CuZero: Embracing the Frontier of Interactive Visual Search for Informed Users

Eric Zavesky Shih-Fu Chang
Dept. of Electrical Engineering
Columbia University
1312 S.W.Mudd, 500 W. 120th St, New York, NY 10027
{emz,sfchang}@ee.columbia.edu

ABSTRACT

Users of most visual search systems suffer from two primary sources of frustration. Before a search over this data is executed, a query must be formulated. Traditional keyword search systems offer only passive, non-interactive input, which frustrates users that are unfamiliar with the search topic or the target data set. Additionally, after query formulation, result inspection is often relegated to a tiresome, linear inspection of results bound to a single query. In this paper, we reexamine the struggles that users encounter with existing paradigms and present a solution prototype system, CuZero. CuZero employs a unique query process that allows zero-latency query formulation for an informed human search. Relevant visual concepts discovered from various strategies (lexical mapping, statistical occurrence, and search result mining) are automatically recommended in real time after users enter each single word. CuZero also introduces a new intuitive visualization system that allows users to navigate seamlessly in the concept space at-will and simultaneously while displaying the results corresponding to arbitrary permutations of multiple concepts in real time. The result is the creation of an environment that allows the user to rapidly scan many different query permutations without additional query reformulation. Such a navigation system also allows efficient exploration of different types of queries, such as semantic concepts, visual descriptors, and example content, all within one navigation session as opposed to the repetitive trials used in conventional systems.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces and Interaction Styles

General Terms

Algorithms, Design, Experimentation, Interaction styles, Performance

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

Keywords

CuZero, Interactive Video Search & Retrieval, Query Formulation, Result Visualization, Semantic Concepts

1. INTRODUCTION

As the amount of available digital media (image and video) grows exponentially, an inability to efficiently search this media content becomes more apparent. In the past, research has focused on the extraction of features at either the low level or semantic level to aide in indexing and retrieval. However, efficient ways to interactively search (or query) the large media databases still lack satisfactory solutions and remain significant challenges.

Exploration of a large collection of image and video data is a non-trivial task. When any user approaches a new search task, formulating a query (search criterion) can be quite difficult. Most modern search systems provide the ability to search with textual input, which has been studied by the information retrieval community at large, but several problems still remain when searching visual content, as illustrated in fig. 1. First, the choice of correct

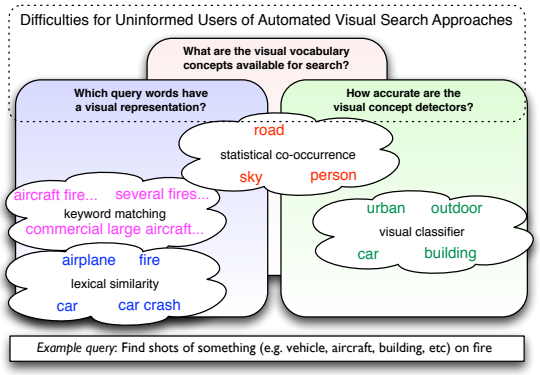


Figure 1: Typical automatic search methods and their potential difficulties for users unfamiliar with the data to search or methods available.

query words can significantly effect the output of a video search system. Generally, the user lacks information about which words would best match the content he or she is looking for (i.e. visual descriptions of people, objects, scenes, etc.). Second, if using automatically detected visual concepts derived from low level image features and trained with a labeled set of data as in [2],[17], non-expert users lack knowledge about the concept vocabulary and accuracy of the concept detectors. While techniques have been proposed for fully automated approaches to combining descriptors of multiple modalities (text, low level features, and concepts) [9], [10], [20] these solutions are not well-suited to be directly used in

an interactive search system. In general, these difficulties experienced by expert and novice due to the aforementioned reasons are symptomatic of an *uninformed user*.

Once search results are returned, the user often struggles in navigating through the large set of images or video content and a typical interface showing a linear list of thumbnails is not sufficient. There is little information to help users understand why the given set of results, how each returned image/video is related to the concepts chosen in the query, and most importantly, how to efficiently adjust the strategies (fast skimming v.s. in-depth browsing) in exploring the result set. Such difficulties arise from the fundamental problem of *disconnection between result navigation interfaces and the query criteria*. Once the query is evaluated, the query criteria are lost and the user is presented with a set of results without showing their relations with the concepts he/she has chosen. Some impressive visualization techniques (such as [3] and [18]) have been proposed to assist users in fast browsing and exploration of result sets. The system reported in [11] has also shown impressive performance in refining search results by relevance feedback and active learning. However, the user's frustration in relating search results to search criteria and inability to dynamically adjust the influence of each query criterion remain unresolved.

Finally, after the initial query and result exploration, users often need to modify their current query or develop a new query to find the best results. In this *blind exploration* process, the user may fall back to searching in a trial and error fashion, as shown in fig. 2. To preempt this behavior, it would be most advantageous for the system to utilize the patterns and behaviors the user has shown in exploring the previous set of results and then judiciously recommend new strategies to aide query formulation.

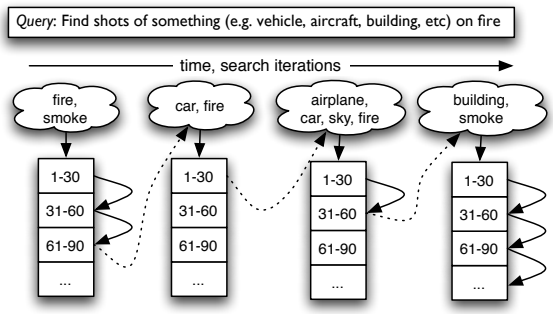


Figure 2: A typical interaction with a search system where the user resorts to trial and error query formulation to explore different parts of the data set.

