(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 August 2010 (12.08.2010)

PCT

(10) International Publication Number
WO 2010/091358 A1

(51) International Patent Classification:
H04H 60/33 (2008.01)

(21) International Application Number:
PCT/US2010/023494

(22) International Filing Date:
8 February 2010 (08.02.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/151,124	9 February 2009 (09.02.2009)	US
61/171,789	22 April 2009 (22.04.2009)	US
61/233,325	12 August 2009 (12.08.2009)	US
61/233,675	31 August 2009 (31.08.2009)	US
PCT/US09/069237	22 December 2009 (22.12.2009)	US

(71) Applicant *(for all designated States except US)*: **THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK** [US/US]; 116 Street and Broadway, New York, NY 10027 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **CHANG, Shih-Fu** [US/US]; 560 Riverside Drive, Apt. 20J, New York, NY 10027 (US). **WANG, Jun** [CN/US]; 3 Hobart Court, New York, NY 10956 (US). **SAJDA, Paul** [US/US]; 101 West End Avenue, New York, NY 10027 (US). **POHLMEYER, Eric** [US/US]. **HANNA, Barbara** [FR/US].

(74) Agents: **RAGUSA, Paul, A.** et al.; Baker Botts LLP, 30 Rockefeller Plaza, New York, NY 10112-4498 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[Continued on next page]

(54) Title: RAPID IMAGE ANNOTATION VIA BRAIN STATE DECODING AND VISUAL PATTERN MINING

(57) Abstract: Human visual perception is able to recognize a wide range of targets but has limited throughput. Machine vision can process images at a high speed but suffers from inadequate recognition accuracy of general target classes. Systems and methods are provided that combine the strengths of both systems and improve upon existing multimedia processing systems and methods to provide enhanced multimedia labeling, categorization, and searching.

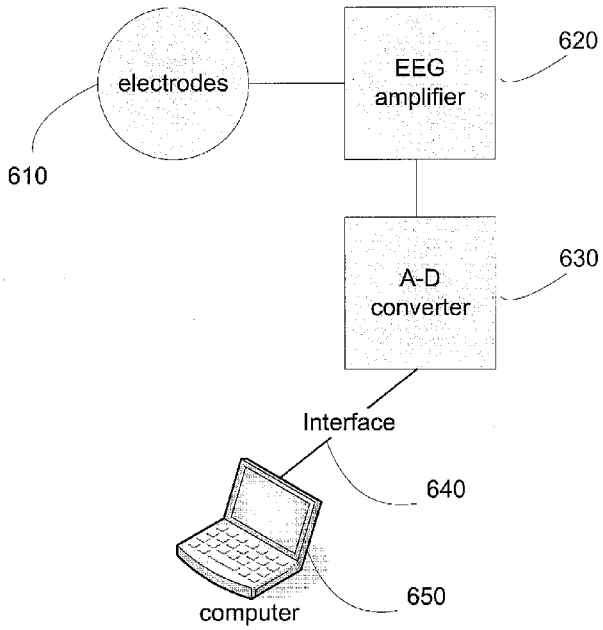

FIGURE 6

WO 2010/091358 A1

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

RAPID IMAGE ANNOTATION VIA BRAIN STATE DECODING AND VISUAL PATTERN MINING

SPECIFICATION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Application Nos. 61/151,124, filed on February 9, 2009, entitled, "System and Method for Arranging Media;" 61/171,789, filed on April 22, 2009, entitled "Rapid Image Annotation via Brain State Decoding and Visual Pattern Mining;" 61/233/325, filed August 12, 2009, entitled "System and Methods for Image Annotation and Label Refinement by Graph;" and PCT Patent Application No. PCT/US09/069237, filed on December 22, 2009, entitled "System and Method for Annotating and Searching Media," which are incorporated herein by reference in their entirety.

BACKGROUND

As the volume of digital multimedia collections grow, techniques for efficient and accurate labeling, searching and retrieval of data from those collections have become increasingly important. As a result, tools such as multimedia labeling and classification systems and methods that allow users to accurately and efficiently categorize and sort such data have also become increasingly important. Unfortunately, previous labeling and classification methods and systems tend to suffer deficiencies in several respects, as they can be inaccurate, inefficient and/or incomplete, and are, accordingly, not sufficiently effective to address the issues associated with voluminous collections of multimedia.

