Multimodal Ranking for Non-Compliance Detection in Retail Surveillance

Hoang Trinh Sharath Pankanti Quanfu Fan
IBM T. J. Watson Research Center
19 Skylikne Dr, Hawthorne, NY 10532

Abstract

In retail stores, cashier non-compliance activities at the Point of Sale (POS) are one of the prevalent sources of retail loss. In this paper, we propose a novel approach to reliably rank the list of detected non-compliance activities of a given retail surveillance system, thereby provide a means of significantly reducing the false alarms and improving the precision in non-compliance detection. Our approach represents each detected non-compliance activity using multimodal features coming from video data, transaction logs (TLog) data and intermediate results of the video analytics. We then learn a binary classifier that successfully separate true positives and false positives in a labeled training set. A confidence score for each detection can then be computed using the decision value of the trained classifier, and a ranked list of detections can be formed based on this score. The benefit from having this ranked list is two-fold. First, a large number of false alarms can be avoided by simply keeping the top part of the list and discarding the rest. Second, a trade off between precision and recall can easily be performed by sliding the discarding threshold along this ranked list.

Experimental results on a large scale dataset captured from real stores demonstrate that our approach achieves better precision than a state-of-the-art system at the same recall. Our approach can also reach an operating point that exceeds the retailers' expectation in terms of precision, while retaining an acceptable recall of more than 60%.

1. Introduction

Checkout non-compliance at the Point of Sale (POS) in grocery stores is one of the major sources of retail loss, costing retailer billions of dollars annually. A checkout non-compliance occurs when an item passes the POS without being registered, e.g., failing to trigger the barcode scanner. This non-compliance event, either done intentionally or unintentionally by the cashier, leads to the same outcome: the item is not charged to the customer, causing loss to the store. For the rest of this paper we will refer to checkout non-compliance activities (both intentional and unintentional) as *nonscan*.

Many video analytics systems such as [14, 4, 1, 5, 15] have been introduced to detect nonscans, showing more advantages than human surveillance in effectiveness, efficiency and scalability. The ultimate goal of these systems is to detect nonscans, create a real-time alert for each of them for human verification. These systems accomplish this task by detecting checkout activities of cashiers during transactions and identify unmatched evidence through joint analysis of cashier activities and logged transaction data (TLog). The predominant cashier activity at the checkout is a repetitive activity called *visual scan*, which is constituted by three distinctive primitive checkout activities: *pick-up*, *scan* and *drop-off*, corresponding to the process of registering one item by the cashier in a transaction, as illustrated in Figure 3. By aligning these detected visual scans with the barcode signals from the POS device, nonscan events can be isolated and detected. The aforementioned process is illustrated in Figure 1.

However when it comes to the real-world deployment of such systems, there are still vast technical challenges to overcome. Changing viewpoints, occlusions, and cluttered background are just a few of those challenges from a realistic environment that any automatic video surveillance system has to handle. In addition, in the low-margin retail business (nonscans occur with much lower frequency than regular visual scans), it is crucial that such a system be designed with careful control of alarms rate (AR) while still being scalable. The alarms rate is defined as the total number of detected nonscans divided by the total number of scanned items. A high alarms rate will make it almost impossible for a human verifier to scan through all the alarms, which means that the probability of finding true nonscans would decrease.

In this paper, we describe a novel algorithm to significantly reduce the false alarms and enhance the precision in nonscan detection of a given retail surveillance system. It can be implemented as an optional postprocessing component of a given retail surveillance system, as represented by the dashed blue box in Figure 1. Our approach takes as in-

