Exploratory Analysis of Visual Search Data

David J. Stracuzzi(✉), Ann Speed, Austin Silva,
Michael Haass, and Derek Trumbo

Sandia National Laboratories, Albuquerque, NM 87185, USA
{djstrac,aespeed,aussilv,mjhaass,dtrumbo}@sandia.gov

Abstract. Visual search data describe people's performance on the common perceptual problem of identifying target objects in a complex scene. Technological advances in areas such as eye tracking now provide researchers with a wealth of data not previously available. The goal of this work is to support researchers in analyzing this complex and multimodal data and in developing new insights into visual search techniques. We discuss several methods drawn from the statistics and machine learning literature for integrating visual search data derived from multiple sources and performing exploratory data analysis. We ground our discussion in a specific task performed by officers at the Transportation Security Administration and consider the applicability, likely issues, and possible adaptations of several candidate analysis methods.

Keywords: Visual search · Eye tracking · Data analysis · Transportation Security Administration

1 Introduction

Visual search research considers the perceptual problem of actively scanning a scene to locate target objects embedded in a complex visual scene. Examples range from everyday activities such as finding a friend among a crowd to detail-oriented activities such as radiologists reading X-ray or MRI images. Methods by which people engage in visual search have been studied for several decades. However, recent advances in eye tracking systems enable collection of fine-grained data in visual search domains under realistic conditions. For example, eye tracking systems can now be embedded into operational environments that include multiple monitors and a variety of display manipulation tools. As a result, visual search studies now produce increasingly rich data in increasingly large volumes. These data may include the eye tracking observations, participant interactions with the environment (such as image manipulation software), performance results with respect to the given task, and the stimulus itself (such as images).

Such rich data requires increasingly sophisticated analysis techniques. Traditional statistical analysis remains an appropriate choice for characterizing variables such as overall performance, changes in performance over time, participant usage of available image manipulation tools, and so on. Importantly, tools for viewing and analyzing the eye tracking results have improved in step with hardware advances. For example, researchers can easily generate and compare heat

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 537–547, 2015.
DOI: 10.1007/978-3-319-20816-9_52

maps and temporally annotated gaze patterns among both individuals and selected groups. They can also identify specific regions of interest and analyze the participants' search with respect to those regions. The available tools support analysis of visual search patterns, differences in search patterns among selected populations, and visual salience of specific image features.

Nevertheless, important limitations remain for the analysis and modeling of visual search data. In particular, data from different sources, such as eye tracking and user software interactions, are typically analyzed separately. Methods for integrating all data sources into a single analysis that can provide deep insights into visual search processes remain elusive. In this paper, we discuss an ongoing effort to apply state of the art machine learning and statistical modeling methods to perform exploratory analysis of complex and multimodal visual search data. We hypothesize that querying and comparing the resulting models will provide deeper insights into the methods and performance of the study participants than otherwise possible. We ground the discussion of our proposed approach with an ongoing study of Transportation Security Administration (TSA) officers who are tasked with interrogating X-ray images of carry-on luggage for threats at airport security checkpoints. Our discussion then focuses on the proposed analysis algorithms, their justification, and anticipated adaptations as the data for this study is still being collected as of this writing.

2 The TSA Task and Data Collection

One of the primary features of airport security is the passenger checkpoint at which both the passengers and their carry-on luggage are checked for items that may pose a threat. The task of reading the X-ray imagery of the carry-on luggage is a perfect example of a visual search task. The Transportation Safety Officers (TSOs) must scan the imagery associated with each bag to determine whether it contains any potential threats.

The TSO's task is difficult for several reasons. For example, the X-ray image converts a three-dimensional collection of objects (the bag and its contents) into a two-dimensional image, thereby superimposing many objects onto one another and obscuring the defining features of many objects. Also challenging is that the class of threat objects is fairly large and varied (consider all of the different items that are not allowed in carry-on luggage). Finally, TSOs must perform their analysis under time pressure, as checkpoint throughput is a significant concern.

Given these issues, the goal of studying the visual search performance of TSOs is to develop insight into how they interrogate the bags, how they identify threat objects, and what changes might be made to the X-ray display system to improve overall performance with respect to measures such as throughput, probability of detection, and probability of false alarms. The exploratory analysis methods that we discuss below focus on identifying features that are predictive of conditions of interest, and identifying specific, testable hypotheses. Sample questions of interest might consider differences among common and anomalous

image interrogation patterns, differences between high- and low-performing analysts, differences between easy and hard to interrogate objects, and predictors of impending analyst errors.

