images are modeled by concatenated visual keywords and if any, annotations.Semantic Analysis (LSA) is applied to compute semantic similarity between an unannotated image and the annotated image corpus. The annotation is then propagated from the ranked documents.In addition, PLSA is applied to compute distribution of the terms of the vocabulary given an unannotated image. Social tagging is a manualimage tagging approach where a community of users participate in tagging of images. Different users can tag the same image and the end tags for an image are decided according to some policy. For instance, when a certain number of users submit the same tag for an image, the tag is assigned to the image. Diaz et al. in investigates ways to improve tag interoperability across various existing tagging systems by providing a global view of the tagging sites. By utilizing a query language, it is possible to assign new tags, change existing ones, and perform other operations. The system uses RDF graph as its data model and assumes that existing tagging systems will eventually become RDF graph providers.

2.3 Our Past Work

The differences of this paper compared with its initial version include:

1. Related work is now covered;
2. More in-depth coverage of the problem definition, including the pseudo code of the mentioned algorithm;
3. More in-depth coverage of the Max Entropy solution;
4. More in-depth coverage of correlation, Including new material related to indirect correlation and correlation and membership scores;
5. The Branch and Bound algorithm that makes the approach scale to large data sets, thus making it feasible in practice;
6. A method for combining the results of the global and local models, that leads to higher quality of annotations;
7. Five new experiments that study various aspects of the proposed solution. Some of our past entityresolution work is also related, but not directly applicable and uses different methodologies [4], [5], [6], [12], [13], [14], [15], [16], [17], [25], [26].

3 NOTATION AND PROBLEM DEFINITION

We consider a setting wherein the user intends to annotate an image with a sequence G ={g_1; g_2; . . . ; $g_{K\}}$ of K ground truth tags. Each tag g_i can be either a single word or a short phrase of multiple words, such as Niagara Falls, Golden Gate Bridge, and so on. Since a tag is typically a single word, we will use “tag” and “word” interchangeably.

3.1 Sequences

Let us define a sequence as a K-dimensional vector W = (w1; w2; . . . ; wK), where wi can be of three types:

a). wi E Li, that is, wi is one of the N words from list Li;

b). wi = null, which encodes the fact that the algorithm believes list Li does not contain gi;

c). wi = ‘‘_’’, that is, the algorithm has not yet decided the value of the ith tag. The cardinality |W| of sequence W is defined as the number of The elements of the first category that the Sequence has: |W| =|{wi E W : wi E Li}|. Sequence W is an answer sequence or a complete sequence; if none of its elements wi is equal to “_”. In other words, an answer sequence cannot contain undecided tags, only words from the N-best lists or null values.

3.2 Notational Example

As an example, suppose that the user takes a picture of her friends Jane in a garden full of roses, and provides the utterances of K = 5 words: G = (g1 = Jane; g2 = rose; g3 = garden; g4 =flower; g5 = red). Then, the corresponding set of five N-best lists for N = 4.

If the recognizer has to commit to a single word per utterance, its output would be (pain; prose; garden; flower; sad). That is, only “garden” and “flower” would be chosen correctly. This motivates the need for an approach that can disambiguate between the different alternatives in the list. For the types of the algorithms being considered, the best possible answer would be (Jane; rose; garden; flower; null). The last word is null since list L5 does not contain the ground truth tag g5 = red. Therefore, the maximum achievable precision is 1 and recall is 4 / 5 . Suppose some approach is applied to this case, and its answer is W =(Jane; rose; garden; power; null), that is, it picks “power” instead of “flower” and thus only “Jane,” “rose,” and “garden” tags are correct. Then, Precision (W) = 3 / 4 and Recall (W) = 3/5.

4. EXPERIMENTS

In this section, we empirically evaluate the proposed approach in terms of both the quality and efficiency on real and synthetic data sets. Data sets. We test the proposed approach on three data sets. The data sets have been generated by web crawling of a popular image hosting website Flickr.

1. Global is a data set consisting of 60,000 Flickr images. We randomly set aside 20 percent of the data for testing (will be called Gtest) and 80 percent for training (Gtrain). We will use portions of Gtest for testing, e.g., 500 random images. The size of the global vocabulary is |VG| = 18;285. Since it is infeasible to provide speech annotations for a large collection of images, the N-best list for this data set have been generated synthetically. Namely, we use the Metaphone algorithm to generate three to four alternatives for the ground truth tags. We also have used parameters to control the uncertainty of the data: the probability that an N-best list will contain the ground truth tag.

2. Local is a data set consisting of images of 65 randomly picked prolific picture takers (at least 100 distinct tags and 100 distinct images). For each user, we randomly set aside 20 percent of the data for testing (Ltest) and use various portions of the remaining 80 percent for training (Ltrain) the local model. The Metaphone algorithm is employed to generate alternatives for the ground truth tags in Ltest.

Experiment 1: (Quality for various noise levels). We randomly picked 20 images from Real and created N –best lists for their annotations using Dragon in two additional noise levels: Medium and High. Medium and High levels have been produced by introducing white Gaussian noise through a speaker.