In this paper, solutions for two difficulties mentioned above are proposed: guided query formulation to help uniformed users and dynamic query adaptation to break free of the rigid visualization interfaces. First, to assist in query formulation, the proposed system truly engages the user and provides new concept-based suggestions as the textual query is entered. With human participation, new concepts proposed by many automatic methods is simultaneously presented and user can intelligently formulate a better query. Second, for a visualization alternative, many different minor permutations (slight variations of concept weights) are simultaneously evaluated and presented. Using two intuitive displays, the user can quickly inspect a high number of query permutations in one panel and simultaneously evaluate the results of each permutation in a second panel. These permutations can be inspected with no precomputed order constraints, a clear advantage over traditional linear page-based browsing. Additionally, all results presented to the user in the instant update panel are guaranteed to be unique allowing deeper and faster exploration of large data sets. Implementations for each method are proposed that create a fast, efficient, and fully engaging interactive search environment.

2. MODERN VIDEO SEARCH

Modern video search systems attempt to tackle the problems described above but often fall short in achieving a coherent solution for all problems. Interactive systems often evolve from previously created fully-automatic systems, which require no user input. Unfortunately, originating from an automatic system can cause problems like inaccurate or non-specific mapping of keywords to a search target, a quick loss of query relevance with deep result inspection, and result lists that are static and thereby restrict fast inspection of alternative query strategies.

2.1 Automatic Keyword to Query Mapping

The input mechanism for most search systems is largely non-interactive. When asked to input a textual query, the user may struggle because of his or her unfamiliarity with the target data set. Additionally, if the user has no specific query topic (i.e. *Find shots of something (e.g. vehicle, aircraft, building, etc.) on fire*) and instead merely has a rough idea to search for, it can be challenging to know what the first query should be. One solution to this problem is to provide users with tools that present information relevant to the user's query by utilizing various approaches for query expansion from words given by a user and expand them into a large set of relevant concepts.

Traditional text-based or synonym-based mapping to a concept vocabulary may not be satisfactory if there is no directly relevant concept. For example, in a query for specific named entity, there may or may not be a specific concept mapping for that person. While most techniques correctly map famous political figures to concepts like *politics*, *speakers*, or *heads of state*, buildings or locations will not be expected to have similarly suitable matches. Work has been done to map different types of queries into different query classes [9],[10],[20] using lexical expansion, direct keyword matching, and automated part of speech tagging. Strategies for concept mapping will then be optimally adapted for each specific query class, e.g., named entity v.s. scenes and objects. However, when applied to more diverse topics, these methods achieved fewer benefits, indicating that either more training data was necessary or that this method may not scale well. Methods were also developed for mining relevant concepts based solely on the scores of visual classifiers. Concepts were ranked by decreasing mutual information in [8] and the number of visual results scoring above a pre-computed threshold in [22] but in both methods strong relationships between concepts were discovered from the current set of results. Often a user would overlook these concepts because of a perceived tangential relationship (i.e. *bleachers and benches* when searching for *basketball*), but the data-driven nature of these suggestions offers a valuable connection between the user and system. Unfortunately, the suggested concepts provided by this approach depend entirely on the quality of results returned from an initial query; without a set of prior results, there are no available concept scores for visual expansion.

2.2 Query Expansion with External Corpora

Query expansion is not limited to a fixed vocabulary of concepts. With the availability of many large corpora on the internet, some works mine these corpora for keyword expansions from a user's query text. Originally formulated as part of a completely automatic search system, [5] developed a method that would propose new, highly relevant named entities and events for a text query. In this

method, a corpus of external documents (i.e. online news publications) not included in the user's target data set are collected for the same period of time as the user's target data. Using a histogram-based method, the most frequent names of people, locations, and events that exist in both the external and target sets are discovered and proposed as new query keywords for the user.

While an innovative use of expansive resources, keyword expansion based on external corpora fails in two general cases: the user's search target is not a named entity or the expansion corpora is not topical with respect to the target. Named entities are distinct and powerful search keywords for text-based queries. However, without a specific named entity or with incorrectly expanded entities result quality can be adversely effected, like an expansion that maps *soccer goal posts* to *World Cup star Pele*. Similarly, a non-topical expansion can occur if the expansion corpora was collected at a period in time different than the target video, like an expansion using news articles from 2008 for video captured in the 1970's.

2.3 The Result Dead-end

Another limitation enforced by traditional interactive systems is the order of result inspection. Most implementations rely on a predefined linear order to help the user traverse through a set of results and require that the user modify certain query parameters to change the course of inspection. This fixed list definition can frustrate users because items deeper in a result list, are generally less relevant than those that precede it, which effectively creates a result "dead-end".

Providing multiple navigation alternatives, [18] actively updates several lists based on textual, visual, temporal, and mid-level concept similarity to the most recently inspected result. This novel approach allows users to break away from the static results of any one query, but it simultaneously diffuses the impact of the user's query formulation. For example, if a user begins a query looking for *something on fire*, it would be possible to find related images by using temporal or textual clues. However, as a user switches between the different lists, the set of results that are actively being inspected may drift further away from the original query topic such that the user no longer understands the results at all; the lists allow a break from the original query, but simultaneously prevent the user from 'anchoring' to a concrete and precise query target.