Various methods have been used to improve the labeling of multimedia data. For example, there has been work exploring the use of user feedback to improve the image retrieval experience. In some systems, relevance feedback provided by the user is used to indicate which images in the returned results are relevant or irrelevant to the users' search target. Such feedback can be indicated explicitly (by marking labels of relevance or irrelevance) or implicitly (by tracking specific images viewed by the user). Given such feedback information, the initial query can be modified. Alternatively, the underlying features and distance metrics used in representing and matching images can be refined using the relevance feedback information. Ultimately, though, the manual labeling by humans of multimedia data, such as

images and video, can be time consuming and inefficient, particularly when applied to large data libraries. Some solutions to the problems described above are disclosed in PCT Patent Application No. PCT/US09/069237, filed on December 22, 2009, the entirety of which is incorporated herein by reference.

The human brain is an exceptionally powerful visual information processing system. Humans can recognize objects at a glance, under varying poses, illuminations and scales, and are able to rapidly learn and recognize new configurations of objects and exploit relevant context even in highly cluttered scenes. While human visual systems can recognize a wide range of targets under challenging conditions, they generally have limited throughput. Human visual information processing happens with neurons which are extremely slow relative to state-of-the-art digital electronics–i.e. the frequency of a neuron's firing is measured in Hertz whereas modern digital computers have transistors which switch at Gigahertz speeds. Though there is some debate on whether the fundamental processing unit in the nervous system is the neuron or whether ensembles of neurons constitute the fundamental unit of processing, it is nonetheless widely believed that the human visual system is bestowed with its robust and general purpose processing capabilities not from the speed of its individual processing elements but from its massively parallel architecture.

Computer vision systems present their own unique benefits and potential issues. While computer vision systems can process images at a high speed, they often suffer from inadequate recognition accuracy for general target classes. Since the early 1960's there have been substantial efforts directed at creating computer vision systems which possess the same information processing capabilities as the human visual system. These efforts have yielded some successes, though mostly for highly constrained problems. One of the challenges in prior research has been in developing a machine capable of general purpose vision and mimicking human vision. Specifically, an important property of the human visual system is its ability to learn and exploit invariances.

SUMMARY

Both human and computer vision systems offer their own unique benefits and disadvantages. The presently disclosed subject matter combines the

benefits of brain state decoding and visual content analysis to improve multimedia data processing efficiency.

Certain embodiments of the disclosed subject matter use brain signals measured by EEG to detect and classify generic objects of interest in multimedia data.

Certain embodiments of the disclosed subject matter are designed to facilitate rapid retrieval and exploration of image and video collections. The disclosed subject matter incorporates graph-based label propagation methods and intuitive graphic user interfaces ("GUIs") that allow users to quickly browse and annotate a small set of multimedia data, and then in real or near-real time provide refined labels for all remaining unlabeled data in the collection. Using such refined labels, additional positive results matching a user's interest can be identified. Such a system can be used, e.g., as a bootstrapping system for developing additional target recognition tools needed in critical image application domains such as in intelligence, surveillance, consumer applications, biomedical applications, and in Internet applications.

Starting with a small number of labels, certain disclosed systems and methods can be implemented to propagate the initial labels to the remaining data and predict the most likely labels (or scores) for each data point on the graph. The propagation process is optimized with respect to several criteria. For example, the system can be implemented to consider factors such as: how well the predictions fit the already-known labels; the regularity of the predictions over data in the graph; the balance of labels from different classes; if the results are sensitive to quality of the initial labels and specific ways the labeled data are selected.

The processes providing the initial labels to label propagation systems can come from various sources, such as other classifiers using different modalities (for example, text, visual, or metadata), models (for example, supervised computer vision models or a brain computer interface), rank information regarding the data from other search engines, or even other manual annotation tools. In some systems and methods, when dealing with labels/scores from imperfect sources (e.g., search engines), additional processes can be implemented to filter the initial labels and assess their reliability before using them as inputs for the propagation process.

Certain embodiments of the disclosed subject matter use the output of the brain signal analysis as an input to a label propagation system which propagates the initial brain-signal based labels to the remaining data and predict the most likely

labels (or scores) for each data point on the graph or novel data not included in the graph.