Data collection focuses on TSOs reading imagery in a high-fidelity, simulated setting. Each TSO reads a sequence of X-ray images for two hours or a maximum of 1,000 luggage items. The luggage contents generally mirror the stream of commerce at a typical airport security checkpoint as determined by the TSA, except that 10 percent of the items include simulated threats. Limited ground truth information is available for each luggage item, including the presence and location of any threat items, the presence of prohibited items such as toothpaste tubes and full water bottles, and the presence of other items of interest such as laptops. Each luggage item includes a set of 31 image products that the TSOs can toggle on and off. Each image product is derived from the X-ray scan and highlights different aspects of the scanned item. These images are identical to those used at the checkpoint and represent the most commonly used subset of all available image products.

We used a simulator to collect data because the scanners used at the checkpoint are closed, proprietary systems. It provides the same viewing environment and most of the same controls as the checkpoint scanners. TSOs view the imagery on two, side-by-side monitors showing a top and side view of the luggage items. The images scroll right-to-left across the monitors until either stopping automatically, or the officer stops them with a key press (depending on which operating procedure is in force). An important side effect of the scrolling is that a signification portion of the eye tracking occurs with respect to a moving image. See Speed et al. [13] for an in depth discussion of the stimulus creation, experimental design, and data collection procedures.

The collected data are illustrated in Fig. 1 includes gaze locations and durations from an integrated SmartEye Pro 6.0 three camera eye tracking system. The data also includes user keystroke information such as image product selections, and TSO judgments such as threat versus no threat for each luggage item. All of the data sources include time stamps. The 31 X-ray image products presented to the TSOs are also available to the analysis algorithms.

3 Data Preparation

Before we can apply the selected statistical and machine learning algorithms to the data, we first need to encode the data in an appropriate format. Given the sequential nature of visual search in general, and the eye tracking data in particular, we elect to represent the data as a time series and apply analysis methods that assume a temporal component. Preparing the data entails several steps. For example, the eye trackers produce data at a high sampling rate that requires significant aggregation before it can be combined with the viewed imagery. Likewise, the raw image pixel values may be less useful during analysis than aggregate image properties such as edges or regions of uniform color.

Also important will be efforts to establish relationships between the different data streams. For example, associations of gaze locations with properties of the

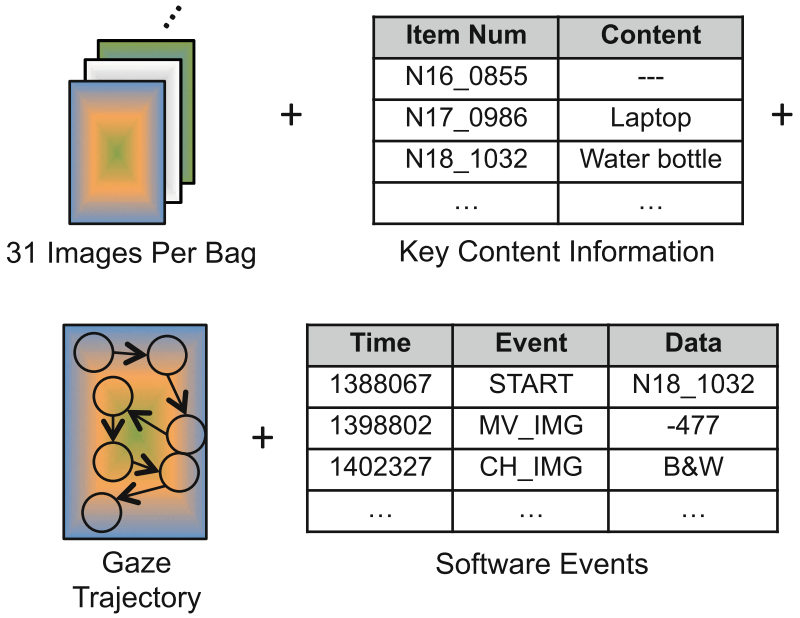

Fig. 1. Illustration of the stimulus imagery, associated ground truth, eye tracking and software event data collected by the checkpoint simulator.

underlying image may together relate to the TSOs decision to switch image products. Identifying these associations, known as feature extraction, plays a critical role in data analysis. In the remainder of this section, we discuss the steps necessary to transform the raw data into a form that facilitates the learning, modeling, and generalization steps that follow.

