[[1]](#footnote-1)

Template for Preparation of Papers for IEEE Sponsored Conferences & Symposia\*

First A. Author, Second B. Author, Jr., and Third C. Author, Member, IEEE

*Abstract*—In the supervised learning setting, sufficient labeled data is essential for strong performance; 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 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 given coarse-grain ground-truth annotations we can achieve performance metrics comparable with standard supervised Machine Learning approaches that require 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 boost performance significantly.

# 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 and self-tracking fitness industry. In particular, there is growing interest in detecting smoking gestures with the goal to prevent relapses for patients seeking smoking cessation. Health providers are also interested in detecting eating behaviors to minimize risk of obesity and type II diabetes. Robust systems built on these models are becoming the backbone to our prospective personalized health care system. However, these systems remain far from realization because their underlying models continue to rely on large amounts of quality labeled data, which is often difficult to acquire. The standard approach to such gestural recognition tasks employs a suitable supervised classifier, which often performs exceptionally well given substantial labeled gestures and a well-engineered feature representation. The bottleneck to this approach is that acquiring sufficient labeled gestures may be challenging, time-consuming or costly. This is of particular interest for data collected in the field, which is essential for building a generalizable, deployable system. In this work we explore a class of models Multiple-Instance Learning that we show can be used to learn to classify gestures given coarsely labeled sessions containing those gestures at a tolerable decrease in performance.

In the general sense, Multi-Instance Learning (MIL) is a class of supervised learning techniques for problems where only partial knowledge of the training labels is available. It was first introduced by Dietterich et al. [10] to address the problem of predicting activity levels of drug components and has since been used in many applications, including content-based image retrieval [9], scene classification [11] and mutagenesis predictions [12]. However, these techniques have rarely been used in the gesture recognition setting, because gestural data exhibits strong temporal dependencies that violate the usual MIL assumption that bags as well as their instances are independently and identically distributed [13]. Although analogous spatial dependencies exist in images, MIL techniques work exceptionally well for tasks such as classifying objects within images, given only training labels associated with the entire image. In the same spirit, we examine the effectiveness of MIL for identifying fine-grained gestures when only coarse-grained sessions labels are available at training time. We demonstrate that if a small fixed amount of labeled gestures are provided, then MIL techniques can be used to reduce the need for additional fine-grained labels while achieving competitive performance metrics on two existing datasets: Lab-20 Eating dataset [5] and RisQ Smoking dataset [1]. We now proceed with recent work on existing data collection techniques, introduce the Multiple-Instance framework, describe the datasets and report results and analysis of several relevant evaluations.

# Related Work

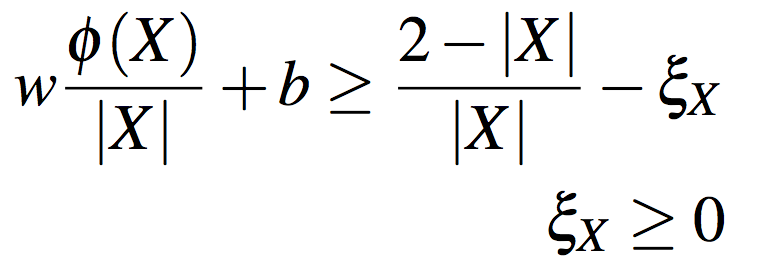
While many techniques have been adopted to reduce the data annotation effort, this often comes at the expense of noisy labels due to factors such as human error.
A commonly used lightweight approach to gesture annotation is experience sampling [6] where human subjects are prompted to label their current activity or recount their previous activity break-down. This is often best suited when the activities span a large time interval comprising mostly of positive gestures; this is often not the case for the smoking and eating detection tasks.
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, introduces privacy concerns, and consumes significant power, making it impractical for collecting large-scale field data. 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. [5] 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. However, due to the low temporal resolution, they use these labels only as test data for detecting eating sessions (meals), not for detecting individual gestures. 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 the data collection does not scale well to field data, because the additional armband is obtrusive.
Trabelsi et al. [8] 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 abundant annotated data once again becomes essential to realize robust, deployable classification systems.
Recent work by Stikic and Schiele [7] explores the feasibility of using Multiple-Instance Learning 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 effectiveness of MIL 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 in contrast to coarse-grained labels from the test user in a leave-one-participant-out evaluation.

