Multi-Instance Learning for Coarsely Labeled Time-Domain Inference

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

In the supervised learning setting, it is essential to have sufficient labeled data; however, in many domains, such as activity recognition, existing labeled data may not be available and the annotation process is often too cumbersome, time-consuming and prone to human error. In this work, we explore the use of Multiple-Instance Learning (MIL) in order to reduce the need for fine-grained labels. We examine the drop in performance on two existing time-domain gesture-annotated datasets and show that MIL given coarse-grain ground-truth annotations can achieve performance metrics comparable with standard supervised Machine Learning approaches given fine-grain labels. We evaluate the performance in a leave-one-participant-out fashion given (1) coarsely labeled field data, (2) finely labeled lab data and (3) coarsely labeled data from the held-out participant. Our analysis shows that we can achieve competitive performance given a small number of fine-grained labels in addition to many coarse-grained labels and that even very few labeled sessions from the held-out participant improve performance significantly. We use this to design a system that gives recommendations to developers on the granularity of the field data, based on an initial lab dataset.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Multi-Instance Learning; Data Collection; Time-Domain; Activity Recognition; Eating Detection; Smoking Detection

INTRODUCTION

The ubiquity of mobile devices has led to a growing body of research in designing and solving gesture recognition tasks. These efforts have enormous implications in the mobile health community, self-tracking fitness industry and the development of state-of-the-art human-computer interfacing. The standard approach to gestural recognition employs an appropriate supervised classifier, which often performs exceptionally well given large amounts of labeled data and a well-chosen feature representation. The bottleneck to this approach is that acquiring sufficient gesture labels may be challenging, time-consuming or costly. While many techniques have been adopted to reduce the data annotation effort, this often comes at the expense of noisy labels due to factors such as human error.

A commonly used lightweight approach to gesture annotation is experience sampling [-1], where human subjects are prompted to label their current activity or recount their previous activity break-down. This is often best suited when the activities span a large enough time interval; otherwise, acquiring fine-grained labels remains difficult and especially prone to human error.

One of the most common solutions to reduce human error in data collection is video annotation. Although video labeling is relatively robust to human error, it is time-consuming, it introduces privacy concerns, and its power consumption is significantly large, making it impractical for collecting large-scale data in the field. Thus, there has been a significant effort to reduce the use of video recordings for annotated data collection while minimizing the label noise. Thomaz et al. [-1] employ an upward-facing camera mounted on a necklace to capture eating gestures in the field; the camera takes a snapshot of the subject every 30 seconds, significantly reducing the power consumption and labeling efforts required. Parate et al. [-1] use a 6-axis inertial sensor equipped on the upper arm in addition to a wrist-worn sensor in order to visualize the arm movements in a virtual 3D environment. This eliminates the need for video recordings while minimally increasing the risk of error. However, the annotation effort remains cumbersome and does not scale well to field data, because the additional armband is obtrusive.

Trabelsi et al. [-1] eliminate the need for training data altogether by using an unsupervised learning approach based on a Hidden Markov Model. While this technique achieves performance comparable to supervised learning approaches, it only provides a partition of the data by class and does not make precise label predictions in the absence of labeled data. When a large number of classes are present or positive labels are sparse, then sufficient annotated data once again becomes essential to realize robust, deployable classification systems.

Every submission will be assigned their own unique DOI string to be included here.

Recent work by Stikic and Schiele [-1] explores the feasibility of using Multi-Instance Learning (MIL) to reduce the labeling effort of activity recognition tasks while incurring minimal additional classification error. Although they show that comparable performance can be achieved with coarse-grained labels, they do not consider the case when the developers provide a small number of fine-grained labels in addition to field data.

In this work we demonstrate the on time-domain inertial data and evaluate the extent to which session-level and gesture-level labels improve performance. We additionally assess the boost in performance given a small number of fine-grained labels from the test user in a leave-one-participant-out evaluation. The

MULTI-INSTANCE LEARNING

In the Multi-Instance Learning (MIL) framework, we jointly consider instances, the atomic units over which predictions are made (i.e. gestures), and bags of instances, which may correspond to sessions or longer, manageable time intervals over which an activity is performed. In the binary setting, each bag is assigned a positive label if at least one instance in the bag is positive; bags with no positive instances are assumed to be negative.

The most naive MIL approach is Single-Instance Learning (SIL) [-1], which makes the usually false assumption that every instance in a positive bag is positive. This reduces the problem to a supervised instance-level classification task, which is generally done using a Support Vector Machine (SVM). When positive instances are sparse, the SIL assumption significantly hurts the classification performance.