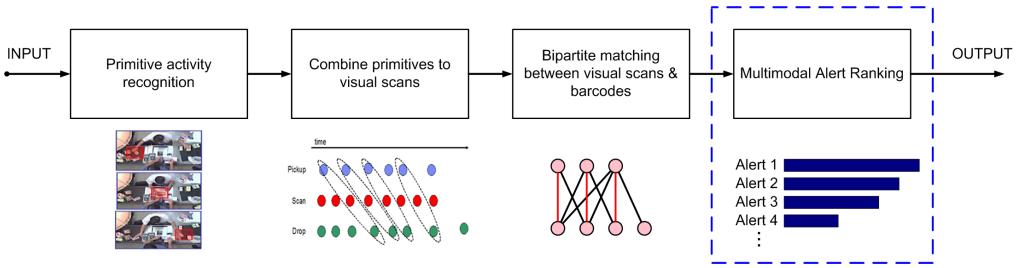


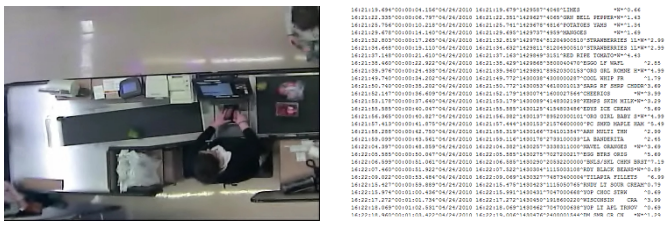
Figure 1. The workflow of a typical retail video surveillance system. Our main contribution is the multimodal alert ranking component (the dashed blue box), which takes as input the list of detected nonscans of the current system, and returns a ranked list of those nonscans based on their confidence scores.

put the list of detected nonscans of the current system, and returns a ranked list of those nonscans in descending order of their confidence scores. Although our approach is not able to improve recall of the system, precision can be much improved in the following sense: given a specific number of detected nonscans starting from the top of the list, more true nonscans should be found in the ranked list than in the original (unranked) list. The intuition is that with the ranking mechanism, the human verifiers only need to review much fewer detected nonscans, while still being able to find a significant portion of true cases.

Our approach starts by representing each detected noncompliance activity using multimodal features coming from video data, TLog data and intermediate results of the video analytics. We use both local features, which are associated with the nonscan, and contextual features associated with the transaction containing that nonscan, such as: time, duration and proximity of the nonscan; temporal relationship with other events. We then learn a binary Support Vector Machine (SVM) classifier that successfully separate true positives and false positives in a labeled training set. A confidence score for each detection can then be computed using the decision value of the trained SVM classifier, and a ranked list of detections can be formed based on this score.

For retail video surveillance systems, there are usually two available input data sources: the video stream and the TLog data stream, as illustrated in Figure 2. This paper advocates that by combining the two streams (video and TLog) of data with the intermediate system results, much can be learned about the nature of a nonscan event performed by a cashier. By analyzing local and contextual features around the detected nonscan, more discriminative power can be obtained that helps identifying a true nonscan versus its false counterpart in a robust way against the presence of noise in either the TLog or the video data.

Our testing results on a large scale video dataset from real-world retail stores demonstrate that our approach achieves a significantly better detection precision, as well as alarms rate, at the same detection rate (recall) compared to the state-of-the-art surveillance systems. Our approach also has the flexibility to tune the ROC curve to the desired operating point that best suits the user's expectation.



Video data Transaction log

Figure 2. The input data for a retail video surveillance system consists of the video data stream from the camera and the TLog data stream from the POS device.

2. Related Work

There has been many works on retail video surveillance for sweethearting detection [14, 4, 1, 5, 6, 15, 7, 11, 8]. However, none of the approaches above employed a feature-based representation of nonscans. These approaches also lack the ability to assign a meaningful confidence score for each returned alert, therefore lack the ability to rank and threshold them. The main contribution of our approach is to achieve significantly better precision with this ability.

Previous approaches using SVM with multiple data streams for ranking purposes exist. In [10], the authors learned retrieval functions from the query-log of the search engine in combination with the log of links the users clicked on. SVM was used to learn these functions, which was shown to improve the query retrieval quality. In [16], the authors present a multimodal active learning approach, with application to semantic feature extraction in multimedia data. The approach in [2] combines multiple modalities and applies it to the problem of recognizing the affective state of interest. In [12], an approach to segmenting news video based on the perceived shift in content using features spanning multiple modalities was introduced. However, since their goal is segmentation rather than ranking, they used a maximum entropy story boundary selection process. Each of these work bear some differences to our approach. In our approach, we train a binary classifier on multimodal

features that can provide a confidence score for each new example based on the classification decision value, which then enables us to perform ranking.

Research works on image and video search [13, 9] also make use of multimodal features such as visual features and user tags for object representation. However, in the search application, all the objects in the database are ranked by their distance w.r.t to the querying object, i.e.: ranking itself is only meaningful relative to the given query. Furthermore, since the searching task has to be able to handle different types of object categories, the features in use have to be sufficiently generic. In our approach, the selected features are application-specific, which particularly capture the information directly and naturally related to retail activities of interest.

