A Developmental Approach to Concept Learning

Liesl Wigand, Monica Nicolescu, Mircea Nicolescu
Department of Computer Science and Engineering, University of Nevada, Reno
{*lieslw, monica, mircea*}*@cse.unr.edu*

Keywords: Machine Learning : Support Vector Machine : Isomap : Dimensionality Reduction : Classification

Abstract: The ability to learn new concepts is essential for any robot to be successful in real-world applications. This is due to the fact that it is impractical for a robot designer to pre-endow it with all the concepts that it would encounter during its operational lifetime. In this context, it becomes necessary that the robot is able to acquire new concepts, in a real-world context, from cues provided in natural, unconstrained interactions, similar to a human-teaching approach. However, existing approaches on concept learning from visual images and abstract concept learning address this problem in a manner that makes them unsuitable for learning in an embodied, real-world environment. This paper presents a developmental approach to concept learning. The proposed system learns abstract, generic features of objects and associates words from sentences referring to those objects with the features, thus providing a grounding for the meaning of the words. The method thus allows the system to later identify such features in previously unseen images. The paper presents results obtained on data acquired with a Kinect camera and on synthetic images.

1 INTRODUCTION

The ability to learn new concepts is essential for any robot to be successful in real-world applications. This is due to the fact that it is impractical for a robot designer to endow the robot with all the concepts that it would encounter during its operational lifetime. In this context, it becomes necessary that the robot is able to acquire new concepts, in a real-world context, from cues provided in a natural, human-like teaching approach. The problem of *concept learning* has been widely studied in the fields of computer vision and machine learning, both as *concept learning from visual images*, or as *abstract concept learning*. However, the existing approaches address this problem in a manner that makes them unsuitable for learning in an embodied, real-world environment.

The field of computer vision has provided a wealth of approaches for learning of concepts from visual images. In the vast majority, the learning problem is to associate each image with a unique class, of which the object is a member. The focus is on creating algorithms that better discriminate between the members of different classes. While being important for a large class of applications, this approach limits the learning problem to a single "feature" of the object, which in most cases is the object's name. Psychological research (Bloom, 2002; Horst et al., 2005) indicates that humans perceive and use objects from the perspective of their multiple characteristic features or possible uses. Features such as size, texture, weight, etc. are typically present in all objects. However, the standard classification approach prevents any generalization of objects at this level. For example, a system that would be trained with images of *blue books* (class 1) and *red balls* (class 2) would only be able to distinguish between those classes. If presented with a *red book* the system would be unable to say anything about the new object, although it had seen *books* and *red* objects before. Humans learn to recognize these separate features and are able to generalize from them, although they might have not seen an identical object before. The ability to generalize at this level is essential for any robotic system that is to be used in real-world applications, given that it is impractical to pre-train a robot with all possible objects it might need to recognize for its tasks.

The concept learning problem has also been widely addressed in the field of machine learning. The learning problem is stated as inducing *abstract concepts* from combinations of multiple features. For each feature, appropriate values are provided in a *feature-value* pairs array, which are used to generalize the characteristic features of the concept. While these methods are very successful when given training information in the above form, they could not be di-

rectly applicable to a system that needs to learn from real-world interactions. First, the symbol grounding problem is avoided by providing the learning system with direct labels for the training samples. Second, an assumption is made that it is known which value corresponds to each feature, for example such that *small* and *large* are values of the *size* feature. Third, it is also assumed that a single training sample contains information about all the relevant features of the concept, which are known in advance and pre-selected by the user. In practice, robots would need to learn from information extracted from the sensors, and would have to solve the symbol grounding problem by relating their observations with verbal cues provided by a human user. The learning information is also not structured as complete *feature-value* arrays, but rather comes from natural means of communication (such as speech or visual cues), which by nature cannot express information in this way. A human user might only mention that an object is *large* or *small*, but not that these values relate to the object's size. In addition, it is not ensured that a human would enumerate all the attributes of an object of interest, to provide the complete information.

This paper takes a *developmental approach to concept learning*, in order to address the above limitations. The hypothesis is that robots need to be able to learn from visual and auditory cues during interactions with human teachers, in an incremental fashion, in a manner similar to how young children acquire new concepts.