3.1 Preprocessing of Eye Tracking Data

The SmartEye eye tracking system records raw gaze and head orientation samples at a rate of 60 Hz. These samples are time-synchronized with the stimulus display and user actions (such as keyboard events) using the MAPPS[1] analysis software package. The MAPPS algorithms are used to calculate fixation points and durations from the eye tracking data stream.

For the TSA data, we set the *fixation minimum duration* parameter to 0.13 seconds and the *size of point cloud* parameter to 71 pixels. Importantly, the point cloud parameter controls the trade-off between fixation size and noise tolerance. Larger values group more gaze locations into a single fixation (possibly combining multiple fixations) while smaller values provide a more fine-grained separation of fixations at the expense of adding noise to the preprocessing results.

Given the display resolution, display size, and typical subject viewing distance (66 cm), these settings define the maximum angular velocity for fixations

[1] http://www.eyesdx.com/products/mapps/.

to be 15.25 degrees/second, which is consistent with published values [3]. As an additional sanity check, we also compared the maximum angular velocity for fixations to the angular velocity (from the participant's viewing position) of the image as it scrolls across the screen. We found that the velocity of the scrolling images was three orders of magnitude slower than that allowable within a fixation. This implies that our preprocessing methods should recognize any visual tracking of the scrolling image as a single fixation and not a sequence.

3.2 Feature Extraction

In the context of data analysis, features are variables derived from the raw data that are more informative and discriminating than the constituent variables. The objective of feature extraction is to simultaneously reduce the size of a data set while highlighting properties of the data that will improve both the accuracy and structure of classifiers and probabilistic models. Feature extraction is therefore a process of both discovery and relevance evaluation. Exploring the space of candidate features requires consideration of a variety of methods and algorithms.

The visual search literature provides extensive discussion of possible image features. For example, Wolfe [15] discusses asymmetry in visual search, which is the notion that detecting the absence of a feature is much more difficult than detecting its presence. This implies that having variables that indicate the absence of a pattern in the imagery may be of value. Itti [7] has conducted extensive research on visual salience, including algorithms for computing heat maps that indicate the bottom-up salience of imagery.

The computer vision literature focuses on identifying objects and relationships in images, irrespective of how the human visual system processes the imagery. Yu and Ng's tutorial [16] introduces a variety of established methods for identifying features such as object boundaries and classes, and for estimating the relative contributions of all features. Recent work on statistical relational learning (see Getoor and Taskar [5] for an extensive overview) is particularly relevant because it focuses on modeling relationships among objects and events. Nowozin and Lampert [9] provide an extensive review of how such methods have been applied to a variety of problems in image processing.

A third class of relevant features derives from the literature on trajectory analysis, of which the eye tracking data sequences are an example. For example, recent work by Raschke et al. [11] considers how to identify and compare scan path structures. More generally, Rintoul and Wilson [12] discuss domain agnostic methods for extracting geometric features from trajectories which they then use to characterize and compare them. The approach has shown success both as a similarity metric among trajectories and as an outlier detector.

3.3 Representing Visual Search Data as a Time Series

The goal of this work, which is to develop deep insights into visual search processes through exploratory data analysis, requires that we identify features that relate the different data streams to each other. Thus, we need to extend the

above image and trajectory analysis methods to identify relationships between the trajectories and the underlying image content. Designing a data representation that facilitates this type of multi-source analysis is therefore critical. For the visual search data, a time series may be the most appropriate choice. This entails representing all of the data as a sequence of observations based on the features identified in the previous step.

The primary issue in constructing the time series relates to encoding features of the static stimulus images into a sequential format. In practice, this requires a two step process. First, we process the images separately from the eye trajectories to identify edges, regions of uniform color, or other identifiable object (laptops, for example). Given these, we can then encode the specific content of the images associated with each fixation area. Thus, the time series represents the sequence of observations made by the participant, as opposed to the entire image.

Traditional time series analysis focuses on sequences of uniformly-spaced, typically univariate measurements [4]. In the context of eye tracking applications, the measurements correspond to both eye events (fixations in particular) and user interactions with the software (such as key presses), which are neither uniformly spaced nor of uniform duration. This raises several issues related to the time series representation, such as whether to explicitly represent event time and duration, or interpolate between individual eye events. The former approach is relevant in the context of trajectory analysis, while the latter may be sufficient for relating fixation locations to features of the underlying image.