One way to avoid being stuck in the result dead-end while going deeper in the result list is to constantly revise the results by reranking the list at hand or to quickly issue a new query to retrieve more relevant results. The reranked list or new query may be optimized based on feedback that users have provided through interactive labeling to indicate images that are relevant or irrelevant. Relevance feedback, as described in an early work [14] and used in some recent systems [11], can be further improved by "actively" selecting the data for subsequent user inspection in order for the system to collect the most beneficial training samples (compared to random passive selection of samples) in learning improved semantic models [19]. Despite the potential benefits of such approaches, users may often become "puzzled" or "uninformed" about the specific criteria being used in any given point of querying or reranking - leading to a equivalent "blindspot" that will likely disengage users from continued exploration.

2.4 Display of Results

After a query has been formulated, traditional systems fail to provide a mechanism for dynamically refining and personalizing the query, or adding breadth to the query. As shown in fig. 2, results from a single query are inherently static in nature. Although various visualization and filtering tools can be applied to explore the given result set, the specific query cannot be easily changed.

One branch of research on result displays focuses on the projection of a result space according to a well-defined distance metric. Works in this branch seek to preserve distances between images with respect to both the global result space and the local neighborhood of results. After selecting an ideal distance metric, authors of [12] derive a new layout using a graph-based projection to explore image similarity in a multi-resolution fashion (i.e. global, neighborhood, and local relationships). While this approach can quickly expose users to a very dynamic set of result images, the user may not understand how the projected space relates to his or her initial query and become frustrated.

In one of our prior works [21], *Visual Islands* utilized a grid-based display (most familiar to users of modern visual search systems) to organize images according to their dominant feature relationships within a small subset of results without losing the relevance ranking information from the original list. This work also introduced a fast method for context-based organization that allowed for non-linear navigation. However, it does not explicitly fill the gap between the two disconnected processes: query criterion specification and result inspection.

3. APPROACHING THE FRONTIER

This work describes CuZero, an interactive system that embraces new and fast alternatives for two areas of interactive systems: query formulation and dynamic result exploration. This work also describes ways to adapt existing search techniques to be very engaging, highly responsive, and dramatically decrease the time needed to explore a large set of image or video data. Fig. 3 demonstrates an implementation of this system, CuZero, that provides two innovative interfaces: an informed concept query formulation panel and a real-time result exploration panel.

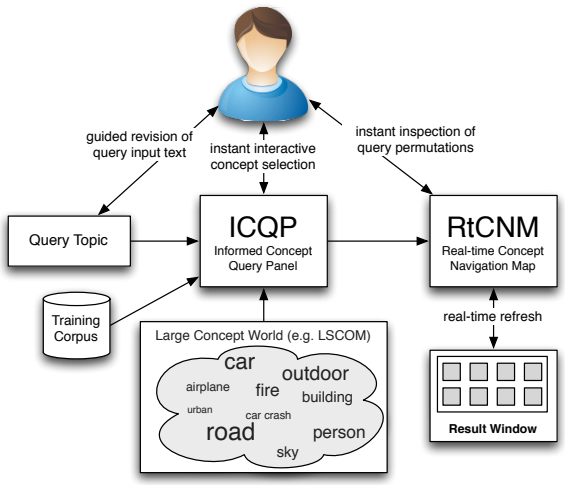


Figure 3: CuZero's components and interactive process pipeline.

3.1 Instant Assistance in Query Formulation

CuZero's informed concept query panel (ICQP) efficiently transitions the purely automatic methods mentioned in sec. 2.1 into an interactive setting with two refinements: the usefulness of each concept suggestion should be provided and suggestions should be updated as instantly and frequently as possible to reflect changes in the user's entered query text. To measure usefulness of a suggested concept, frequency and performance of the concept over a training set can be analyzed. For example, a suggested concept will be shown with a larger font if it has a high occurrence frequency

and high accuracy in automatic detection. Demonstrated in fig. 4, the magnitude of this measurement can be conveyed to the user by increasing or decreasing the font size of the suggested concept in the display presented to the user. High-frequency suggestion updating can be easily accomplished by monitoring user input (e.g. word completion or sojourn time).

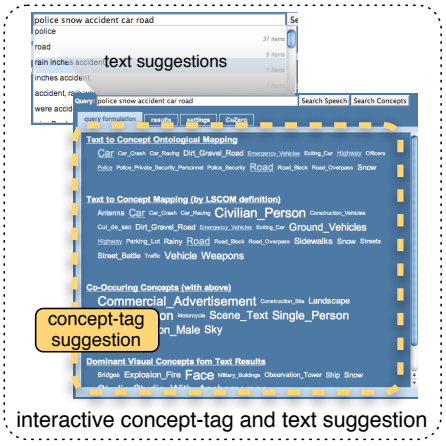


Figure 4: CuZero's instant concept suggestion panel dynamically updates during query formulation.

In all automatic methods reviewed, the user must enter query text, click 'search', then wait for the system to execute the query and review a set of results. In the proposed system, when a certain key-press is detected (e.g. the spacebar), the system automatically updates the currently displayed set of suggestions. Alternatively, timers measuring user inactivity could also be employed, but the use of timers to trigger an event may cause confusion or frustration if the user is unable to understand this behavior. The main benefit of CuZero's approach to suggestions is that CuZero presents instant information assistance through many suggestions to the user, which drives a guided and interactive query formulation instead of directly mapping the user's query into a specific set of automatically suggested concept-tags. Similar functions for automatic query text suggestion are also found in modern online services like automatic query suggestion in *Youtube*'s keyword search and *Firefox* version 3. Applying this idea in our system, an automatic and instant suggestion of visual concepts related to the typed query words is returned to the user.