The output of certain disclosed system embodiments can include refined or predicted labels (or scores indicating likelihood of positive detection) of some or all the images in the collection. These outputs can be used to identify additional positive samples matching targets of interest, which in turn can be used for a variety of functions, such as to train more robust classifiers, arrange the best presentation order for image browsing, or rearrange image presentations.

Certain embodiments of the disclosed subject matter use a refined or predicted label set to modify the initial set of data to be presented to a user in a brain signal based target detection system.

In a disclosed embodiment of a system and method in accordance with the disclosed subject matter, a partially labeled multimedia data set is received and an iterative graph-based optimization method is employed resulting in improved label propagation results and an updated data set with refined labels.

Embodiments of the disclosed systems and methods are able to handle label sets of unbalanced class size and weigh labeled samples based on their degrees of connectivity or other importance measures.

In certain embodiments of the disclosed methods and systems, after the propagation process is completed, the predicted labels of all the nodes of the graph can be used to determine the best order of presenting the results to the user. For example, the images can be ranked in the database in a descending order of likelihood so that user can quickly find additional relevant images. Alternatively, the most informative samples can be displayed to the user to obtain the user's feedback, so that the feedback and labels can be collected for those critical samples. These functions can be useful to maximize the utility of the user interaction so that the best prediction model and classification results can be obtained with the least amount of manual user input.

The graph propagation process can also be applied to predict labels for new data that is not yet included in the graph. Such processes can be based, for example, on nearest neighbor voting or some form of extrapolation from an existing graph to external nodes.

In some embodiments of the disclosed subject matter, to implement an interactive and real-time system and method, the graph based label propagation can

use a graph superposition method to incrementally update the label propagation results, without needing to repeat computations associated with previously labeled samples.

BRIEF DESCRIPTION OF THE DRAWINGS

Further features and advantages of the presently disclosed subject matter will become apparent from the following detailed description taken in conjunction with the accompanying figures showing illustrative embodiments of the disclosed subject matter, in which:

Fig. 1 is a diagram illustrating exemplary aspects of computer vision system modes in accordance with the presently disclosed subject matter;

Fig. 2 is diagram illustrating an exemplary graphic user interface (GUI) portion of a computer vision module in accordance with the presently disclosed subject matter;

Fig. 3 is a flow chart illustrating an exemplary computer vision labeling propagation and refining method in accordance with the presently disclosed subject matter;

Fig. 4 is a diagram illustrating a fraction of a computer vision constructed graph and computation of a computer vision node regularizer method in accordance with the presently disclosed subject matter;

Fig. 5 is a flow chart illustrating an exemplary computer vision labeling diagnosis method in accordance with the presently disclosed subject matter.

Fig. 6 illustrates hardware and functional components of an exemplary system for brain signal acquisition and processing;

Fig. 7 is a diagram illustrating exemplary aspects of a combined brain-computer multimedia processing system in accordance with the presently disclosed subject matter;

Fig. 8 is a flow chart illustrating an exemplary method according to the presently disclosed subject matter.

DETAILED DESCRIPTION

Systems and methods as disclosed herein can be used to overcome the labeling and classification deficiencies of prior systems and methods described above

by coupling both computer vision and human vision components in various configurations. Computer vision components will first be described in accordance with the present disclosure.

Figure 1 illustrates a system and various exemplary usage modes in accordance with the presently disclosed subject matter.

Given a collection of multimedia files, the exemplary computer vision components of Figure 1 can be used to build an affinity graph to capture the relationship among individual images, video, or other multimedia data. One exemplary computer vision system can be a transductive annotation by graph (TAG) data processing system. The affinity between multimedia files can be represented graphically in various ways, for example: a continuous valued similarity measurement or logic associations (e.g., relevance or irrelevance) to a query target, or other constraints (e.g., images taken at the same location). The graph can also be used to propagate information from labeled data to unlabeled data in the same collection.