Encoding the data as a single time series also raises a variety of other questions. For example, calculating and encoding features of the image content under the participant's gaze becomes tricky when the image moves or the participant changes image products during a fixation. Likewise, the data contains a variety of non-events, such as when the participant looks away from the monitors during their analysis. There is also the question of whether saccades provide any valuable information beyond the fixations. In practice all of these are empirical questions that we will answer as a part of the exploratory analysis. Nevertheless, these questions illustrate the scope of research remaining to be done with respect to visual search data analysis.

4 Visual Search Data Analysis

Having constructed the time series data representation that will support the remainder of the proposed analyses, we now consider specific analytic methods. Given the complexity of the visual search data, the exploratory data analysis process will necessarily be both incremental and iterative. The process is incremental in the sense that identifying and extracting useful patterns and features of the data may require sequential application of several different algorithms and techniques. The process is also iterative in that both individual algorithms and subsequences of them may need to be revisited to achieve the desired results, particularly in light of results achieved later in the process.

Figure 2 summarizes the analysis process (single line) and data flow (double line). The numbers in the figure identify the sections of the paper that provide

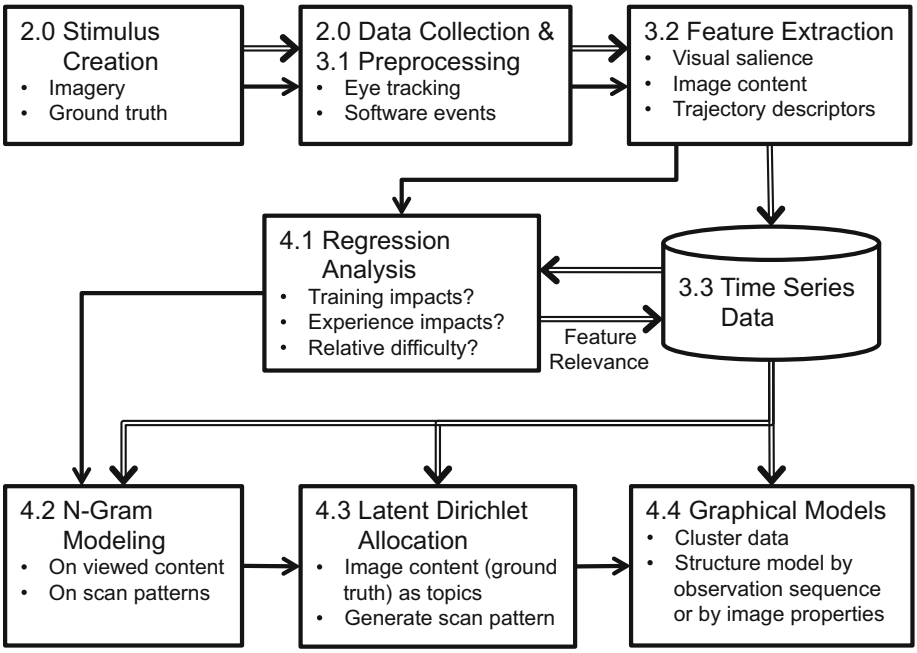

Fig. 2. Process flow (single line) and data flow (double line) for the envisioned visual search data analysis. Numbers indicate the sections of the paper in which we discuss the details of each step.

details for each step. Many of the analytic steps may identify new features that can be added into the time series description of the data. We highlight some of these opportunities in the discussion below. Finally, note that for the remainder of the paper, we use the word *trial* to indicate the interrogation of and decision for a single luggage item by a single TSO.

4.1 Bayesian Logistic Regression

Our first pass at analysis will gain domain insight by modeling questions of interest and by determining feature relevancies based on their contributions to models. Although the models themselves may prove useful, we are most interested in understanding the extent to which various features predict classes of interest, such as participant experience level or the difficulty of a given luggage item. Logistic regression predicts the probability that a trial belongs to a given class by optimizing the distribution over the parameters for a linear combination of predictor variables (features). Though similar in form to traditional regression analysis, the Bayesian approach offers several benefits such as a reduced chance of overfitting and information about the uncertainty associated with both parameter values and predictions [1]. In many cases, this uncertainty can say more about the predictiveness of a feature than the feature's weight.

One expected challenge in modeling the data concerns feature selection. Specifically, the number of features created in earlier steps will be large, as will the number of model parameters. This makes both estimating the parameter values and interpreting the contribution of the associated variable more difficult. Down selecting from among the set of available features, will therefore play an important role in modeling the data. See Guyon and Elisseeff [6] for an extensive review of the topic.