3.1.1 Concept-based Query Relevance Model

CuZero, adopts a concept-based query relevance model. Before discussing several options for automatic concept suggestions, we first describe the linearly fused concept query model currently used in CuZero. Search systems return documents ranked by relevance to a user's query, as determined by a function, R. Whereas text search systems use a keyword k_q to compute relevance to a document d_k, CuZero, like visual search systems approximate visual relevance to image results with a set of high-level concept-tags $\{t_i\}$ as in eq. 1. The exact, automated selection of $\{t_i\}$ is a non-trivial problem (sec. 2.1), so instead of guessing at user intentions CuZero presents the results of many different mapping methods to bridge the semantic gap between visual concepts and keywords (eq. 2).

$$R(d_k; k_q) \simeq R_{vis}(d_k; \{t_i\}) \quad (1)$$

$$\{t_i\} = mappped\_concept\_tag(k_q) \quad (2)$$

$$R_{vis}(d_k; \{t_i\}) = \sum_{i=1}^{N} \omega_i R(d_k; t_i) \quad (3)$$

In CuZero, scores of the user-selected visual concept classifiers are combined with individual weights, ω_i, to derive a single relevance score for an entire document (eq. 3).

3.1.2 Automatic Recommendation Systems

Automatic concept recommendation allows the user to perform a preliminary inspection of the visual semantic space before executing his or her query. CuZero presents the textual and concept recommendations of several automatic systems as explained below, allowing the user to select and formulate a diverse concept-tag world. For all concept statistics and models, the Columbia374 is used, which was trained and evaluated on the TRECVID2005 data set [1], [16]. Note that CuZero's recommendation system offers a general framework and other concept suggestion techniques can be readily incorporated.

Text Completion with Speech Phrases: Text completion is a recommendation method that allows the user to automatically discover relevant speech phrases from the searched corpus with only a few keyword letters. Candidates for text completion are found from uni-, bi-, and tri-gram phrases found in speech transcripts and indexed in an easily searchable database. To improve responsiveness, text recommendations are not ranked by their relevance score at query time.

Text to Concept Ontological Mapping: Suggestions for direct ontological mapping are derived in two ways. The system first searches for exact textual matches of words found in the query text against the stemmed names of all available concepts [13]. Next, a boolean search (returns relevant if any word matches) against synonyms found in WordNet is executed [6]. For a single list, a logical union of both search result sets is performed.

Text to Concept Mapping (by LSCOM definition): Using a boolean search, the system searches stemmed concept definitions as provided by LSCOM [7]. This recommendation type can provide concepts that are only tangentially related like finding the concept *sports* from the LSCOM definition for *athlete*, "a person whose job it is to perform in a sport". The quality of these matches can vary dramatically because of definition specificity, so this set of suggestions is visually ranked with lower confidence than direct ontological matches.

Co-Occuring Concepts: Co-occuring concepts are those concepts that tend to appear together in a certain corpus. Using human-provided labels from the Columbia374 data over the TRECVID corpora, pair-wise mutual information scores between all concepts are computed. Using initial concepts from the above ontological mapping and definition expansions, additional concepts with high mutual information scores are suggested.

Dominant Visual Concepts from Text Results: Visual concept-tags are recommended from a text query that uses the initial N text search results to determine dominant semantic concepts [22]. In an off-line process before user interaction, concept detection score distributions for each concept in the Columbia374 are analyzed. The mean of image relevance scores in the top K detection results is computed and stored as a dominance threshold. From the top N text search results at query-time, concepts that have the most results with scores above their respective *dominance threshold* are recommended to the user. In this work, CuZero uses the settings $N = 40$ and $K = 4000$.

3.2 Parallel Multi-Query Exploration

While humans are capable of understanding how different parameters can effect a query, manipulating and visually representing these parameters can be burdensome. Prior work in [4] allowed a user to visually vary query parameters with a small set of pre-

defined setting pairs like *faces* and *only faces* or *no outdoor* and *only outdoor*. While this interaction frees the user from meticulously altering numerical values, non-expert users may not understand the underlying meaning of this control, as discussed in [15]. Ideally, users should be able to interact with a visual system to strategically arrange different concepts allowing a better exploration of how the results of each concept interact. Starting from a coarse placement of selected concepts in sec. 3.1 onto a 2D map, the system can then measure the contribution (or weight) of each concept-tag at many points (or cells), as shown in fig. 5 allowing the parallel exploration of multiple queries. CuZero employs

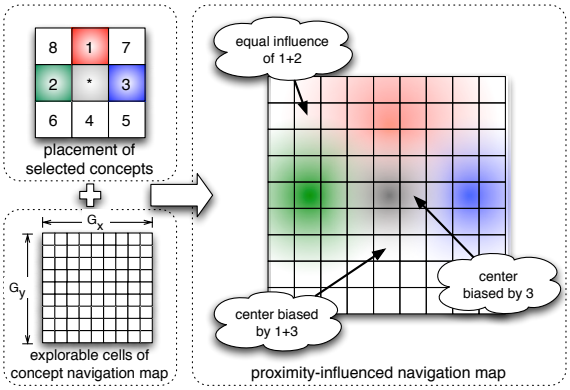


Figure 5: Combination of fixed concept placement and a quantized navigation map to produce proximity-sensitive exploration.

this visualization strategy with its *real-time concept navigation map* (RtCNM) so that the human is only required to understand the simple relationship that for every point *closer proximity* = *increasing influence*. With this naturally intuitive computation, the process of delicately mixing different parameters, regardless of their original dimensionality, is completely hidden from the user. In this work, the user is constrained to 8 concept-tag dimensions, but the proposed framework is generic enough to handle an arbitrary number. The addition of more simultaneous concepts decreases the intuitiveness of the RtCNM: differences between image results of cells in the RtCNM are less obvious and managing the 2D layout of many concepts is more difficult for the user.