# Multiple-Instance Learning

# In the Multiple-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) [4], 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.

# When positive instances are sparse, the SIL assumption significantly hurts the classification performance. In the activity recognition setting, Stikic and Schiele [7] use the Maximum Pattern Margin Formulation (miSVM) originally proposed by Andrews et al. [2] 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 [3] 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

# 



# where 1/|X| w (X) + b is the normalized prediction scores under the feature function , weights *w* and bias *b*, and 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 , indicating the expected class distribution of instances within bags. The sparse balancing MIL (sbMIL) approach initially trains a sMIL classifier, then relabels |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 implementation provided in [4] due to the sparsity of positive instances.

# 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 [5] 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.

## A. 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. This generates 12379 labeled instances, of which 1480 (11.96%) are eating. He reports a 0.42 average f1 score over LOPO evaluations. We achieve the same performance using a linear SVM under the same experimental protocol.

## B. RisQ data

The RisQ smoking dataset consists of 50Hz fused 9-axis inertial data in the form of quaternions from 14 subjects. The raw data stream is converted into a local trajectory in 3D space. Classification is done using a Random Forest, followed by a Conditional Random Field for smoothing predictions, over feature vectors of candidate windows identified by locating peak-trough-peak patterns indicative of smoking gestures. There are 11,900 candidate windows, of which 358 (3.00%) are smoking gestures. A total of 37 features are extracted, including angular, velocity, displacement and duration features. Parate et al. report a LOPO precision of 91% and recall of 81%, which corresponds to f1 score of 85.7%.
In our work, we use the same computational pipeline but replace the Random Forest classifier with a sbMIL classifier to allow for sparse labels.

# 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. Fig. 1 shows that for the Lab-20 eating dataset 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 over a fixed number of participants and *N* coarse-grained labels are provided by the remaining participants. In the Lab-20 dataset, fine-grained labels are acquired from 5 participants and coarse-grained labels from the remaining 14 participants. In the RisQ dataset, they are acquired from 5 and 13 participants respectively. 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, bags or sessions from the held-out participant in the training data, which are excluded from the test set. Our experiments involve varying the values of *N*, *M* and *K*.
In each of the experiments, we run several trials to smooth out noisy evaluations. 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  and the SVM regularization constant *C*. These parameters are learned using 5-fold cross-validation, where each validation fold consists of data from one of the 5 instance-level participants, in order to mimic the true test setting as accurately as possible from the training data. We use a randomized grid search with 20 iterations over the parameter space.

## A. Lab-20 Eating

Fig. 1 shows in blue the average F1 score over all LOPO evaluations, varying the label granularity over a fixed subset of 1500 instances (15-20%) of the Lab-20 eating data. The baseline standard SVM performance is shown in green for comparison. From fig. 1 it is clear that the performance drops very quickly as the granularity of the labels decreases.
However, fig. 2 demonstrates that this drop in performance is minimal even for large bag sizes, if in addition to coarse-grained labels, fine-grained labels are provided. More precisely, fig. 2 shows the average F1 score over all LOPO evaluations, varying the label granularity but fixing the number of finely labeled training instances $M$ from the lab data. This is shown when 150, 300 or 450 labeled training instances are provided from the lab data. In each case, the number of training bags from the field remains constant but the granularity of the labels over those bags is varied. When *M* = 150, the F1 score drops noticeably; however, remains much larger than when no single instances are included, as shown in fig. 1. When *M* = 300, the performance drop is insignificant, and when *M* = 450, the performance remains roughly the same, indicating that it may be acceptable to use field data with bag sizes of up to 300.
This alone could alternatively suggest that the additional field data we are providing does not give a significant boost in performance. To show that it indeed does increase the performance, we consider the case when the number of labeled training instances *M* is fixed and the number of training bags *N* varies.
Fig. 3 shows the average LOPO F1 score 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. The number of labeled training instances is fixed at *M* = 1500. 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.