3. Our Approach

3.1. Representing Nonscans using Multimodal Features

A nonscan is defined as a visual scan activity that cannot be associated to any barcode signal. Therefore it is a complex human activity that can only be defined in the joint domain of multiple modalities, in this case, by both the video data and the TLog data, or more precisely, by the misalignment between these two data streams. This is the key difference between nonscan detection and other human activity recognition tasks.

We represent each detected nonscan by a feature vector of 11 dimensions. We employ both local features, which is associated with the nonscan, and contextual features - associated with the transaction containing that nonscan. Specifically we use the following 12 features:

1. Temporal distance from the detected nonscan to the nearest *key-in* event.

   A *key-in* event usually happens with some particular types of items without barcodes (produce, vegetable, fruits, etc) in which cases the cashier has to interact with the scanning device (keyboard, touch-screen, etc) instead of just passing the item through the scanning area. This is one of the main causes of false positives. Therefore, this feature has positive correlation with the probability of the detected nonscan being a true positive (The larger the distance, the more likely that it is a true nonscan).

2. Temporal distance from the detected nonscan to the nearest barcode signal.

   Since a true nonscan tends not to occur around a barcode signal, this feature has positive correlation with the probability of the detected nonscan being a true positive.

3. Temporal distance from the nonscan to the nearest other nonscan.

   Based on the intuition that true nonscans tend not to happen in quick succession, this feature has positive correlation with the probability of the detected nonscan being a true positive.

4. The scan gap difference computed by Equation (1).

$$D_i^{scangap} = |T_{i-1} + T_{i+1} - 2 * T_i| \quad (1)$$

   where T_i is the scan gap of the i^{th} nonscan.

   This feature can be considered the first-order derivative of the scan gap.

5. Distance from the nonscan to the start of transaction, normalized by the duration of the transaction.

6. Distance from the nonscan to the start of transaction, normalized by the duration of the transaction.

   A lot of false positives are caused by non-checkout activities (e.g.: membership card, payment, receipt handling, etc), which tend to occur around start and end of a transaction. Therefore these features (5) and (6) have positive correlation with the probability of the detected nonscan being a true positive.

7. Total number of overlapping *pickup* events inside the nonscan event.

8. Total number of overlapping *scan* events inside the nonscan event.

9. Total number of overlapping *drop* events inside the nonscan event.

   A valid visual scan usually has one of each primitives. The occurrence of multiple primitives inside one single visual scan is usually caused by noisy primitive detection result. Therefore these three features $(7-9)$ have negative correlation with the probability of the detected nonscan being a true positive.

10. Total number of barcode signals in the transaction.

11. Average scangap in the transaction.

12. Average scan gap difference in the transaction. We compute the average of scan gap difference over all visual scans in the transaction, using the same formula in equation (1).

The features (1) through (9) are local features. i.e. features directly associated with the nonscan of interest. Features (10) through (12) are contextual features, taking into

Figure 3. A repetitive activity called "visual scan" includes three distinctive activities: (a) *pick-up*, (b) *scan* and (c) *drop-off*, corresponding to the process of registering one item by the cashier in a transaction. A "visual scan" that does not correspond to any barcode signal is considered a *nonscan*. (d) A *key-in* event occurs when the cashier interacts with the POS devices such as keyboard, touch screen, which is usually performed in place of the *scan* event for items without barcodes (e.g.: produce). (*Best viewed in color*).

account the global information of the whole transaction in which the nonscan is detected.

Please note that the computation of these features requires the detections of all primitive activities such as *key-in*, *pick*, *scan* and *drop* (Figure 3). Since we implement our approach as a post-processing stage of an existing video surveillance system, all these four primitive events are intermediate results of the video analytics being employed. Both [5] and in particular [15] can detect such activities with high accuracy and robustness.

Figure 4 illustrated our data analysis based on the three contextual transaction-level features. Figure 4 (a), (b), (c) shows the variations in each feature of the transactions with true nonscans, and demonstrates that these features have discriminative power to transactions containing true nonscans. Specifically, in Figure 4 (a), most of the transactions containing nonscans have the average scan gap between the 30^{th} percentile and the 80^{th} percentile. In Figure 4 (b), most of the transactions containing nonscans have the average scan gap difference between the 30^{th} percentile and the 80^{th} percentile. In Figure 4 (c), most of the transactions containing nonscans have the number of items above the 60^{th} percentile.