The remainder of the paper is structured as follows: Section 2 discusses related research, Section 3 describes our approach, Section 4 presents our results and Section 5 gives a summary of our work.

2 RELATED WORK

Concept learning is a significant research problem, which has been addressed in computer science, cognitive science, neuroscience and psychology. This section presents related research in these areas.

The goal of psychology and child development research, as it relates to concept learning, is to understand the mechanisms that underlie the formation of concepts in children and humans in general. Various aspects of this problem have been explored. (Schyns et al., 1998) explore the interplay between the high-level cognitive process over the perceptual system, which gives rise to new concept formation. (Feldman, 2003) proposes a principle that indicates that people induce the simplest categories consistent with a given set of examples and introduces an algebra for representing human concept learning. (Kaplan and Murphy, 200) evaluate the effect of prior knowledge on category learning and suggest that the category exemplars as well as prior knowledge about the category's domain influence the learning process. The concept learning approach proposed in this paper aims to build a feature space for representing the concepts. The issue of category dimensionality has been examined in (Hoffman and Murphy, 2006), supporting the motivation to address this problem at the level of the object feature space. This approach is consistent with findings in child psychology research, which indicate that children start by learning the individual features and only form a single category after more extensive familiarization (Horst et al., 2005).

This paper takes the view that a robot should learn by using both language and vision input, which studies in neuroscience and psychology have found likely in human children (Scholl, 2005), (Pinker, 2007). The simple comparison of sights and sounds may allow an infant to develop a world model, and development relies on interaction with people and the environment. For more information on developmental robotics see (Lungarella et al., 2003). Most previous work done with images and text has been done in data mining. For example, images from the internet can be automatically associated with labels, as those on websites like Flickr, or webpages related to keywords can be retrieved. Usually the focus of these works is not to learn the meaning of the words but to accurately label the images so that a user may find them quickly with a text search. However this is really the same problem, and many of these techniques may be applied here, especially methods used to eliminate poor labels which are common in internet databases (Brodley and Friedl, 1999). There has been much work done purely on images or on text, such as in the cases of document retrieval and content-based image retrieval, which rely on word features or image features, not on both. For a more complete review of existing methods see (Lew et al., 2006).

The field of computer vision provides a wide spectrum of approaches to this problem. (Huang and LeCun, 2006) proposes a combination of support vector machines (SVMs) and convolutional nets to characterize objects in variable conditions of illuminations and with multiple different viewpoints. (Yang and Kuo, 2000) uses content as the relevant feature to categorize images. (Wolfgang Einhauser and Konig, 2002) demonstrates how a hierarchical neural network evolves structures invariant to features such as color and orientation, consistent with physiological findings. (Piater and Grupen, 2000) presents an incremental approach to learning the set of features nec-

essary for visual classification; whenever the system has difficulty classifying an image, it seeks new features that are capable of helping differentiate between the multiple classes. Our proposed system assumes a one class classification problem, where we can only guarantee positive examples, not negatives. We make use of one class SVMs, although it may be beneficial to try neural networks in the future, which are similar to SVMs in behavior, but less sensitive to parameter choice (Khan and Madden, 2010). Techniques from these methods may be used in a system that relies on both images and text to find correspondences. For example the features used to relate images may be used to represent images in our system, and the processes used to eliminate unimportant words could be used the same way in our system. For now we use features provided by Isomap in order to avoid features that may be specific to a word, and assume that the result of classification can be used to indicate whether a word should be eliminated (Uzwyshyn, 2009).

The work proposed in this paper departs from the standard computer vision and machine learning techniques in several important directions. With regards to computer vision, the difference in our techniques is that we aim at learning concepts at the level of their main characteristic features, rather than at the level of a single class. In the proposed approach, one object will be a member of multiple classes (e.g., a *book* can be both *large* and *red*) and different objects will be members of the same class (e.g. a *book* or a *ball* could be both *large*). The goal is to learn a multi-dimensional space of *features*, which could be used to characterize previously unseen objects. With respect to the machine learning techniques, this work departs from the assumptions related to the structure of the training data and propose an approach that uses visual and text input, similar to that which would be provided in natural interactive scenarios.