A second challenge relates to the form of logistic regression models. Strictly speaking, logistic regression ignores the temporal relationship between the observations, treating the set of observations associated with a trial as vector of variable values. This implies that a regression model may not capture the relationships among the observations well enough to reveal the predictiveness of individual features. As an alternative, we can consider linear-chain conditional random fields (CRFs), which are a natural extension of logistic regression to sequential observations and alleviate the problem at the expense of added model and computational complexity. Sutton and McCallum [14] provide an accessible description of the relationship between logistic regression and linear-chain CRFs. For this work, we will start with the simplest model, and then expand to more complex methods depending on the results.

4.2 N-Gram Modeling

The remaining three analysis methods draw substantial inspiration from the text analysis literature. We hypothesize that, although text analysis is more constrained than image analysis — English must be read left-to-right and top-to-bottom while images features can be considered in any order — a number of similarities remain. For example, local structures and relationships play a major role in determining meaning. N-gram models exploit combinations of temporally adjacent features to predict larger patterns, such as "political news article" or "luggage containing water bottle".

N-gram models estimate the probability of the next observation in a sequence based on the previous $n-1$ observations by counting the number of times the same combination appears in training data. For text analysis, this corresponds to the probability of observing a word given a small number of words that precede it in the sentence. By extension, we can use n-grams to estimate the probability of observing certain image features given the previous few observations (fixations).

This provides a simple mechanism for finding local structure and relationships in sequences, and then using them to evaluate much longer sequences. For example, the presence of certain fixation sequences may predict whether a TSO will mark a bag as a threat. Discovery of such sequences would provide substantial insight into what image features trigger officers to recognize threats, whether correctly or incorrectly. This information can then be used to improve the algorithms that process the raw X-ray data into images.

Extending n-grams to the visual search domains requires a few key extensions to the method. For example, no two fixation descriptors (points in the time series) will be exactly the same. In order to calculate the probability of observing

a given sequence of observations, we will need to develop a notion of similarity. Similarly, only a few of the possible fixation sequences will be observed in the collected data, so we need a way to account for the sequences that we didn't observe (or make the unreasonable assumption that they never occur). Jurafsky and Manning [8] provide an accessible introduction to this issue and possible solutions. Also consider that, like language analysis, image analysis may have long-distance dependencies. Although in practice local dependencies are sufficient to produce useful models, we still need determine a reasonable value for n in the context of visual search. Finally, note that the n-gram features may be added into the time series representation for use by other models.

4.3 Latent Dirichlet Allocation

Continuing with the text-processing analogy, latent Dirichlet allocation (LDA) treats each example as a mixture of *topics*. Traditionally used to classify documents, each topic is considered responsible for generating portions of the document text. Thus, by identifying the topics associated with a document, we can draw conclusions about the document's content. Blei [2] provides an extensive introduction to LDA and its relatives.

We can apply the same method to visual search data by viewing the sequence of fixations with associated image features as the document text. The goal then is to cluster the fixations from the trials, such that the clusters represent the key features of the image and the TSO's analysis of it. When finished, we can identify the set of topics that governs any observation sequence based on the available ground truth data. Given that we know the contents of the image (such as laptop, water bottle, or threat item) and the outcome of the analysis (false positive, false negative, and so on), we can identify descriptive features for each topic and use them for insight into the visual search process. For example, we may find that experienced TSOs consistently focus on a different set of image features than novices.

As with the n-gram models, LDA will rely heavily on defining a notion of similarity between two fixation descriptors. Likewise, we also need to account for the many possible fixation descriptors that will not appear in the collected data. Although not strictly required, providing meaningful names for the identified topics may require a significant visualization effort. Topic names ultimately derive from the content of the fixation clusters, which will not be easy to interpret. Finally, LDA assumes that the ordering of the observations does not matter (equivalent to bag-of-words in text analysis). More sophisticated variants of LDA can take this ordering into account, but as with the regression analysis, we will begin with the simplest models first.

4.4 Discriminative Graphical Models

Extending now beyond models designed for text classification, discriminative graphical models represent a generalization of the linear-chain CRFs noted

above [14]. The method treats the trials as a sequence of interdependent variables. The structure of the graph indicates how the individual observations depend on each other, and many structures are possible. In particular, features of the underlying imagery that are independent of the TSO's fixations can be incorporated into the model. This allows the model to incorporate more context from the underlying imagery than was possible with the previously described methods.