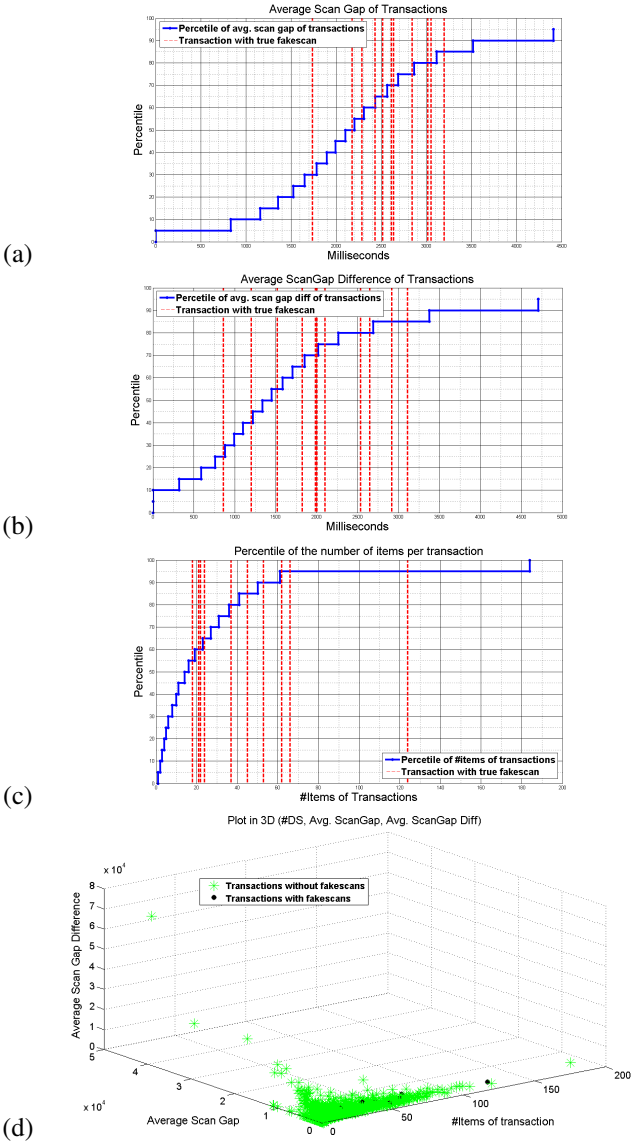


Figure 4. (a) Percentile plot of average transaction scan gap. (b) Percentile plot of average transaction scan gap difference. (c) Percentile plot of the number of items per transaction. Vertical red lines indicates transactions containing true nonscan events. (d) Distribution of transaction over the 3-D features formed by the three features above.(*Best viewed in color*).

3.2. SVM-based Multimodal Ranking

We extract training data from a large retail surveillance video dataset capturing one-day long checkout activities from 6 checkout lanes in two real retail stores to train a two-class SVM classifier. The dataset accounts for a total of $32,969$ items being checked out within 1660 transactions (note that each transaction can have multiple checked out items). This challenging dataset presents large variances in cashiers, backgrounds, camera angles, with significant occlusions, and distractions from customers. The videos are at frame rate 20 FPS and low resolution (320x240). Ground truth nonscans are manually annotated for these videos.

The training data include:

- 65 positive examples: these are conceptually true nonscans annotated from the training data.
- 370 negative examples: these are conceptually true visual scans annotated from the training data.

We use the SVM implementation in [3], with RBF kernel. For numerical stability, we scale the training data to the

range $[-1, +1]$. The package also provides a tool for parameter selection, using grid search with cross-validation. We apply this tool to select the optimal values for the parameters C and γ, where C is the penalty parameter of the error term in the SVM objective function, and γ is the scaling parameter of the RBF kernel. The classifier trained by SVM is a hyperplane in the feature space that successfully separate the positive examples from the negative examples.

In test time, we can compute a confidence score of a detected nonscan based on how well that nonscan is classified as a true case by the SVM classifier. Specifically, the score is computed based on the decision values or probability estimates from the SVM prediction results. After the classifier is trained, we propose a ranking mechanism as follows.