As illustrated in Figure 1, each node in the graph **150** can represent a basic entity (data sample) for retrieval and annotation. In certain embodiments, nodes in the graph **150** can be associated with either a binary label (e.g., positive vs. negative) or a continuous-valued score approximating the likelihood of detecting a given target. The represented entity can be, for example, an image, a video clip, a multimedia document, or an object contained in an image or video. In an ingestion process, each data sample can first be pre-processed **120** (e.g., using operations such as scaling, partitioning, noise reduction, smoothing, quality enhancement, and other operations as are known in the art). Pre-filters can also be used to filter likely candidates of interest (e.g., images that are likely to contain targets of interest). After pre-processing and filtering, features can be extracted from each sample **130**. TAG systems and methods in accordance with the disclosed subject matter do not necessarily require usage of any specific features. A variety of feature sets preferred by practical applications can be used. For example, feature sets can be global (e.g., color, texture, edge), local (e.g., local interest points), temporal (e.g. motion), and/or spatial (e.g., layout). Also, multiple types and modalities of features can be aggregated or combined. Given the extracted features, affinity (or similarity) between each pair of samples is computed **140**. No specific metrics are required by TAG, though judicious choices of features and similarity metrics can significantly impact the quality of the final label prediction results. The pair-wise affinity values can then

be assigned and used as weights of the corresponding edges in the graph **150**. Weak edges with small weights can be pruned to reduce the complexity of the affinity graph **150**. Alternatively, a fixed number of edges can be set for each node by finding a fixed number of nearest neighbors for each node.

Once the affinity graph **150** is created, a TAG system can be used for retrieval and annotation. A variety of modes and usages could be implemented in accordance with the teachings herein. Two possible modes include: interactive **160** and automatic **170** modes. In the Interactive Mode **160**, users can browse, view, inspect, and label images or videos using a graphic user interface (GUI), an embodiment of which is described in more detail hereinafter in connection with Figure 2.

Initially, before any label is assigned, a subset of default data can be displayed in the browsing window of the GUI based on, for example, certain metadata (e.g., time, ID, etc.) or a random sampling of the data collection. Using the GUI, a user can view an image of interest and then provide feedback about relevance of the result (e.g., marking the image as "relevant" or "irrelevant" or with multi-grade relevance labels). Such feedback can then be used to encode labels which are assigned to the corresponding nodes in the graph.

In Automatic Mode **170**, the initial labels of a subset of nodes in the graph can be provided by external filters, classifiers, or ranking systems. For example, for a given target, an external classifier using image features and computer vision classification models can be used to predict whether the target is present in an image and assign the image to the most likely class (positive vs. negative or one of multiple classes). As another example, if the target of interest is a product image search for web-based images, external web image search engines can be used to retrieve most likely image results using a keyword search. The rank information of each returned image can then be used to estimate the likelihood of detecting the target in the image and approximate the class scores which can be assigned to the corresponding node in the graph. An initial label set can also be generated based on the initial output of the human vision components of Figure 1.

In this particular embodiment, the TAG system hardware configuration can include an audio-visual (AV) terminal, which can be used to form, present or display audio-visual content. Such terminals can include (but are not limited to) end-user terminals equipped with a monitor screen and speakers, as well as server and

mainframe computer facilities in which audio-visual information is processed. In such an AV terminal, desired functionality can be achieved using any combination of hardware, firmware or software, as would be understood by one of ordinary skill in the art. The system can also include input circuitry for receiving information to be processed. Information to be processed can be furnished to the terminal from a remote information source via a telecommunications channel, or it can be retrieved from a local archive, for example. The system further can include processor circuitry capable of processing the multimedia and related data and performing computational algorithms. The processor circuitry may be a microprocessor, such as those manufactured by Intel, or any other processing unit capable of performing the processing described herein. Additionally, the disclosed system can include computer memory comprising RAM, ROM, hard disk, cache memory, buffer memory, tape drive, or any other computer memory media capable of storing electronic data. Notably, the memory chosen in connection with an implementation of the claimed subject matter can be a single memory or multiple memories, and can be comprised of a single computer-readable medium or multiple different computer-readable media, as would be understood by one of ordinary skill in the art.