3.2.1 Cell-based Relevance Weighting

Concept relevance weights are assigned to each cell of the RtCNM with a straight-forward Euclidean computation as illustrated in fig. 5 and explained in algorithm 1. Selected concepts $n = 1, ..., N$ are assigned to anchor cells a_n in the RtCNM space defined by the width G_x and height G_y. For every cell c_i in the RtCNM, the Euclidean distance to each concept anchor is computed as $d_{i,n}$. The minimum $d_{i,n}$ of each cell is also stored as ρ_i to better schedule subsequent cell processing operations. Next, each distance $d_{i,n}$ is non-linearly scaled into a weight $\omega_{i,n}$, using a Gaussian parameterized by σ. The tuning of the σ parameter provides high-precision control of how dramatically each concept anchor a_n influences its neighboring cells c_i. Next, all cell weights are normalized by the max weight for each concept n. With this formulation, concept weights are exponentially distributed around the anchor cell a_n for each concept. Finally, to guarantee that an anchor cell a_n contains the most relevant results for its concept n, zero all other concept weights for cell c_i if $\omega_{i,n}$ is exactly one. This weighting strategy provides precise permutations of many concept weights for each cell c_i in the RtCNM.

Algorithm 1 Cell Weighting for RtCNM

Input: Map cells $c_{i=1,...,I}$, concept anchor cells $a_{n=1,...,N}$, and Gaussian re-weighting factor σ.

1. Compute the Euclidean distance $d_{i,n}$ between cell c_i and each concept, a_n, according to its position $a_n^{x,y}$ and the cell's position $c_i^{x,y}$.
2. Derive cell priority from distance, $\rho_i = min(d_{i,n=1,...,N})$
3. Compute Gaussian weight for each cell i and concept n.

$$\omega_{i,n} = \frac{1}{\sigma\sqrt{2\pi}} exp\Big\{\frac{-(d_{i,n})^2}{2\sigma^2}\Big\} \quad (4)$$

4. Normalize all cell weights by max weight for each concept

$$\omega_{i,n} = \frac{\omega_{i,n}}{\max_i(\omega_{i,n})} \quad (5)$$

5. If weight $\omega_{i,n}$ is exactly one, concept n anchored at cell i, so zero all other weights for cell i, $\omega_{i,n\neq k} = 0$.

3.2.2 Planning Cell Results

After computing concept weights for the RtCNM, the system must assign or *plan* the most relevant result images for each cell. A straight-forward approach is to use weights $\omega_{i,n}$ from algorithm 1 and relevance scoring from eq. 3 where documents d_k are realized as image results. In CuZero, this computation is trivial because each concept n has a single concept detection list that is predetermined by visual classifiers. Thus, the planned results for each cell are simply a weighted combination of many concept result lists as indicated below.

$$R_{vis}(d_k; c_i) = \sum_{n=1}^{N} \omega_{i,n} R_{vis}(d_k; a_n) \quad (6)$$

3.2.3 Repeated Result Exposure

With the ability to explore many query permutations in parallel, a previously unexplored problem presents itself: how to best present the most unique image results to the user. Evaluating multiple queries with traditional weighted score fusion (eq. 6) produces many degenerate, overlapping result lists as depicted in fig. 6. CuZero therefore employs a special form of result planning to

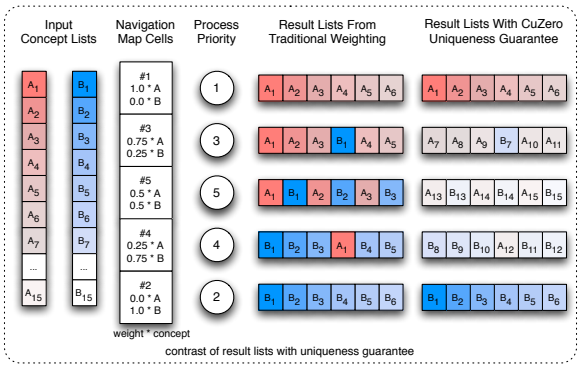


Figure 6: Using traditional methods to render results from multiple queries will lead to many repeated exposures. CuZero guarantees that the first visible page of results is unique across all cells in the concept navigation map.

guarantee that the first page of results (those immediately visible to the user for each cell) are unique for every cell in the naviga-

tion map. In the example in fig. 6, exploration depth is doubled if lists from concept A and B do not overlap. However, if an overlap of image results does occur (common with two highly related concepts like *outdoor* and *grass*), CuZero prioritizes cells with the smallest distance to user selected anchor concepts and numerical ties between minimum distances are broken by prioritizing concepts in the order that the user selected them. For example, cell contents closer to anchor $a_{n=1}$ will be planned before those for $a_{n=2}$. The suppression of repeat exposures improves the user experience in two ways: it encourages deeper exploration of concept results and creates a more engaging browsing experience because the user can instantly inspect a more diverse set of images.