- Use the trained SVM classifier to predict the label of each detected nonscan (as true or false nonscan).
- Compute the confidence score of each detected nonscan using the decision values or probability estimates from SVM classification.
- Rank the list of detected nonscans based on the confidence scores of nonscans.

4. Experimental Results

Dataset: The testing data include 443 nonscans detected from the same large scale training dataset described in Section 3.2. There is no overlap between the nonscans selected for testing and the nonscans selected for training. We apply the same scaling parameters that were used for the training set to scale the test set to the range $[-1, +1]$.

Evaluation Metrics: The detection rate (recall) is defined to be the number of detected true nonscan divided by the total number of true nonscans in the dataset. The precision is defined to be the number of detected true nonscans divided by the total number of detected nonscans. The alarms rate is defined as the total number of detected nonscans divided by the total number of scanned items in the dataset. Although precision is a more meaningful concept in terms of performance evaluation, in practice the users are more interested in the alarms rate. The alarms rate and the precision are related to each other following Equation (2).

$$Precision = \frac{N_{tp}}{AR \times N_{items}} \quad (2)$$

where N_{tp} is the number of detected true positive nonscans , N_{items} is the total number of scanned items.

Comparative evaluation: We apply our approach as a postprocessing component of a state-of-the-art surveillance system described in [15]. We rank the list of nonscans detected by [15] based on their confidence scores, and returns nonscans with scores above a certain threshold.

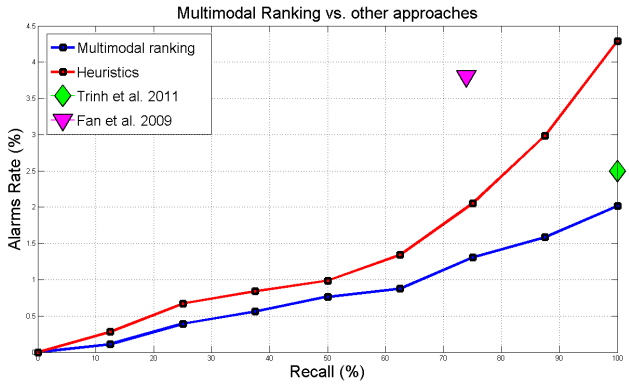


Figure 5. Performance of the multimodal ranking approach in terms of recall and alarms rate. For the same recall (detection rate), our approach achieves better alarms rate compared to the heuristics-based approach, as well as compared to the best performance of [15] and [5].

We compare our results against: (i) a baseline heuristic ranking method that will be described later in this section, (ii) best performance of [15], and (iii) best performance of [5].

After the ranked list of nonscans is obtained, we have the flexibility to select a certain score threshold and return all the nonscans with scores above that threshold. For each of these selection, we have a different precision/recall value pairs. A higher the threshold will result in a higher precision and a lower recall. A full ROC curve can be generated by sliding the threshold along the ranked list.

Figure 5 illustrates the comparative evaluation in terms of alarms rate and detection rate.

It is demonstrated in Figure 5 that our approach significantly reduce the alarms rate, at the same detection rate compared to the best performance of [5] and [15]. Please note that since we take the list of detected nonscans from [15] as input, we can only achieve the same recall as them at best. However, we achieve a better precision than [15] at 100% recall.

The results strongly suggests that our approach is an efficient way to improve the precision of the system. Since we argued earlier that the probability of finding true nonscan cases by human verification will increase as the alarms rate decreases, this achievement of our approach brings direct profit to the retailers.

Figure 6 illustrates a comparative evaluation in terms of precision/recall. Again, at the same recall, our approach achieves the best precision compared to all other methods.

Heuristic Ranking Method: We implement a baseline heuristic ranking method as follows. For the heuristic rank-

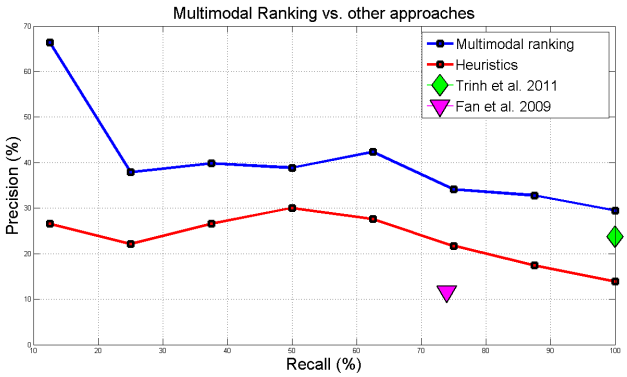


Figure 6. Performance of the multimodal ranking approach in terms of recall and precision. For the same recall (detection rate), our approach achieves better precision compared to the heuristics-based approach, as well as compared to the best performance of [15] and [5].