3 APPROACH

Our general approach is similar to how parents teach young children things about the world: while pointing to an object that captures the childs attention, or by showing an object to the child, a parent describes the object saying things like "That's *yellow*", "Look at the *big* box!", "Keep this stick *vertical*". Over time, and with sufficient examples, the child learns the meaning of *yellow*, *big* and *vertical* and is able to recognize these features in objects previously unseen.

The hypothesis of our work, as indicated by methods such as Isomap (Tenebaum et al., 2000) is that lower dimensionality spaces obtained through such methods incorporate significant relevant features of the data, such as for example an *object's size*, *orientation*, or *shading*. However, these algorithms stop at the level of classification, and do not attempt to automatically infer the relation that the actual object features have with the reduced feature space. The goal of this work is to build on the dimensionality reduction paradigm in order to provide an automated way for learning the correspondence between the reduced feature space and the physical features of the data.

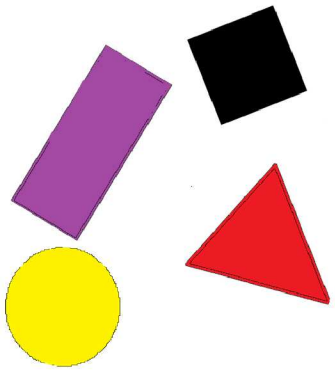

Figure 1: Example synthetic images.

Our methodology follows a two stage approach: (1) unsupervised dimensionality reduction for feature extraction and (2) learning of word-feature relations. The validation consists of characterizing a previously unseen object in terms of its features, based on the learned feature space.

We created a database of images, combining artificial images and images of objects gathered using a Kinect camera. We described all the images using complete English sentences, similar to utterances that a parent would say to a child when talking about the object in the image. The artificial data consists of 500 by 500 color images of shapes of varying colors, sizes and positions on a plain white background, as seen in Figure 1. For each image a text file is created by a human which consists of a sentence per line describing the shape or some feature of the image. At the moment these sentences are assumed to be positive to avoid any language processing, for example a sentence may be "The shape is *round*" but cannot be "The shape isn't *round*." There are also suplemental files which attempt to list negative labels, which are used to check results.

In practice this may not be complete enough: one user may say a thing is round, and another may disagree. For now we assume that labels are not subjective. We also have a greyscale generated dataset, in order to test accuracy with color removed. The

Figure 2: Example masked Kinect images from the RGB-D database.

non-synthetic dataset is a collection of Kinect images gathered in the lab and combined with the RGB-D dataset (Lai et al., 2012), at 640 by 480 resolution. These images were masked to remove the background and every object is roughly at center, with some noise from the background still present (see Figure 2). These were similarly labeled by users who desrcibed them in complete sentences, and supplemented with a set of negative labels for truth testing. All images are currently stored as PNG files. A simple data collection system was constructed to get images of more objects. This system used a Kinect and the Point Cloud Library to grab point clouds from the Kinect (Bogdan Rusu and Cousins, 2011). These point clouds are then processed using a depth filter to remove most of the background, and a planar segmenter based on RANSAC which removes the flat surface on which objects are assumed to lie. From this data an image is generated with a white value given to all empty space. There is no additional noise removal, since it did not seem necessary in this setup. In future projects the depth information may be used as well to more accurately describe the concepts to be learned.

3.1 Feature extraction

In this step we perform feature extraction by applying Isomap to the image training data, which reduces the feature space from several hundred thousand to around a hundred features. This improves generalization and shortens training times. Isomap is a form of multi-dimensional scaling which preserves geodesic distance rather than Euclidean. Generally, the user selects an initial goal number of dimensions to reduce to, although some algorithms reduce dimensions until an error threshold is crossed. This algorithm may have problems if too few or too many neighbors are used for the search, and also has trouble if points are moved off the manifold by noise. This does not have a significant effect in the current system, but for better results on real data it may be necessary to change to another method.