Figure 2 shows an exemplary system GUI that can optionally be implemented in accordance with the presently disclosed subject matter. The disclosed GUI can include a variety of components. For example, image browsing area **210**, as shown in the upper left corner of the GUI, can be provided to allow users to browse and label images and provide feedback about displayed images. During the incremental annotation procedure, the image browsing area can present the top ranked images from left to right and from top to bottom, or in any other fashion as would be advantageous depending on the particulars of the application. System status bar **220** can be used to display information about the prediction model used, the status of current propagation process and other helpful information. The system processing status as illustrated in Figure 2 can provide system status descriptions such as, for example, '*Ready*', '*Updating*' or *'Re-ranking.'* The top right area **230** of the GUI can be implemented to indicate the name of current target class, e.g., "statue of liberty" as shown in Figure 3. For semantic targets that do not have prior definition, this field can be left blank or can be populated with general default text such as "target of interest." Annotation function area **240** can be provided below the target name area **230**. In this embodiment, a user can choose from labels such as '*Positive*', '*Negative*',

and '*Unlabeled.*' Also, statistical information, such as the number of positive, negative and unlabeled samples can be shown. The function button in this embodiment includes labels *'Next Page'*, *'Previous Page'*, *'Model Update'*, '*Clear Annotation*', and '*System Info.*'

Various additional components and functions can be implemented. For example, image browsing functions can be implemented in connection with such a system and method. After reviewing the current ranking results or the initial ranking, in this embodiment, such functionality can be implemented to allow a user to browse additional images by clicking the buttons '*Next Page*' and '*Previous Page.*' Additionally, a user can also use the sliding bar to move through more pages at once.

Manual annotation functions can also optionally be implemented. In certain embodiments, after an annotation target is chosen, the user can annotate specific images by clicking on them. For example, in such a system, positive images can be marked with a check mark, negative images can be marked with a cross mark '×', and unlabeled images can be marked with a circle 'o'.

Automatic propagation functions can also be implemented in connection with certain embodiments. After a user inputs some labels, clicking the button '*Model Update*' can trigger the label propagation process and the system will thereafter automatically infer the labels and generate a refined ranking score for each image. A user can reset the system to its initial status by clicking the button labeled '*Clear Annotation.*' A user can also click the button labeled '*System Info*' to generate system information, and output the ranking results in various formats that would be useful to one of ordinary skill in the art, such as, for example, a MATLAB-compatible format.

In the GUI embodiment shown in Figure 2, two auxiliary functions are provided which are controlled by checking boxes '*Instant Update*' and '*Hide Labels.*' When a user selects '*Instant Update,*' the shown system will respond to each individual labeling operation and instantly update the ranking list. The user can also hide the labeled images and only show the ranking results of unlabeled images by checking '*Hide Labels.*'

Given assigned labels or scores for some subset of the nodes in the graph (the subset is usually but not necessarily a small portion of the entire graph), embodiments of the disclosed systems can propagate the labels to other nodes in the graph accurately and efficiently.

Figure 3 is a chart illustrating a TAG labeling propagation method in accordance with an exemplary implementation of the presently disclosed subject matter. At **310**, the similarity or association relations between data samples are computed or acquired to construct an affinity graph. In **320**, some graph quantities, including a propagation matrix and gradient coefficient matrix are computed based on the affinity graph. At **330**, an initial label or score set over a subset of graph data is acquired. In various embodiments, this can be done via either interactive or automatic mode, or by some other mode implemented in connection with the disclosed subject matter. At **340**, one or more new labels are selected and added to the label set. Procedure **350** is optional, where one or more unreliable labels are selected and removed from the existing label set. In **360**, cleaned label set are obtained and a node regularization matrix is updated to handle the unbalanced class size problem of label data set. Procedures **340**, **350**, and **360** can be repeated until a certain number of iterations or some stop criteria are met. In step **370**, the final classification function and prediction scores over the data samples are computed.

Additional description of computer vision algorithms and graph data generally described above is now provided. In an embodiment in accordance with the disclosed subject matter, an image set $\mathbf{X}=(\mathbf{X}_L,\mathbf{X}_U)$ can include labeled samples $\mathbf{X}_L=\{\mathrm{x}_1,\cdots,\mathrm{x}_l\}$ and unlabeled samples $\mathbf{X}_U=\{\mathrm{x}_{l+1},\cdots,\mathrm{x}_n\}$, where l is the number of labels. The corresponding labels for the labeled data set can be denoted as $\{y_1,\cdots,y_l\}$, where $y\in\{1,\cdots,c\}$ and c is the number of classes. For transductive learning, an objective is to infer the labels $\{y_{l+1},\cdots,y_n\}$ of the unlabeled data $\mathbf{X}_U=\{\mathrm{x}_{l+1},\cdots,\mathrm{x}_n\}$, where typically $l<<n$, namely only a very small portion of data are labeled. Embodiments can define an undirected graph represented by $G=\{\mathbf{X},\mathbf{E}\}$, where the set of node or vertices is $\mathbf{X}=\{\mathrm{x}_i\}$ and the set of edges is $\mathbf{E}=\{e_{ij}\}$. Each sample x_i can be treated as the node on the graph and the weight of edge e_{ij} can be represented as w_{ij}. Typically, one uses a kernel function $k(\cdot)$ over pairs of points to calculate weights, in other words $w_{ij}=k(x_i,x_j)$ with the RBF kernel being a popular choice. The weights for edges can be used to build a weight matrix which can be denoted by $\mathbf{W}=\{w_{ij}\}$. Similarly, the node degree matrix $\mathbf{D}=diag(d_1,...,d_n)$ can be defined as