In the activity recognition setting, Stikic and Schiele use the Maximum Pattern Margin Formulation (miSVM) originally proposed by Andrews et al. [-1] in order to account for the sparsity of positive bags. Due to the non-convexity of the objective function, they use a heuristic to learn the separating hyperplane. They initially train an SIL SVM, whose decision hyperplane is used to relabel the most positive predictions within positive bags. The SVM is then retrained on the relabeled data and the process is repeated until the labels converge. Although this approach accounts for the sparsity of positive gestures, it tends to over-predict the positive class [?] and has no mechanism to adjust the sensitivity based on known density.

Bunescu and Mooney [-1] deal with the challenge of sparse positive bags by using an adaptive SVM constraint (sMIL). In particular, they formulate the MIL constraint that there exists at least one positive instance in every positive bag X as follows

$$\begin{aligned} w\frac{\phi(X)}{|X|} + b \geq \frac{2-|X|}{|X|} - \xi_X \\ \xi_X \geq 0 \end{aligned}$$

where $w\frac{\phi(X)}{|X|} + b$ is the normalized prediction scores under the feature function ϕ, weights w and bias b, and ξ_X is the non-negative slack parameter that allows some extent of misclassification of instances in X to avoid over-fitting the model to the training data. When the bag size $|X|$ is small, the right-hand side becomes larger, suggesting that smaller positive bags are more informative.

Bunescu and Mooney additionally introduce a balancing parameter η, indicating the expected class distribution of instances within bags. The sparse balancing MIL (sbMIL) approach initially trains a sMIL classifier, then relabels the $\eta\,|X|$ most positive instances as positive and the remaining instances as negative. The final hyperplane is then learned using SIL given the relabeled data.

In this work we employ the sbMIL due to the sparsity of positive labels.

DATA

In order to reason in a practical sense about the trade-off between performance and labeling effort under the MIL formulation, we perform several evaluations on two existing datasets: the lab-20 eating dataset developed by Edison Thomaz [-1] and the RisQ smoking dataset developed by Parate et al. [-1]. In order to assess how well the model generalizes to unseen users, we perform leave-one-participant-out (LOPO) evaluations; that is, the model is trained on all but one participant and then evaluated on the held out participant.

Lab-20 Eating

The lab-20 eating dataset comprises of 25Hz 3-axis accelerometer data collected using a wrist-worn inertial sensor from 20 individuals. Individuals were provided food to eat and were asked to perform other possible confounding actions as they please, including talking on the phone, brushing their teeth and combing their hair. The average duration across participants is 31 minutes 21 seconds and comprises of approximately 48% eating sessions. Note, however, that the proportion of eating gestures is much smaller, since non-eating gestures are frequently present within eating sessions.

We use Thomaz's evaluation as the baseline result for comparison. In his work, he uses a Random Forest classifier over 15 statistical features (mean, variance, skew, kurtosis and root mean square over each axis) extracted over windows of 6 seconds with 50% overlap. He reports a 0.42 average LOPO f1 score. We achieve similar performance using a linear SVM.

RisQ Dataset

The RisQ smoking dataset contains 50Hz fused 9-axis inertial data in the form of quaternions from 15 subjects. Parate reports a precision of 91% and recall of 81%. The pipeline consists of (1) computing the trajectory from the quaternion stream, (2) identifying candidate windows by locating peak-trough-peak patterns, (3) extracting 37 angle, velocity, displacement and duration features, (4) classifying windows using a Random Forest and (5) smoothing the predictions using a Conditional Random Field.

In our work, we use the same pipeline but replace the Random Forest classifier with a sbMIL classifier to allow for sparse labels.

(a) M = 150 (b) M = 300 (c) M = 450

Figure 2: Average LOPO performance of sbMIL on Lab-20 dataset as a function of the bag size given 150, 300 and 450 additional labeled training instances respectively

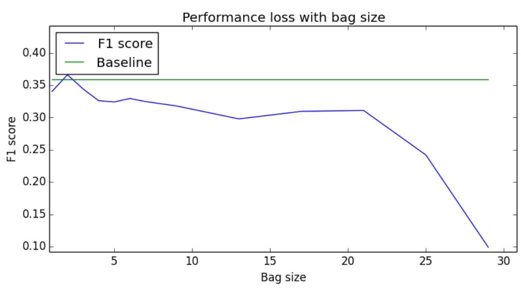


Figure 1: Average LOPO f1 score of sbMIL on Lab-20 dataset as a function of the bag size

EXPERIMENTAL SETUP

In order to reason about the effectiveness of MIL techniques in gesture recognition, we evaluate the average LOPO performance for various bag sizes. Figure 1 shows for the Lab-20 eating dataset that as the bag size decreases, the performance of each MIL technique drops, and it is upper bounded by the baseline SVM performance.