Several other dimensionality reduction techniques were considered, and may be used in the future. One which was implemented and tested, created overlapping patches of the images, and used K-Means Clustering to determine template types of patches in the images, as seen in (Coates et al., 2010). These types were then used as features, so that patches would be collected from an image, and its representative feature vector would indicate which patch types were present in the image. It was determined that k would need to be very large to avoid loosing information, especially with real data. Isomap, in this case, was better for our data, but also requires more memory and time. This method may be useful on larger datasets, where the gain in dimensions becomes less important then speed and space taken up by the reduction method.

3.2 Learning word-feature relations

Once the feature vectors for the images have been created, a one class SVM is trained for each word, as follows. The sentences describing the images are stripped of punctuation, and transformed to all lower case. In the future both punctuation and capitalization may be useful to differentiate words, but for now they are assumed to be unimportant. Words that occur too frequently or infrequently are removed and no SVM is trained. For now this limit is set at less than 5% of samples or more than 95%. This is to avoid both "stop" words which occur everywhere (*is, a, the*), and rare words from which the SVM can learn little (*vertically* for example, occured once.) This is a "bag of words" approach, and it may be useful at a later time to use bigrams (word pairs) as labels to determine correlation of features. For each word, for example *red*, we collected all of the images for which that word has occured in any sentence relative to it. This produced a set of images in which there is a high probability that an object with the color red may be present. For each word, using this subset of images we used a one class support vector machine. Due to the assumption that words in sentences are positive labels we do not have any images labeled as negative examples. Furthermore, as mentioned previously, deciding negative examples may prove difficult as descriptions may be subjective. SVMs are good at classifying high dimensional data, and are efficiently represented as well, although one class SVMs may not perform as well. A one class SVM assumes that the training data is only of a single class, and attempts to fit a hyperplane to that class. This results in a very strict border on the training data, which can be a problem if the training

data does not represent that class completely. At the present the system is using a radial basis function with a degree of 3, and 0.3 as the upper bound on training error, although degree of 2, 4 and errors of .2, .4 and .5 were also tested, and did not perform as well across the data. In future versions these parameters will be set using grid search on a per word basis.

4 RESULTS

The system was tested on a set of 128 generated images and 134 real objects. This data was split into positive training cases and a mix of positive and unknown testing cases. These included features and objects that were not seen in the training data. The test cases were in some cases labeled negative in the afore mentioned ground truth, and other times unlabeled if the true label was ambiguous (for example, a label such as "several" which may or may not describe a learnable feature). Where the true label was undefined, the result was not counted for or against the system's accuracy, but was used as input for manual adjustment of the system. This included updating the ground truth, and testing methods for adding positive examples. The dictionary for generated data was 91 words, and the real data was 195 words. It was implemented in Python, using the Python Image Library and numpy to store and process images, and scipy to plot data. Scikit learn was used for scaling, Isomap, and one class SVMs (Pedregosa et al., 2011). Training takes roughly 40 minutes without memory limits, but nearly an hour and a half with memory limits.

4.1 Feature extraction results

The isomap dimensionality reduction was given a limit of 100 dimensions to reduce to. The generated data reduced from 750000 features (raw color images) to 75 features. This is essential for the success of the SVM, which is generally good for datasets where the number of features exceed the number of samples, but still performs poorly if the number of features greatly exceeds the number of samples. It is possible that the raising the dimension of the final feature space will aid in learning certain complex words, but this may also cause the SVMs to fail.

4.2 Word-feature learning results

The results for the synthetic data indicate that features such as *color*, *shape* and *location* can all be learned in this simple manner. Many of the words had too many or too few examples, but these words were mainly stop words. For example *shape* occurs in almost every case, as in "The shape is *yellow*." We calculate precision, recall and F1 score based on the predicted labels compared to the ground truth. If a label is not indicated in the ground truth, the example is not used in calculating the scores, but is listed as an "unknown." The results are summarized in Table 1 for a subset of the data, with Table 2 containing several words which performed well, but with many unknown ground truth labels. Essentially these words may be learnable, but should have improved ground-truth for example *corners* was used with rectangles, squares and triangles, but due to 18 images not being labeled as either having corners or not, further testing is needed to verify these results.