$d_l = \sum_{j=l}^{n} \mathbf{W}_{ij}$. A graph related quantity $\Delta = \mathbf{D} - \mathbf{W}$ is called graph Laplacian and its normalized version is

$$\mathbf{L} = \mathbf{D}^{-\frac{1}{2}} \Delta \mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{S}$$

where $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}$. The binary label matrix $\mathbf{Y}$ can be described as $\mathbf{Y} \in \mathrm{B}^{n \times c}$ with $\mathbf{Y}_{ij} = 1$ if x_i has label $y_i = j$ (means data x_i belongs to class j) and $\mathbf{Y}_{ij} = 0$ otherwise (means data x_i is unlabeled). A data sample can belong to multiple classes simultaneously and thus multiple elements in the same row of Y can be equal to 1. Figure 4 shows a fraction of a representative constructed graph with weight matrix $\mathbf{W}$, node degree matrix $\mathbf{D}$, and label matrix $\mathbf{Y}$. A classification function $\mathbf{F}$, can then be estimated on the graph to minimize a cost function. The cost function typically enforces a tradeoff between the smoothness of the function over the graph and the accuracy of the function at fitting the label information for the labeled nodes.

Embodiments of the disclosed TAG systems and methods can provide improved quality of label propagation results. For example, disclosed embodiments can include: superposition law based incremental label propagation; a node regularizer for balancing label imbalance and weighting label importance; alternating minimization based label propagation; and label diagnosis through self tuning.

Embodiments of the disclosed TAG systems and methods can also include an incremental learning method that allows for efficient addition of newly labeled samples. Results can be quickly updated using a superposition process without repeating the computation associated with the labeled samples already used in the previous iterations of propagation. Contributions from the new labels can be easily added to update the final prediction results. Such incremental learning capabilities can be useful for achieving real-time responses to a user's interaction. Since the optimal prediction can be decomposed into a series of parallel problems, and the prediction score for individual class can be formulated as component terms that only depend on individual columns of a classification matrix $\mathbf{F}$:

$$\mathbf{F} = (\mathbf{I} - \alpha \mathbf{S})^{-1} \sum_{i=1}^{l} \hat{\mathbf{Y}}_i = \sum_{i=1}^{l} (\mathbf{I} - \alpha \hat{\mathbf{S}})^{-1} \mathbf{Y}_i = \sum_{i=1}^{l} \hat{\mathbf{F}}_i$$

where $\alpha \in (0,1)$ is a constant parameter. Because each column of $\mathbf{F}$ encodes the label information of each individual class, such decomposition reveals that biases can arise

if the input labels are disproportionately imbalanced. Prior propagation algorithms often failed in this unbalanced case, as the results tended to be biased towards the dominant class. To overcome this problem, embodiments disclosed herein can apply a graph regularization method to effectively address the class imbalance issue. Specifically, each class can be assigned an equal amount of weight and each member of a class can be assigned a weight (termed as node regularizer) proportional to its connection density and inversely proportional to the number of samples sharing the same class.