3.3 Achieving Zero Latency Interaction

CuZero utilizes multimedia search components in a light-weight and dynamic way with several highly responsive, low-latency solutions. Fig. 7 illustrates the main processing pipeline in CuZero. The final pipeline stage primarily involves cell planning and the

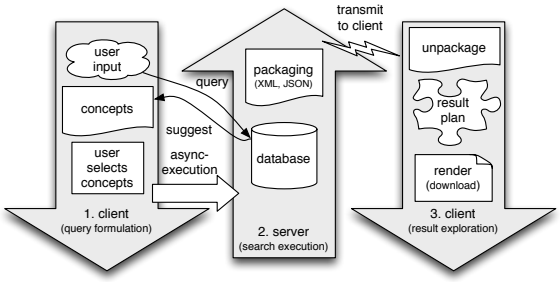


Figure 7: CuZero's asynchronous pipeline.

rendering (retrieval and graphical display) of result images. Unfortunately, the execution time of this stage is disproportionally long, consuming an average of 93% of the total time required for a new query. Fortunately, by utilizing a user's idle time and prioritizing rendering order, CuZero can reduce the user-perceived wait time by 81% to an average of 10 seconds for complete query execution.

3.3.1 Significant Navigation Map Changes

CuZero allows users to change the RtCNM at-will by adding, removing, or rearanging concepts in the query formulation panel. However, executing procedures to re-weight, re-plan, and re-render the RtCNM for every minor change would be prohibitively expensive. Instead, CuZero monitors changes to the RtCNM and computes new cell weights, but the planning and rendering stages are only executed for an individual cell c_i if its cumulative weight change, $\Delta\omega_i$

$$\Delta\omega_i = \sum_{n=1}^{N} |\omega_{i,n}^{new} - \omega_{i,n}| \quad (7)$$

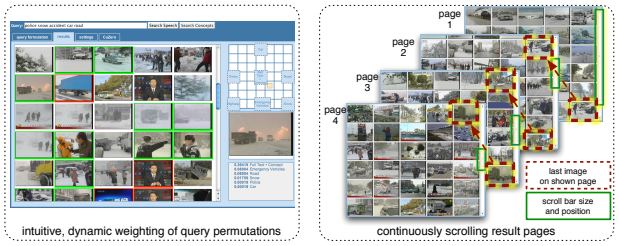
exceeds a pre-determined stability threshold. This fixed threshold is heuristically chosen to guarantee that the user's modification did not significantly modify the previously planned and rendered RtCNM layout.

3.3.2 Instancy in Interactions

To keep the user engaged in the process, modern search systems execute a single query in only 3-5 seconds and CuZero was developed with this lower-bound in mind. To explore many query permutations in parallel, CuZero fully utilizes the user's system resources; several asynchronous processes simultaneously run to render new image results and fully engage users during a search.

Priority-based Planning and Rendering: While computing the weights for each cell in the navigation map (algorithm 1), a priority score ρ_i for each cell is also computed. CuZero plans and renders the first page results for each cell according to its priority. An adjustable time-out counter prevents the user interface from being overloaded by the rendering process and if a user explores an un-rendered navigation cell, CuZero will attempt to immediately render that cell.

Intuitive Interactions: Common actions like labeling an image result positive or negative, reviewing its speech transcript, or inspecting the image result with a small animated icon are all click-free operations. Users can quickly inspect or label image results by simply hovering over an image anywhere within the interface allowing for fast interaction with the system without actively focusing on a specific keystroke or clicking within a small visual region. Additionally, the simultaneous availability of both the concept navigation map (fig. 8(a)) and the result panel (fig. 8(b)) allows the user to instantly switch between navigation for depth or breadth without breaking focus on the result exploration at hand.



(a) (b)

Figure 8: CuZero combines interfaces for both (a) a wide-breadth inspection of many query permutation cells and (b) a continuously scrolling window for exploration of one query cell at high-depth.

3.3.3 Caching While Idle

Wherever possible, tasks that require heavy computation are executed on the server (or back-end) whereas tasks that involve dynamic updates are executed on the client (or front-end). The following sections describe how the result exploration stage in fig. 7 is further optimized.

Cell Weights and Planning On-Click: In the query formulation panel, users are presented with concept suggestions, as described in sec. 3.1.2. When the user clicks (or selects) a concept, it is added to the navigation map and the client initiates a request for new weights and a result list for the newly selected concept. When the result list is received, the client continues to plan all cells whose result lists might have been changed due to the newly added concepts. The list of affected cells are estimated by using the weight change monitoring criterion in sec. 3.3.1. This technique is effective because users are usually idle for several seconds while carefully formulating their informed query instead of directly starting a new search.

Continuous Scroll Rendering: When the user wants to deeply explore a particular query permutation in the 2D map, CuZero employs a *continuously scrolling* interface that dynamically renders more results as the user inspects and depletes the current set of image results that have been cached, as in fig. 8(b). It is important to note that cells are already planned and only rendering of unseen results is completed by request. Every time the interface detects a scroll operation, CuZero verifies that there are at least two additional pages of unseen results rendered in the user's scroll buffer. This dynamic loading reduces demands on the system for