Since we created real in a Low noise level on 100 images, for a fair comparison, the points corresponding to Low noise levels in the plots are averages over these 20 images, as opposed to the all 100 images. As anticipated, higher noise levels negatively affect performance of all the approaches. In this experiment, the results are consistent in terms of precision, recall, and F-measure: at the bottom is Recognizer, then Direct, thenIndirect, followed by ME, and then by Upper Bound. As expected, Indirect is slightly better than Direct. In turn, ME tends to dominate Indirect. ME consistently outperforms Recognizerby 11-22 percent of F-measure across the noise levels and it is also within 7-20 percent of F-measure from Upper Bound. In the subsequent discussion, we will refer to Real data with the Low level of noise as just Real.

Experiment 2: (Quality versus size of N-Best lists). The F-measure as a function of N (the size of the N -best list) on Real data. For a given N , the N-best lists are generated by taking the original N-best lists from Real data and keeping at most N first elements in them. Increasing N presents a trade-off. Namely, as N increases, the greater is the chance that the ground truth element will appear in the list. At the same time, Direct, Indirect, and ME algorithms are faced with more uncertainty as there will be more options to disambiguate among. The results demonstrate that the potential benefit from the former outweighs the potential loss due to the latter, as the F-measure increases with N . As expected, the results of Indirect are slightly better than those of Direct. As in the previous experiment, ME tends to outperform Indirect.

Experiment 3: (Correlation of direct and indirect scores). we discussed that one of the requirements for the indirect score function is that it should behave similar to the direct score function. The correlation between the two scoring functions. It plots Hit Ratio as a function of Best M, which is the probability that the top sequence according to the direct score is contained within the best M sequences according to the indirect score on Real data set.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes an approach for using speech for text annotation of images. The proposed solution employs semantics captured in the form of correlations among image tags to better disambiguate between alternatives that the speech recognizer suggests. We show that semantics used in this fashion significantly improves the quality of recognition, which, in turn, leads to more accurate annotation. As future work, we plan to incorporate other sources of semantic information, including but not restricted to social network of the picture taker, the picture taker's address book, domainontologies, visual properties of the image, etc

REFERENCES

[1] R. Bayeza-Yates and B. Riberto-Neto, Modern Information Retrieval. Addison-Wesley, 1999.

[2] D.M. Blei and M.I. Jordan, "Modeling Annotated Data," Proc. ACM SIGIR, 2003.

[3] J. Chen, T. Tan, and P. Mulhem, "A Method for Photograph Indexing Using Speech Annotation," Proc. Second IEEE Pacific Rim Conf. Multimedia: Advances in Multimedia Information Processing (PCM), 2001.

[4] S. Chen, D.V. Kalashnikov, and S. Mehrotra, "Adaptive Graphical Approach to Entity Resolution," Proc. ACM/IEEE Joint Conf. Digital Libraries (JCDL), June 2007.

[5] Z. Chen, D.V. Kalashnikov, and S. Mehrotra, "Exploiting Relationships for Object Consolidation," Proc. ACM SIGMOD Workshop Information Quality in Information Systems (IQIS '05), June 2005.

[6] Z.S. Chen, D.V. Kalashnikov, and S. Mehrotra, "Exploiting Context Analysis for Combining Multiple Entity Resolution Systems," Proc. ACM SIGMOD, June/July 2009.

[7] C. Desai, D.V. Kalashnikov, S. Mehrotra, and N. Venkatasubramanian, "Using Semantics for Speech Annotation of Images," Proc. IEEE Int'l Conf. Data Eng. (ICDE), Mar./Apr. 2009.

[8] O. Dı´az, J. Iturrioz, and C. Arellano, "Facing Tagging Data Scattering," Proc. Int'l Conf. Web Information Systems Eng. (WISE), 2009.

[9] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, vol. 42, nos. 1/2, pp. 177- 196, 2001.

[10] Y. Jin, L. Khan, L. Wang, and M. Awad, Image Annotations by Combining Multiple Evidence & Wordnet," Proc. ACM Int'l Conf. Multimedia, pp. 706-715, 2005.

About Authors:

Mr.HarishBabu Kalidasu has completed his Bachelor of Technology in Andhra University. He is pursuing Post graduation in M.Tech (CSE) at JNTU Kakinada University, At present he is working as Asst.Professor CSE Dept, in Priyadarshini Institute of Technology & Science, Tenali. He has 3 years of experience in teaching field. He is author or co-author of more than five papers includes Networking, software engineering.

Mr B.Prasanna Kumar has completed his Bachelor of Technology of CSE in Rao & Naidu College of Engineering Ongole. He is completed M.Tech (CSE) in University College, Acharya Nagarjuna University, Guntur, Andhra Pradesh. He has 8 Yrs of experience. In teaching field, presently working as Associate Professor & Head, Department of Computer Science & Engineering at Mandava Institute of Engineering and Technology, Jaggayyapet, Krishna (Dt), Andhra Pradesh.

Mr. K.Purna Prakash has completed B.Tech. in Koneru Lakshmaiah College of Engineering, Vaddeswaram, affiliated to Nagarjuna University. He has completed M.Tech. I.T. in S.R.K.R. Engineering College, Bhimavaram, affiliated to Andhra University. He has 4 years of teaching experience. At present he is working as Asst. Professor in Sir C.R.Reddy College of Engineering, Eluru.