$$\mathbf{F} = \sum_{i=1}^{l} v_{ii} \hat{\mathbf{F}}_i = \sum_{i=1}^{l} (\mathbf{I} - \alpha \hat{\mathbf{S}})^{-1} v_{ii} \mathbf{Y}_i = (\mathbf{I} - \alpha \hat{\mathbf{S}})^{-1} \mathbf{V}\mathbf{Y}$$

where the diagonal matrix $\mathbf{V} = \{v_{ii}\}$ is introduced as a node regularizer to balance the influence of labels from different classes. Assume sample x_i is associated with label j, the value of v_{ii} is computed as:

$$v_{ii} = d_i / \sum_{k=1}^{l} d_k \mathbf{Y}_{kj}$$

where d_i is the node degree of labeled sample x_i and $\sum_{k=1}^{l} d_k \mathbf{Y}_{kj}$ is the sum of node degree of the labeled nodes in class j. Figure 5 illustrates the calculation of node regularizer on a fraction of an exemplary constructed graph. The node weighting mechanism described above allows labeled nodes with a high degree to contribute more during the graph diffusion and label propagation process. However, the total diffusion of each class can be kept equal and normalized to be one. Therefore the influence of different classes can be balanced even if the given class labels are unbalanced. If class proportion information is known beforehand, it can be integrated into particular systems and methods by scaling the diffusion with the prior class proportion. Because of the nature of graph transduction and unknown class prior knowledge, however, equal class balancing leads to generally more reliable solutions than label proportional weighting.

Along with the node regularizer, incremental learning by superposition law is described here as another embodiment of the disclosed systems and methods. Let $D_j = \sum_{k=1}^{l} d_k \mathbf{Y}_{kj}$ denotes the total degree of the current labels in class j. Adding a

new labeled sample x_s (the corresponding degree is d_{ss}) to class j, two coefficients λ, γ can be calculated as:

$$\lambda = \frac{D_j}{D_j + d_{ss}} \qquad \gamma = \frac{d_{ss}}{D_j + d_{ss}}$$

Then the new prediction score for class j can be rapidly computed as:

$$\mathbf{F}_{.j}^{new} = \lambda \mathbf{F}_{.j} + \gamma \mathbf{P}_{.s}$$

where $\mathbf{F}_{.j}$ is the j th column of the classification matrix $\mathbf{F}$ and $\mathbf{P}_{.s}$ is the j th column of the propagation matrix $\mathbf{P}$ (The propagation matrix will be defined later). This is in contrast to a brute force approach that uses the whole set of labeled samples, including the new labeled sample and the existing labeled samples, to calculate the classification function from scratch again. The disclosed systems and methods result in a much more efficient implementation of the label propagation process.

Certain embodiments of the disclosed systems and methods make modifications to the cost function used in previously used systems and methods. For example, in certain systems and methods, the optimization is explicitly shown over both the classification function $\mathbf{F}$ and the binary label matrix $\mathbf{Y}$:

$$(\mathbf{F}^*, \mathbf{Y}^*) = \arg\min_{\mathbf{F} \in R^{n \times c}, \mathbf{Y} \in \mathbf{B}^{n \times c}} Q(\mathbf{F}, \mathbf{Y})$$

where B is the set of all binary matrices Y of size $n \times c$ that satisfy $\sum_j \mathbf{Y}_{ij} = 1$ for a single labeling problem, and for the labeled data $x_i \in \mathbf{X}_l$, $\mathbf{Y}_{ij} = 1 \text{if } y_i = j$. However, embodiments of the disclosed systems and methods naturally adapt to a multiple-label problem, where single multimedia file can be associated with multiple semantic tags. More specifically, the loss function is:

$$Q(\mathbf{F}, \mathbf{Y}) = \frac{1}{2} tr \left\{ \mathbf{F}^T \mathbf{L} \mathbf{F} + \mu (\mathbf{F} - \mathbf{V}\mathbf{Y})^T (\mathbf{F} - \mathbf{V}\mathbf{Y}) \right\}$$

where the parameter μ balances two parts of the cost function. The node regularizer $\mathbf{V}$ permits the use of a normalized version of the label matrix $\mathbf{Z}$ defined as: $\mathbf{Z} = \mathbf{V}\mathbf{Y}$. By definition, in certain embodiments, the normalized label matrix satisfies $\sum_i \mathbf{Z}ij = 1$.