Table 3 contains similar results with a different error permitted by the one class SVM. The error for this was increased to 0.3 permitted, which resulted in significantly better scores on some labels such as *four* and *square*, but worse results on others such as *red*. This indicates that the process used here can perform better, but it is impossible to know how significant the change will be.

The real data did not perform as well as the artificial images. While colors which were well represented did well, almost every other word in the data set was used only once, or only with a single object which occured several times. The result is that a word like *shiny* might occur only with regards to a flashlight, and could then be associated with any of the flashlight's features. It is likely that a much larger dataset will be necessary for real data to show any valuable results, since the dictionary is too broad to be captured in a hundred objects. The results from the words that succeeded may need more examples as well, since the features have a broader range on real objects. For example *red* in the generated data set is a single color, however, *red* in real objects is many shades. While the results indicate that the system is learning color across shades, if more shades

Table 1: Generated Data Results: 128 total samples, permitted error of 0.2, degree 3, radial basis kernel.

Word	Samples	Precision	Recall	F1
"yellow"	22	1.0	0.45	0.62
"four"	26	1.0	0.5	0.66
"black"	37	1.0	0.62	0.76
"circle"	37	1.0	0.54	0.70
"triangle"	38	1.0	0.52	0.68
"square"	21	1.0	0.33	0.5
"blue"	18	1.0	0.5	0.66
"red"	17	1.0	0.64	0.78

were present that may change. To see the color results for real data see Table 4.

4.3 Discussion and future work

We found that when collecting sentences about the objects, they were not described completely. So a *red square* may be described as *red*, or *square*, but not always as both. This should not be a concern as long as the training data covers the class well. Another difficulty results from our system's lack of knowledge about the langauge: it cannot make assumptions about the presence of one word implying the absence of another, for example a *red* object may also be described as *orange*. This means that as well as being unable to determine if a label is present or not if the word is missing, we can also make no assumptions about exclusivity of the labels. Since this is in part, a feature of human language, it makes creation of ground truth labels difficult. In training many people may agree that an object is *red*, but another person may say it is *maroon*, making it difficult to say that a label is correct or not. For now we assume that our ground truth set is accurate and does not conflict with the user labels.

The main problem of the one class classifier is a high false negative rate. Since it fits to only positive examples, it can exclude negatives very well. However, if the training data did not cover the class well, or if the hyperplane did not fit to all examples (degree too small, permitted error too small), or there is noise in the data which moves it off of the hyperplane, then the classifier can have a high false negative rate. This can be seen in the analysis of our system's results. Handling this problem is a major goal of future work.

Many of the parameters (SVM error, SVM degree, Isomap neighbors, Isomap dimension) need to be set procedurally to find the best combination. Isomap takes up a large amount of time and space, and is not robust to noise. It may be necesary to replace it with another method, especially if more data is added. More Kinect data is needed, of more objects, and with varying features. More sentences describing objects should also be collected, to broaden the vocabulary and collect more examples of rarely used words. Similarly the ground truth set needs to be updated, due to the fact that many of the words describing the real data were not listed as either true or false, and so these results could not be verified. This may require a better categorization, since several words have multiple meanings. This means it may be necessary to identify multiple clusters within a word's training set and treat them as separate words. Other techniques for processing images, and correcting labels which were found during a literature search were not implemented due to time constraints, but may improve future results. Furthermore, we plan to develop an interactive approach to data collection, in which auditory and visual training data is acquired directly by a robot through interactions with human users.

Table 2: Generated Data Results with many unknowns: 128 total samples, permitted error of 0.2, degree 3, radial basis kernel.

Word	Samples	Precision	Recall	F1
"three"	24	1.0	0.66	0.8
"top"	19	1.0	0.63	0.77
"oval"	27	1.0	0.55	0.71
"right"	19	1.0	0.47	0.64
"corners"	46	0.96	0.56	0.71
"round"	34	1.0	0.55	0.71
"upper"	16	1.0	0.5	0.66
"shape"	77	1.0	0.74	0.85

Table 3: Generated Data Results: 128 total samples, permitted error of 0.2, degree 3, radial basis kernel.