## B. RisQ data

To show that this model generalizes to other datasets, we perform similar tests on the RisQ smoking dataset. Fig. 4 shows the average LOPO F1 score of the sbMIL classifier for various number of labeled sessions. Here, a session is a variable length time period of smoking, in which at least one smoking gesture occurred. The number of labeled instances is fixed at *M* = 1500, approximately 15-20% of the total training data. We see nearly a 15% increase in performance given only 5 additional labeled sessions, resulting in an F1 score comparable to the Random Forest/Condition Random Field baseline performance reported in the original RisQ pipeline.

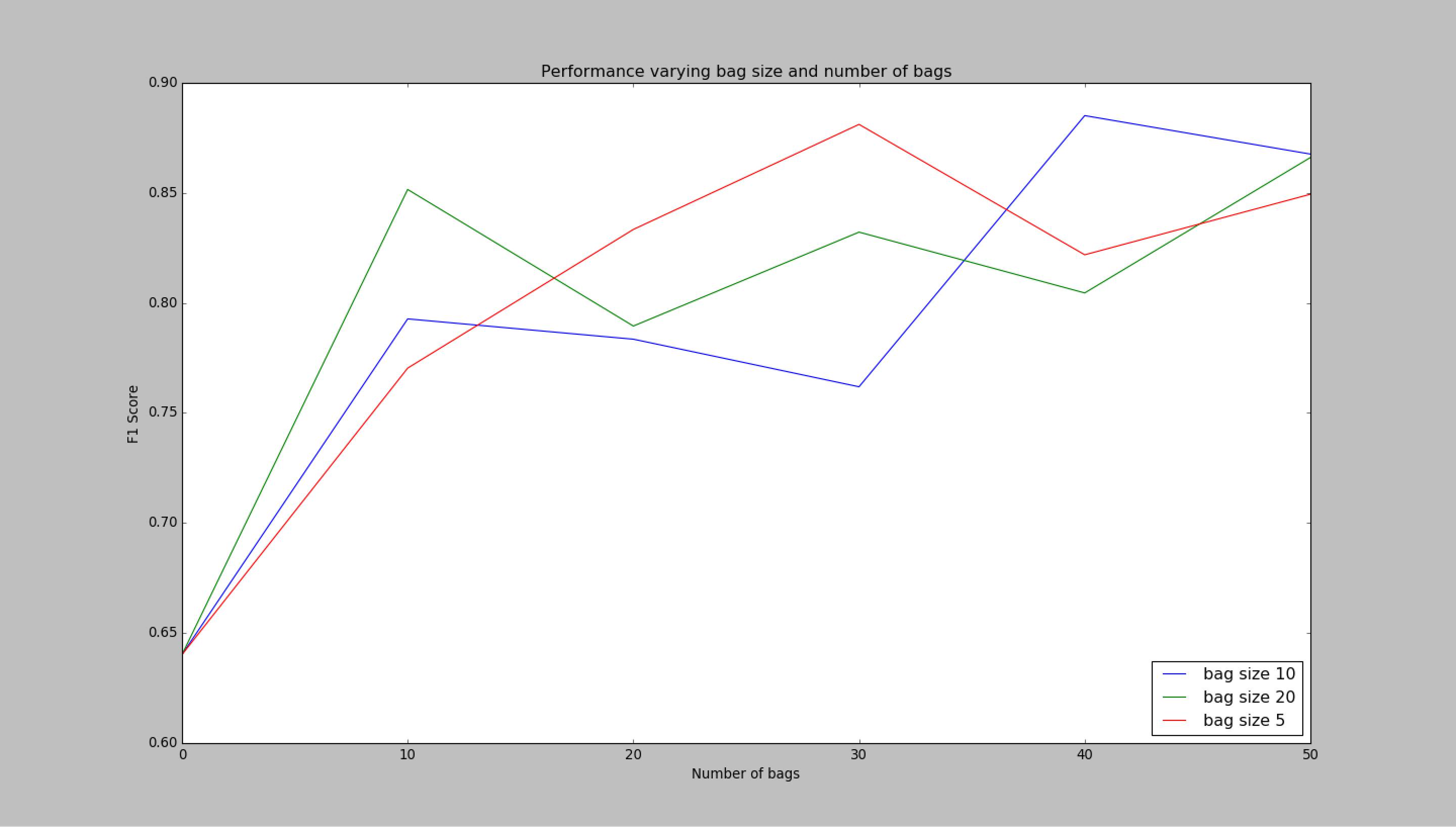
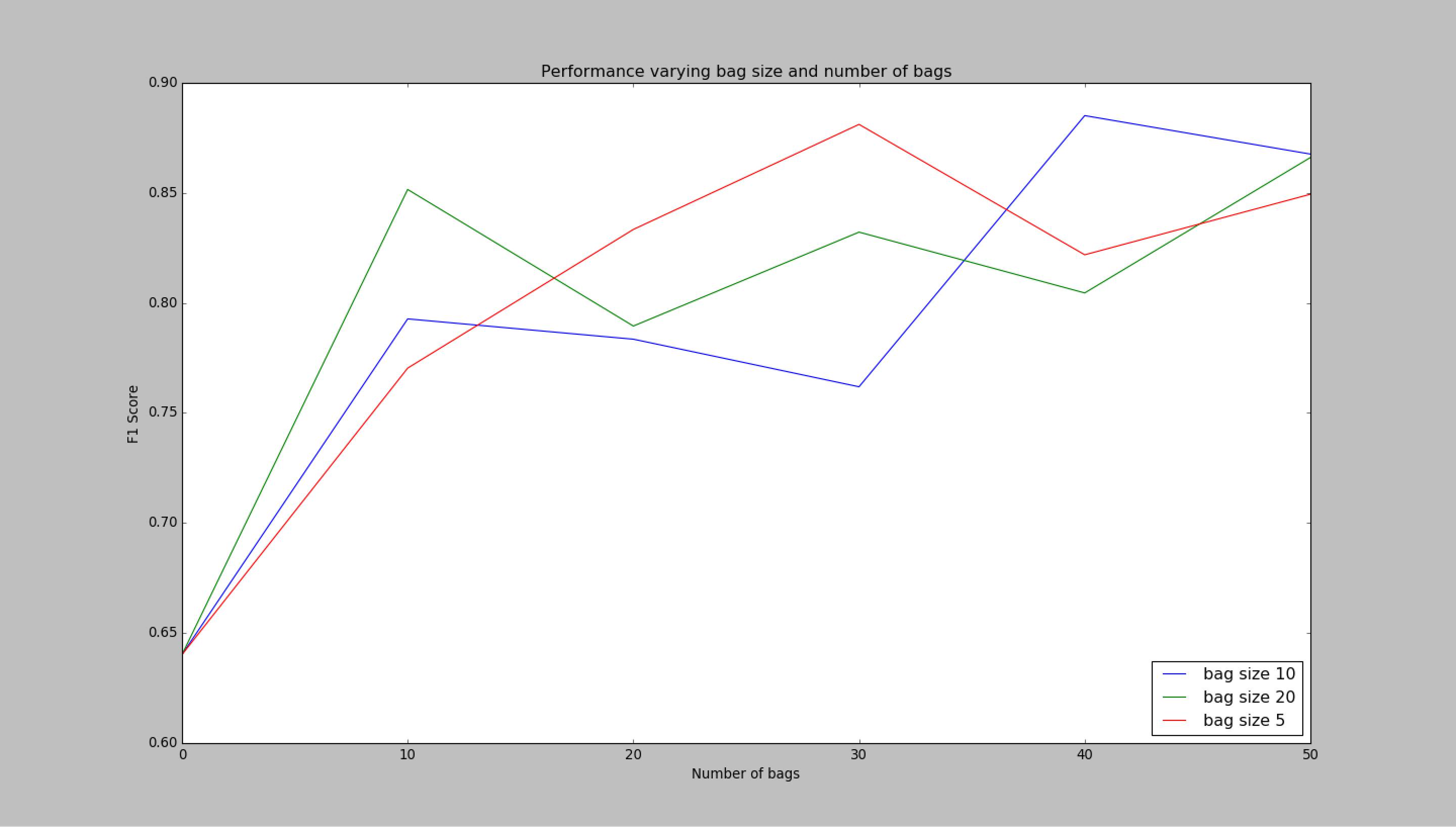
# Evaluation