Evidently, the performance is greater given more finely-grained labels. However, given that these labels may be difficult to acquire, we must ask: How many such labels do we need?

In order to address this, we evaluate several experiments in which M fine-grained labels are provided by 5 participants and N coarse-grained labels are provided by the remaining 14 participants. The coarse-grained labels may either be labeled sessions, which may vary in duration, or partitions of the data with a fixed duration. As a personalization step for enhancing performance, we additionally include K instances from the held-out participant in the training data, which are then excluded from the test set. Our experiments involve varying the values of N, M and K.

In each of the experiments, a subset of the training data is used and is therefore selected uniformly from the entire training data; to smooth out noise introduced by the randomness, the performance is averaged over 10 trials. The performance reported is in each case the best performance achieved using cross-validation over the model hyperparameters. These parameters include the expected class weights, the sparse balancing parameter η and the SVM regularization constant C.

EVALUATION

Lab-20 Eating

From figure 1 it is clear that the performance drops very quickly as the granularity of the labels decreases. However, figure 2 demonstrates that this drop in performance is minimal even for large bag sizes, if in additional to coarse labels, fine grained labels are provided. This is shown when 150, 300 or 450 labeled training instances are provided from the lab data. When $M = 450$, the performance remains roughly the same, meaning it may be acceptable to use field data with bag sizes of up to 300.

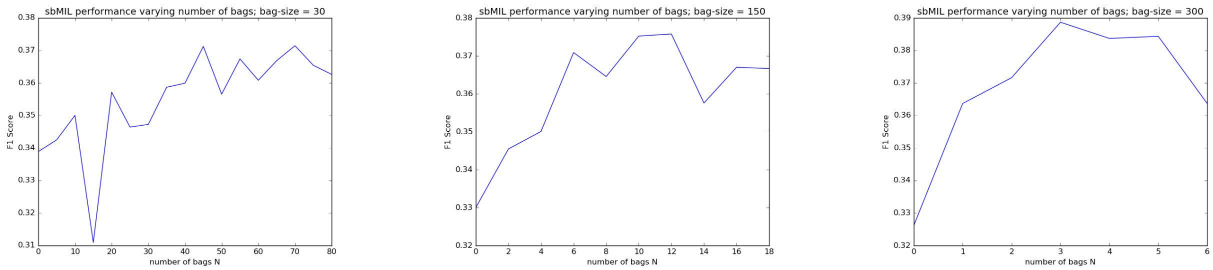
Figure 3 shows the average LOPO performance on the Lab-20 eating dataset as the number of bags increases for bag sizes of 15, 150 and 300 instances. These correspond roughly to 1.5, 15 and 30 minute bags respectively. As the amount of training data increases, the f1 score increases, as expected. Interestingly, the performance is greater given larger bags, even when fewer labels are available. This suggests that many unlabeled instances are preferable to few labeled instances. This is the essential advantage of using MIL techniques.

RisQ Dataset

In the RisQ dataset

TYPESET TEXT

The styles contained in this document have been modified from the default styles to reflect ACM formatting conventions. For example, content paragraphs like this one are formatted using the Normal style.



(a) bag size : 15 (1 min 30 s)　(b) bag size : 150 (15 min)　(c) bag size : 300 (30 min)

Figure 3: Average LOPO performance of sbMIL on Lab-20 dataset as a function of the number of bags given bag sizes of 15, 150 and 300 respectively

	Test Conditions		
Name	*First*	*Second*	*Final*
Marsden	223.0	44	432,321
Nass	22.2	16	234,333
Borriello	22.9	11	93,123
Karat	34.9	2200	103,322

Table 1: Table captions should be placed below the table. We recommend table lines be 1 point, 25% black. Minimize use of table grid lines.

LaTeX sometimes will create overfull lines that extend into columns. To attempt to combat this, the `.cls` file has a command, `\sloppy`, that essentially asks LaTeX to prefer underfull lines with extra whitespace. For more details on this, and info on how to control it more finely, check out `http://www.economics.utoronto.ca/osborne/latex/PMAKEUP.HTM`.

References and Citations

Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets [1, 2, 7]. Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e., you provide the address for obtaining the report within your citation) and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced (e.g., "[Borriello, personal communication]").

References should be in ACM citation format: `http://acm.org/publications/submissions/latex_style`. This includes citations to internet resources [1, 3, 4, 9] according to ACM format, although it is often appropriate to include URLs directly in the text, as above.