Word	Samples	Precision	Recall	F1
"yellow"	22	1.0	0.36	0.53
"four"	26	1.0	0.57	0.73
"black"	37	1.0	0.56	0.72
"circle"	37	1.0	0.62	0.76
"triangle"	38	1.0	0.52	0.68
"square"	21	1.0	0.66	0.8
"blue"	18	1.0	0.33	0.5
"red"	17	1.0	0.41	0.58

5 CONCLUSION

This paper presented an approach to developmental concept learning from images and text, in order to associate attributes extracted from the images with words. The system relies on feature extraction and one class classification to accomplish this goal. The results indicate reasonable success of around .7 to .8

Table 4: Real Data Results: 128 total samples, permitted error of 0.3, degree 3, radial basis kernel.

Word	Samples	Precision	Recall	F1
"yellow"	22	1.0	0.64	0.78
"black"	34	1.0	0.64	0.78
"blue"	35	1.0	0.74	0.85
"purple"	10	1.0	0.6	0.75
"red"	60	0.83	0.73	0.77
"white"	36	1.0	0.58	0.73

F1 score on simple *color*, *shape* and *location* words. In general, lower scores resulted where few examples were given, or the word was more complex, as in the cases of *vertically*, *shiny*, or even *handle*. Overall the results are encouraging, although better results on non-synthetic data is necessary to prove the utility of this approach.

Further extending this approach can lead to a system that can be used as a basis for learning human communication. This will allow for better robot collaborators, which can learn interactively in similar ways in which humans learn from each other.

REFERENCES

Bloom, P. (2002). *How Children Learn the Meaning of Words*. MIT Press.

Bogdan Rusu, R. and Cousins, S. (2011). 3d is here: Point cloud library (pcl). In *IEEE International Robotics and Automation (ICRA)*.

Brodley, C. E. and Friedl, M. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167.

Coates, A., Lee, H., and Ng, A. Y. (2010). An analysis of single-layer networks in unsupervised feature learning. In *Advances in Neural Information Processing Systems*.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, 12(6):227–232.

Hoffman, A. B. and Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3):301–315.

Horst, J., Oakes, L., and Madole, K. (2005). What does it look like and what can it do? category structure influences how infants categorize. *Child Development*, 76(3):614–631.

Huang, F. J. and LeCun, Y. (2006). Large-scale learning with svm and convolutional nets for generic object characterization. In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*.

Kaplan, A. and Murphy, G. (200). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26(4):829–846.

Khan, S. S. and Madden, M. G. (2010). A survey of recent trends in one class classification. In *Articial Intelligence and Cognitive Science - 20th Irish Conference*.

Lai, K., Bo, L., Ren, X., and Fox, D. (2012). Rgb-d (kinect) object database. http://www.cs.washington.edu/rgbd-dataset/index.html.

Lew, M., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art challenges. *ACM Transactions on Multimedia Computing, COmmunications, and Applications*.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: A survey. *Connection Science*, 15:151–190.

Pedregosa, F., Varoquaux, G., Granfort, A., Michel, V., Thirion, B., Grisel, O., M., B., Prettenhoffer, P., and et. al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12:2825–283.

Piater, J. and Grupen, R. (2000). Constructive feature learning and the development of visual expertise. In *Proc., Intl. Conf. on Machine Learning*, Stanford, CA. Morgan Kaufmann.

Pinker, S. (2007). *The Language Instinct*. Harper Perennial Modern Classics, New York.

Scholl, B. (2005). *The innate mind: structure and contents*. Oxford University Press.

Schyns, P., Goldstone, R., and Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21(1):1–54.

Tenebaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–1323.

Uzwyshyn, R. (2009). An arbitrage oppurtunity for image search and retrieval. *Bulletin for the American Society for Information Science and Technology*, 35.

Wolfgang Einhauser, Christoph Kayser, K. K. and Konig, P. (2002). Learning multiple feature representations from natural image sequences. In *Proc., Intl. Conf. on Artificial Neural Networks*.

Yang, Z. and Kuo, J. (2000). Learning image similarities and categories from content analysis and relevance feedback. In *Proc., ACM workshops on Multimedia*, pages 175–178.