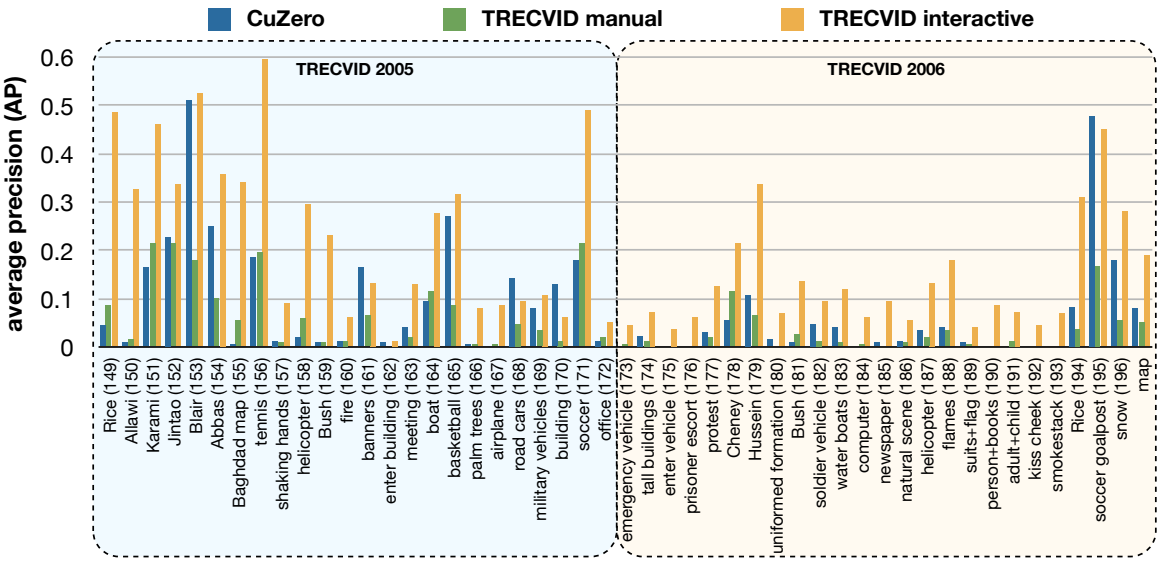


Figure 9: CuZero AP vs average AP achieved by all manual and interactive systems participating in TRECVID 2005 and TRECVID 2006 searches. CuZero results include no user-provided labels and were completed in one third of the alloted TRECVID search time. CuZero uses solely concept-based seach while other systems may include more diverse search tools.

results that the user is not interested in while guaranteeing the instant availability of results for a query permutation that the user is interested in.

4. EXPERIMENTS

CuZero offers two major innovations for interactive search: an instant-feedback, guided query formulation process and a concept navigation map for fast, at-will exploration of many parallel query formulations. We evaluated two experiments with CuZero to measure the benefits of "informed" concept suggestion and the benefit of the instant concept navigation and browser panels. In both experiments the user searched for results relevant to 48 topics evaluated in the TRECVID 2005 and 2006 search evaluation tasks [16]. TRECVID is an annual international evaluation for tasks like high-level feature detection and search over a large set of video data. The TRECVID data in these experiments is multi-lingual broadcast news. For these experiments, we used a reasonable classifier baseline created in [1], which contains 374 visual concept classifiers that were trained from LSCOM (Large Scale Concept Ontology for Multimedia) and publicly released as the *Columbia374*. In both experiments, the average precision (AP) is measured for each of the different search topics. AP is a common metric in information retrieval that measures precision at all depths of a search process and averages all measurements up to a given depth.

In the first experiment, the quality (as measured by AP) of concepts generated from three automatic methods (discussed in sec. 3.1.2) is compared to concepts chosen when using the "informed" query suggestion method in CuZero. Table 1 compares the mean average precision (MAP) scores at a depth of 100 across the 48 TRECVID search topics. For each topic, a concept mapping was performed from keywords entered by the user. The mapped concepts were ranked by their prior performance (over a training corpus) and then classifier scores from the top three concepts were averaged to produce a single result list. When using CuZero, concepts were presented to the user (see fig. 4) and the selected concepts were used to formulate a similar score-averaged result list. Comparing CuZero to traditional methods, it is evident that CuZero's user-guided concept suggestion yields the highest performance with a relative gain of at least 33% over all automatic methods for concept mapping, among which the lexical concept mapping method achieves the best performance.

Method	MAP @ 100	Gain
dominant visual concepts	0.02850	-30%
concept expansion by definition	0.03915	-5%
concept ontological mapping	0.04115	-
CuZero selected concepts	0.05480	33%

Table 1: Comparison of MAP scores (mean average precision) at depth 100 from automatic and CuZero concept suggestion methods. Only concepts (not keywords) were used to compute result lists for each TRECVID 2005 and 2006 query topic.

While high-quality concept suggestion is important, an evaluation was also needed for the new concept navigation and result browsing panels. Fig. 9 plots the performance of CuZero versus the average performance of other manual and interactive systems participating in TRECVID. In these experiments, users of CuZero were not asked to label positive or negative results. Instead, the system simulated construction of a result list based solely on the order of images that the user has viewed and includes both relevant and irrelevant images. To simultaneously measure the speed-based benefits of CuZero result exploration, users had a total of 5 minutes to browse images after query formulation, contrary to the standard 15 minutes available in the official TRECVID evaluation. First, this figure demonstrates that CuZero is competitive when compared to other interactive systems. Search topics that scored poorly (*190:person+books* and *193:smokestack*) can often be attributed to a lack of coverage by the Columbia374 concept classifiers. Additionally, other search systems typically use more diverse search tools (text, concept, visual similarity) than CuZero, which solely uses concept-based search in this experiment. Second, these results are quite impressive when one considers the fact that there was no user labeling involved and irrelevant results are

also included in the simulated CuZero result lists. For most of the search topics, CuZero performance was greater than or equal to the average performance of interactive systems given 300% of the time alloted for image browsing. With these observations, CuZero demonstrates promising benefits for both speed and performance over traditional visual search systems.

5. CONCLUSIONS AND FUTURE WORK

CuZero is a very fast, highly responsive interactive search system that satisfies many goals for an interactive system because it bridges the human-computer interface problem by maximizing the use of auto-recommendation systems and user freedom to examine many query options. First, it employs a zero-latency process to aide in query formulation, as guided by text and concept-tag suggestions derived exclusively from the user's entered query text. While each of these suggestion techniques may be suited for a few specific query topics, a user can interactively choose the best suggestions guided by automated frequency and performance cues. Second, it demonstrates that a dynamic and interactive weighting scheme allows more relevant results to be found (and in less time) when compared to trial-and-error search. With only changes to the way that results are presented, humans are able to more quickly inspect and label relevant image results. Finally, using asynchronous updates and caching, the latency incumbent on search systems can be avoided and a user can explore a data set more deeply in less time. Background navigation map updates and continuous scrolling are two techniques that fully utilize the asynchronous benefits of the proposed client-server topology.