SECTIONS

The heading of a section should be in Helvetica or Arial 9-point bold, all in capitals. Sections should *not* be numbered.

Subsections

Headings of subsections should be in Helvetica or Arial 9-point bold with initial letters capitalized. For sub-sections and sub-subsections, a word like *the* or *of* is not capitalized unless it is the first word of the heading.

Sub-subsections

Headings for sub-subsections should be in Helvetica or Arial 9-point italic with initial letters capitalized. Standard `\section`, `\subsection`, and `\subsubsection` commands will work fine in this template.

FIGURES/CAPTIONS

Place figures and tables at the top or bottom of the appropriate column or columns, on the same page as the relevant text (see Figure 1). A figure or table may extend across both columns to a maximum width of 17.78 cm (7 in.).

Captions should be Times New Roman or Times Roman 9-point bold. They should be numbered (e.g., "Table 1" or "Figure 1"), centered and placed beneath the figure or table. Please note that the words "Figure" and "Table" should be spelled out (e.g., "Figure" rather than "Fig.") wherever they occur. Figures, like Figure 2, may span columns and all figures should also include alt text for improved accessibility. Papers and notes may use color figures, which are included in the page limit; the figures must be usable when printed in black-and-white in the proceedings.

The paper may be accompanied by a short video figure up to five minutes in length. However, the paper should stand on its own without the video figure, as the video may not be available to everyone who reads the paper.

Inserting Images

When possible, include a vector formatted graphic (i.e. PDF or EPS). When including bitmaps, use an image editing tool to resize the image at the appropriate printing resolution (usually 300 dpi).

QUOTATIONS

Quotations may be italicized when *"placed inline"* (Anab, 23F).

> Longer quotes, when placed in their own paragraph, need not be italicized or in quotation marks when indented (Ramon, 39M).

LANGUAGE, STYLE, AND CONTENT

The written and spoken language of SIGCHI is English. Spelling and punctuation may use any dialect of English (e.g., British, Canadian, US, etc.) provided this is done consistently. Hyphenation is optional. To ensure suitability for an international audience, please pay attention to the following:

- Write in a straightforward style.
- Try to avoid long or complex sentence structures.
- Briefly define or explain all technical terms that may be unfamiliar to readers.
- Explain all acronyms the first time they are used in your text—e.g., "Digital Signal Processing (DSP)".
- Explain local references (e.g., not everyone knows all city names in a particular country).
- Explain "insider" comments. Ensure that your whole audience understands any reference whose meaning you do not describe (e.g., do not assume that everyone has used a Macintosh or a particular application).
- Explain colloquial language and puns. Understanding phrases like "red herring" may require a local knowledge of English. Humor and irony are difficult to translate.
- Use unambiguous forms for culturally localized concepts, such as times, dates, currencies, and numbers (e.g., "1–5–97" or "5/1/97" may mean 5 January or 1 May, and "seven o'clock" may mean 7:00 am or 19:00). For currencies, indicate equivalences: "Participants were paid ₩ 25,000, or roughly US $22."
- Be careful with the use of gender-specific pronouns (he, she) and other gendered words (chairman, manpower, man-months). Use inclusive language that is gender-neutral (e.g., she or he, they, s/he, chair, staff, staff-hours, person-years). See the *Guidelines for Bias-Free Writing* for further advice and examples regarding gender and other personal attributes [10]. Be particularly aware of considerations around writing about people with disabilities.
- If possible, use the full (extended) alphabetic character set for names of persons, institutions, and places (e.g., Grønbæk, Lafreniére, Sánchez, Nguyễn, Universität, Weißenbach, Züllighoven, Århus, etc.). These characters are already included in most versions and variants of Times, Helvetica, and Arial fonts.

ACCESSIBILITY

The Executive Council of SIGCHI has committed to making SIGCHI conferences more inclusive for researchers, practitioners, and educators with disabilities. As a part of this goal, the all authors are asked to work on improving the accessibility of their submissions. Specifically, we encourage authors to carry out the following five steps:

1. Add alternative text to all figures
2. Mark table headings
3. Add tags to the PDF
4. Verify the default language
5. Set the tab order to "Use Document Structure"

For more information and links to instructions and resources, please see: `http://chi2016.acm.org/accessibility`. The `\hyperref` package allows you to create well tagged PDF files, please see the preamble of this template for an example.

PAGE NUMBERING, HEADERS AND FOOTERS

Your final submission should not contain footer or header information at the top or bottom of each page. Specifically, your final submission should not include page numbers. Initial submissions may include page numbers, but these must be removed for camera-ready. Page numbers will be added to the PDF when the proceedings are assembled.