ing method, we utilize the same features as described in Section 3.1, and compute the confidence score for each nonscan heuristically as a linear function of the feature values, as in Equation 3.

$$score(s) = \sum_i b_i f_i(s) \quad (3)$$

where $b_i \in \{-1, 1\}$ indicates the direction of correlation of feature f_i with the confidence score of the nonscan s.

Figure 5 and 6 both demonstrate that for the same detection rate, we consistently achieve significantly lower alarms rate and precision than the baseline heuristic method. This superiority should be entirely contributed to the usage of the SVM classifier. While the heuristics approach uses prior domain knowledge about how each feature relates to the confidence score in an intuitive way, SVM can automatically learn the weights of each feature, leading to a more reliable confidence value computation.

5. Conclusion

In this paper, we describe a novel approach to significantly improving the precision in checkout non-compliance activity detection of an existing retail surveillance system. Our approach is based on investigating different features from multiple modalities such as: video data, TLog data and visually detected checkout related events called *primitives*. We represent each fraudulent event detected by the current system as a feature vector in a high-dimensional feature space, and train an SVM classifier that successfully separates the true detections from the false detections. We show that the SVM can determine the most effective features which serve as indicators of true fraudulent events, and model the variations of such events based on these features. A ranking scheme based on the decision values of the trained classifier is employed to obtain a ranked list of detected nonscans, in which the top nonscans in the list are most likely to be the true nonscans.

Testing results on data from real-world retail stores demonstrate that our approach has the flexibility to achieve a significantly better detection precision at the same recall compared to the current state-of-the-art surveillance systems.

References

[1] Agilence. http://www.agilenceinc.com/. 241, 242

[2] Y. I. Ashish Kapoor, Rosalind W. Picard. Probabilistic combination of multiple modalities to detect interest. In *ICPR*, 2004. 242

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm. 244

[4] A. Dynamics. http://www.americandynamics.net/. 241, 242

[5] Q. Fan, R. Bobbit, Y. Zhai, A. Yanagawa, S. Pankanti, and A. Hampapur. Recognition of repetitive sequential human activity. In *CVPR*, 2009. 241, 242, 244, 245, 246

[6] Q. Fan, A. Yanagawa, R. Bobbit, Y. Zhai, R. Kjeldsen, S. Pankanti, and A. Hampapur. Detecting sweethearting in retail surveillance videos. In *ICASSP*, 2009. 242

[7] Q. Fan, A. Yanagawa, R. Bobbit, Y. Zhai, S. Pankanti, and A. Hampapur. Fast detection of retail fraud by using polar touch buttons. In *ICME*, 2009. 242

[8] P. Gabbur, Q. Fan, H. Trinh, and S. Pankanti. A pattern discovery approach to retail fraud detection. In *ACM SIGKDD*, 2011. 242

[9] S. C. H. Hoi and M. R. Lyu. A multimodal and multilevel ranking scheme for large-scale video retrieval. 2008. 243

[10] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, 2002. 242

[11] J. Pan, Q. Fan, S. Pankanti, H. Trinh, P. Gabbur, and S. Miyazawa. Soft margin keyframe comparison: Enhancing precision of fraud detection in retail surveillance. In *WACV*, 2011. 242

[12] G.-J. Poulisse, M.-F. Moens, T. Dekens, and K. Deschacht. News story segmentation in multiple modalities. *Multimedia Tools Appl. Journal*, 2010. 242

[13] F. Richter, S. Romberg, E. Horster, and R. Lienhart. Multimodal ranking for image search on community databases. In *MIR*, 2010. 243

[14] StopLift. http://www.stoplift.com. 241, 242

[15] H. Trinh, Q. Fan, S. Pankanti, P. Gabbur, J. Pan, and S. Miyazawa. Detecting human activities in retail surveillance using hierarchical finite state machine. In *ICASSP*, 2011. 241, 242, 244, 245, 246

[16] M. yu Chen and A. Hauptmann. Active learning in multiple modalities for semantic feature extraction from video. In *AAAI*, 2005. 242