Future revisions of CuZero would further harness its truly interactive nature and include additional search options for the navigation map. CuZero can observe all query revisions in the concept suggestion panel, the movements of the user in the navigation map, and any relevant/non-relevant labels that the user may explicitly provide to proactively suggest alternative query strategies that the user may not have considered otherwise . In this paper, the navigation map focused on concept-based exploration but future revisions of CuZero could include *visual icons* as search entries. Ranked results for a *visual icon* are based on similarity-based image retrieval. This fruitful functional addition is easily accommodated by the algorithms described in section 3.2. Finally, because the navigation map is merely a representation of different concept weights, CuZero could be modified to allow users to load or save navigation maps across many searches, thereby creating templates of navigation maps that may be quickly reapplied in the future by other users.

6. REFERENCES

[1] S.-F. Chang, *et al.*, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts." *Columbia ADVENT Technical Report #222-2006-8*, Columbia University, March 2007. URL: http://www.ee.columbia.edu/dvmm/columbia374.

[2] S.-F. Chang, L. Kennedy, E. Zavesky, "Columbia University's Semantic Video Search Engine." In *ACM International Conference on Image and Video Retrieval*, Amsterdam, Netherlands, July 2007.

[3] M. Christel, "Examining User Interactions with Video Retrieval Systems." In *SPIE Multimedia Content Access: Algorithms and Systems*, Vol. 6506, San Jose, CA, February, 2007.

[4] M. Christel, N. Papernick, N. Moraveji, and C. Huang, "Exploiting Multiple Modalities for Interactive Video Retrieval." In *International Conference on Acoustics, Speech and Signal Processing* (ICASSP'04), Montreal, Quebec, Canada, May 17-21, 2004.

[5] T.-S. Chua, S.-Y. Neo, H.-K. Goh, M. Zhao, Y. Xiao, G. Wang, "TRECVID 2005 by NUS PRIS." In NIST TRECVID workshop, Gaithersburg, MD, Nov 2005.

[6] C. Fellbaum, "WordNet An Electronic Lexical Database. Language, speech, and communication." *Cambridge, Mass: MIT Press*, 1998.

[7] L. Kennedy. "Revision of LSCOM Event/Activity Annotations, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia." ADVENT Technical Report 221-2006-7 Columbia University, December 2006.

[8] L. Kennedy, S.-F. Chang. "A Reranking Approach for Context-based Concept Fusion in Video Indexing and Retrieval." In *ACM International Conference on Image and Video Retrieval*, Amsterdam, Netherlands, July 2007.

[9] L. Kennedy, P. Natsev, S.-F. Chang. "Automatic Discovery of Query Class Dependent Models for Multimodal Search." In *ACM Multimedia*, Singapore, November 2005.

[10] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. "Semantic concept-based query expansion and re-ranking for multimedia retrieval." In *Proceedings of ACM International Conference on Multimedia* (MM'07), pages 991-1000, Augsburg, Germany, Sept. 2007.

[11] S.-Y. Neo, H. Luan, Y. Zheng, H.-K. Goh and T.-S. Chua, "VisionGo: Bridging Users and Multimedia Video Retrieval." In *International Conference On Image And Video Retrieval (CIVR))*, Niagara Falls, Canada, July 2008.

[12] G. P. Nguyen and M. Worring. "Optimization of interactive visual-similarity-based search." In ACM Transactions on Multimedia Computing, Communications and Applications, 4(1):7:1-23, January 2008.

[13] M. F. Porter, "An Algorithm for Suffix Stripping." In *Program*, 14(3): 130-137, 1980.

[14] Y. Rui, T. S. Huang, M. Ortega, S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval." In *IEEE Transactions On Circuits and Systems for Video Technology*, 1998, Vol 8; Number 5, pages 644-655.

[15] R. Singer, "Just one slider, but not so simple - Signal vs. Noise (by 37signals)". [web page] June 02, 2005. URL: http://www.37signals.com/svn/archives2/just_one_slider_but_not_so_simple.php. Accessed: May 20, 2008.

[16] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid." In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, California, October 26 - 27, 2006.

[17] J. R. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors." In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Baltimore, MD, July 2003.

[18] M. Worring, C. G. M. Snoek, O. Rooij, G. P. Nguyen, and A. Smeulders, "The MediaMill semantic video search engine." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, April 2007.

[19] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval." In *ACM International Conference on Multimedia*. Ottawa, Canada, September 2001.

[20] R. Yan, J. Yang, and A. Hauptmann. "Learning query-class dependent weights in automatic video retrieval." In Proceedings of the 12th annual ACM international conference on Multimedia (MM'04), New York, NY, October 10-16, 2004.

[21] E. Zavesky, S.-F. Chang, C.-C. Yang. "Visual Islands: Intuitive Browsing of Visual Search Results." In *ACM International Conference on Image and Video Retrieval*, Niagra Falls, Canada, July 2008.

[22] E. Zavesky, Z. Liu, D. Gibbon, B. Shahraray. "Searching Visual Semantic Spaces with Concept Filters." In *IEEE International Conference on Semantic Computing*, Irvine, California, September 2007.