## A. Coarse-grained Training Data

Fig. \ref{fig:figure1} shows in blue the average f1 score over all LOPO evaluations, varying the label granularity over a fixed subset of 1500 instances (15-20\%) of the

## B. Personalization

Lab-20 eating data. The baseline standard SVM performance is shown in green for comparison. From Figure \ref{fig:figure1} it is clear that the performance drops very quickly as the granularity of the labels decreases.

possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

#### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

## B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## C. Figures and Tables

### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

a. Sample of a Table footnote. (Table footnote)

1. Example of a figure caption. *(figure caption)*

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K.”

# Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Appendix

Appendixes should appear before the acknowledgment.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

1. Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. 2014. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM.
2. Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*. MIT Press, 561–568.
3. Razvan C. Bunescu and Raymond J. Mooney. 2007. Multiple Instance Learning for Sparse Positive Bags. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML-2007)*. Corvallis, OR. http://nn.cs.utexas.edu/?bunescu:icml07
4. Gary Doran and Soumya Ray. 2014. A Theoretical and Empirical Analysis of Support Vector Machine Methods for Multiple-instance Classification. *Mach. Learn.* 97, 1-2 (Oct. 2014), 79–102. DOI: http://dx.doi.org/10.1007/s10994- 013- 5429- 5
5. Edison Thomaz, Irfan Essa, and Gregory D. Abowd. 2015. A Practical Approach for Recognizing Eating Moments with Wrist-Mounted Inertial Sensing. In *Proceedings of the ACM International Joint Conference on Pervasive Ubiquitous Computing*. ACM.
6. Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. 2007. MyExperience: A System for in Situ Tracing and Capturing of User Feedback on Mobile Phones. In *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (MobiSys ’07)*. ACM, New York, NY, USA, 57–70. DOI: http://dx.doi.org/10.1145/1247660.1247670
7. Maja Stikic and Bernt Schiele. 2009. *Location and Context Awareness: 4th International Symposium, LoCA 2009 Tokyo, Japan, May 7-8, 2009 Proceedings*. Springer Berlin Heidelberg, Berlin, Heidelberg, Chapter Activity Recognition from Sparsely Labeled Data Using Multi-Instance Learning, 156–173. DOI: http://dx.doi.org/10.1007/978- 3- 642- 01721- 6\_10
8. D. Trabelsi, Samer Mohammed, F. Chamroukhi, L. Oukhellou, and Yacine Amirat. 2013. An unsupervised approach for automatic activity recognition based on hidden Markov model regression. *IEEE Transactions on Automation Science and Engineering* Regular paper, in press, DOI: 10.1109/TASE.2013.2256349 (2013).
9. Oded Maron , Tomás Lozano-Pérez, A framework for multiple-instance learning, *Proc. of the 1997 Conf. on Advances in Neural Information Processing Systems* 10, p.570-576, 1998
10. T. G. Dietterich, R. H. Lathrop, T. Lozano-Perez. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence Journal, 89, 1997.
11. Zhou, Z.-H. & Zhang, M.-L. 2006. Multi-instance multi-label learning with application to scene classification. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems, MIT Press, 1609–1616.
12. Chevaleyre, Y. & Zucker, J.-D. 2001. Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. Application to the mutagenesis problem. In Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, Springer, 204–214.
13. Dan Zhang, Yan Liu, Luo Si, Jian Zhang, and Richard D. Lawrence. 2011. Multiple Instance Learning on Structured Data. In *Advances in Neural Information Processing Systems 24*, J. Shawe-taylor, R.s. Zemel, P. Bartlett, F.c.n. Pereira, and K.q. Weinberger (Eds.). 145–153. http: //books.nips.cc/papers/files/nips24/NIPS2011\_0127.pdf.

1. \*Research supported by ABC Foundation.

   F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

   S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

   T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp). [↑](#footnote-ref-1)