PRODUCING AND TESTING PDF FILES

We recommend that you produce a PDF version of your submission well before the final deadline. Your PDF file must be ACM DL Compliant. The requirements for an ACM Compliant PDF are available at: `http://www.sheridanprinting.com/typedept/ACM-distilling-settings.htm`.

Test your PDF file by viewing or printing it with the same software we will use when we receive it, Adobe Acrobat Reader Version 10. This is widely available at no cost. Note that most reviewers will use a North American/European version of Acrobat reader, so please check your PDF accordingly.

When creating your PDF from Word, ensure that you generate a tagged PDF from improved accessibility. This can be done by using the Adobe PDF add-in, also called PDFMaker. Select Acrobat | Preferences from the ribbon and ensure that "Enable Accessibility and Reflow with tagged Adobe PDF" is selected. You can then generate a tagged PDF by selecting "Create PDF" from the Acrobat ribbon.

CONCLUSION

It is important that you write for the SIGCHI audience. Please read previous years' proceedings to understand the writing style and conventions that successful authors have used. It is particularly important that you state clearly what you have done, not merely what you plan to do, and explain how your work is different from previously published work, i.e., the unique contribution that your work makes to the field. Please consider what the reader will learn from your submission, and how they will find your work useful. If you write with these questions in mind, your work is more likely to be successful, both in being accepted into the conference, and in influencing the work of our field.

ACKNOWLEDGMENTS

Sample text: We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Authors 1, 2, and 3 gratefully acknowledge the grant from NSF (#1234–2012–ABC). *This whole paragraph is just an example.*

REFERENCES FORMAT

Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced (e.g., [Golovchinsky, personal communication]). References must be the same font size as other body text. References should be in alphabetical order by last name of first author. Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets. For papers from conference proceedings, include the title of the paper and the name of the conference. Do not include the location of the conference or the exact date; do include the page numbers if available.

References should be in ACM citation format: http://www.acm.org/publications/submissions/latex_style. This includes citations to Internet resources [4, 3, 9] according to ACM format, although it is often appropriate to include URLs directly in the text, as above. Example reference formatting for individual journal articles [2], articles in conference proceedings [7], books [10], theses [11], book chapters [12], an entire journal issue [6], websites [1, 3], tweets [4], patents [5], games [8], and online videos [9] is given here. See the examples of citations at the end of this document and in the accompanying BibTeX document. This formatting is a edited version of the format automatically generated by the ACM Digital Library (http://dl.acm.org) as “ACM Ref.” DOI and/or URL links are optional but encouraged as are full first names. Note that the Hyperlink style used throughout this document uses blue links; however, URLs in the references section may optionally appear in black.

REFERENCES

1. ACM. 1998. How to Classify Works Using ACM’s Computing Classification System. (1998). http://www.acm.org/class/how_to_use.html.
2. R. E. Anderson. 1992. Social Impacts of Computing: Codes of Professional Ethics. *Social Science Computer Review December* 10, 4 (1992), 453–469. DOI: http://dx.doi.org/10.1177/089443939201000402
3. Anna Cavender, Shari Trewin, and Vicki Hanson. 2014. Accessible Writing Guide. (2014). http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/.
4. @_CHINOSAUR. 2014. "VENUE IS TOO COLD" #BINGO #CHI2014. Tweet. (1 May 2014). Retrieved Febuary 2, 2015 from https://twitter.com/_CHINOSAUR/status/461864317415989248.
5. Morton L. Heilig. 1962. Sensorama Simulator. U.S. Patent 3,050,870. (28 August 1962). Filed Februrary 22, 1962.
6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal and Ubiquitous Computing* 18, 4 (2014), 765–766. DOI: http://dx.doi.org/10.1007/s00779-014-0773-4
7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where Do Web Sites Come from?: Capturing and Interacting with Design History. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’02)*. ACM, New York, NY, USA, 1–8. DOI:http://dx.doi.org/10.1145/503376.503378
8. Nintendo R&D1 and Intelligent Systems. 1994. *Super Metroid*. Game [SNES]. (18 April 1994). Nintendo, Kyoto, Japan. Played August 2011.
9. Psy. 2012. Gangnam Style. Video. (15 July 2012). Retrieved August 22, 2014 from https://www.youtube.com/watch?v=9bZkp7q19f0.
10. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing*. ERIC, Bloomington, IN, USA.
11. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.
12. Langdon Winner. 1999. *The Social Shaping of Technology* (2nd ed.). Open University Press, UK, Chapter Do artifacts have politics?, 28–40.