Applying graphical models to the visual search domain will require finding an appropriate division of the trials. The TSOs can search an image in a variety of ways, and combining data from trials that differ wildly in interrogation style will likely have a dissonant effect on parameter estimation. The same may hold for images that do not share similar content or visual properties. The ground truth associated with the images and trials, such as luggage content, analyst performance, and analyst experience level, can inform these, but may not be sufficient. An alternative is to first cluster the trials (or images), and then model each cluster separately, as demonstrated by Oates et al. [10].

We can use the cluster-based graphical models to derive insights into visual search processes in multiple ways. One option is to consider the set of trials contained in each cluster and identify similarities. This is similar in spirit to identifying topics in LDA and serves to highlight the distinguishing properties of each trial. A second option is to evaluate trials held separate from those used during clustering for goodness of fit to each cluster. This type of trial classifier can also help to identify the key properties of trials in each cluster.

5 Summary and Conclusion

Insight into visual search processes can impact a variety of research and development areas. Just in the context of TSA domain considered in this paper, deeper insights into how the checkpoint officers read the imagery to recognize and identify threats could impact how the X-ray data is collected, how it is processed, and how it is displayed to the user. In this paper, we outlined an extensive process for collecting and analyzing visual search data from the TSA domain. Our proposed analysis focuses on identifying a broad variety of features in the data and on integrating the different data sources into a single analysis.

Integrating sources such as eye tracking, software interactions, stimulus imagery, and its associated ground truth information makes the analysis process very complex, as illustrated by the large number of steps and open issues described above. Nevertheless, we assert that combining all of the data as described offers the best opportunity to answer important questions such as what predicts how difficult a luggage item will be to assess, or what might indicate an impending TSO error. Progress on these questions will help to improve security at airport checkpoints.

Acknowledgements. The authors thank David Robinson and Travis Bauer for helpful discussions on the data analysis methodologies. This research was funded in part or

whole by an interagency agreement between the Transportation Security Administration and Sandia. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. Bishop, C.M., Tipping, M.E.: Bayesian regression and classification. In: Suykens, J., Horvath, I., Basu, S., Micchelli, C., Vandewalle, J. (eds.) Advances in Learning Theory: Methods, Models and Applications. NATO Science Series III: Computer and Systems Sciences, vol. 190. IOS Press, Amsterdam (2003)
2. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
3. Erkelens, C.J., Sloot, O.B.: Initial directions and landing positions of binocular saccades. Vis. Res. **35**(23–24), 3297–3303 (1995)
4. Fu, T.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**, 164–181 (2011)
5. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)
6. Guyon, I., Elisseef, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
7. Itti, L.: Visual salience. Scholarpedia **2**(9), 3327 (2007). http://www.scholarpedia.org/article/Visual_salience
8. Jurafsky, D., Manning, C.: Natural language processing, January 2012. https://www.coursera.org/course/nlp
9. Nowozin, S., Lampert, C.H.: Structured learning and prediction in computer vision. Found. Trends Comput. Graph. Vis. **6**(3–4), 185–365 (2010)
10. Oates, T., Firoiu, L., Cohen, P.R.: Clustering time series with hidden markov models and dynamic time warping. In: Proceedings of the IJCAI 1999 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning, pp. 17–21 (1999)
11. Raschke, M., Herr, D., Blascheck, T., Ertl, T., Burch, M., Willmann, S., Schrauf, M.: A visual approach for scan path comparison. In: 2014 Symposium on Eye Tracking Research and Applications. ACM, Safety Harbor (2014)
12. Rintoul, M.D., Wilson, A.T.: Trajectory analysis via a geometric feature space approach (2014) (unpublished, in preparation)
13. Speed, A.E., Silva, A., Trumbo, D., Stracuzzi, D.J., Warrender, C., Trumbo, M., Divis, K.: Determining the optimal time on X-ray analysis for transportation security officers. In: Proceedings of the 9th International Conference on Augmented Cognition. Springer, Los Angeles (2015)
14. Sutton, C., McCallum, A.K.: An introduction to conditional random fields. Found. Trends Mach. Learn. **4**(4), 267–373 (2011)
15. Wolfe, J.M.: Asymmetries in visual search: an introduction. Percept. Psychophys. **63**(3), 381–389 (2001)
16. Yu, K., Ng, A.: ECCV-2010 tutorial: feature learning for image classification, September 2010. http://ufldl.stanford.edu/eccv10-tutorial/