An alternating minimization procedure to solve the above optimization problem can also contribute to improvements over prior methods and systems, as disclosed herein. Specifically, the cost function discussed above includes two

variables that can be optimized. While simultaneously recovering both solutions can be difficult due to the mixed integer programming problem over binary $\mathbf{Y}$ and continuous $\mathbf{F}$, a greedy alternating minimization approach can be used instead. The first update of the continuous classification function $\mathbf{F}$ is straightforward since the resulting cost function is convex and unconstrained, which allows the optimal $\mathbf{F}$ to be recovered by setting the partial derivative $\frac{\partial Q}{\partial F}$ equal to zero. However, since $\mathbf{Y} \in \mathrm{B}^{n \times c}$ is a binary matrix and subject to certain linear constraints, another part of another embodiment of the disclosed alternating minimization requires solving a linearly constrained max cut problem which is **NP**. Due to the alternating minimization outer loop, investigating guaranteed approximation schemes to solve a constrained max cut problem for $\mathbf{Y}$ can be unjustified due to the solution's dependence on the dynamically varying classification function $\mathbf{F}$ during an alternating minimization procedure. Instead, embodiments of the currently disclosed methods and systems can use a greedy gradient-based approach to incrementally update $\mathbf{Y}$ while keeping the classification function $\mathbf{F}$ at the corresponding optimal setting. Moreover, because the node regularizer term $\mathbf{V}$ normalizes the labeled data, updates of $\mathbf{V}$ can be interleaved based on the revised $\mathbf{Y}$.

The classification function, $\mathbf{F} \in R^{n \times c}$, as used in certain embodiments of the disclosed subject matter, is continuous and its loss terms are convex, which allows its minimum to be recovered by zeroing the partial derivative:

$$\frac{\partial Q}{\partial F} = 0 \Rightarrow \mathbf{LF} + \mu(\mathbf{F}^* - \mathbf{VY}) = 0$$
$$\Rightarrow \mathrm{F}^* = (\mathrm{L}/\mu + \mathrm{I})^{-1}\,\mathbf{VY} = \mathbf{PVY}$$

where $\mathbf{P} = (\mathbf{L}/\mu + \mathbf{I})^{-1}$ is denoted as the propagation matrix and can assume the graph is symmetrically built. To update $\mathbf{Y}$, first $\mathbf{Y}$ can be replaced by its optimal value $\mathbf{F}^*$ as shown in the equation above. Accordingly:

$$\begin{aligned} Q(\mathbf{Y}) &= \frac{1}{2}\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\,\mathbf{V}^{\mathrm{T}}\,\mathbf{P}^{\mathrm{T}}\mathbf{LPVY} \\ &\quad + \mu(\mathbf{PVY} - \mathbf{VY})^{t}\,(\mathbf{PVY} - \mathbf{VY})) \\ &= \frac{1}{2}\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\,\mathbf{V}^{\mathrm{T}}\,[\mathbf{P}^{\mathrm{T}}\mathbf{LP} + \mu(\mathbf{P}^{\mathrm{t}} - \mathbf{I})(\mathbf{P} - \mathbf{I})]\mathbf{VY}) \end{aligned}$$

This optimization still involves the node regularizer $\mathbf{V}$, which depends on $\mathbf{Y}$ and normalizes the label matrix over columns. Due to the dependence on the current estimate of $\mathbf{F}$ and $\mathbf{V}$, only an incremental step will be taken greedily in certain disclosed embodiments to reduce $Q(\mathbf{Y})$. In each iteration, position (i^*, j^*) in the matrix $\mathbf{Y}$ can be found and the binary value $\mathbf{Y}_{i^*j^*}$ of can be changed from 0 to 1. The direction with the largest negative gradient can guide the choice of binary step on $\mathbf{Y}$. Therefore, $\frac{\partial Q}{\partial \mathbf{Y}}$ can be evaluated and the associated largest negative value can be found to determine (i^*, j^*).

Note that setting $\mathbf{Y}_{i^*j^*} = 1$ is equivalent to a similar operation on the normalized label matrix $\mathbf{Z}$ by setting $\mathbf{Z}_{i^*j^*} = \varepsilon \; 0 < \varepsilon < 1$, and $\mathbf{Y}, \mathbf{Z}$ to have one-to-one correspondence. Thus, the greedy minimization of Q with respect to $\mathbf{Y}$ in this disclosed embodiment is equivalent to the greedy minimization of Q with respect to $\mathbf{Z}$:

$$(i^*, j^*) = \arg\min_{i,j} \frac{\partial Q}{\partial \mathbf{Z}}$$

The loss function can be rewritten using the variable $\mathbf{Z}$ as:

$$Q(\mathbf{Z}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^T\left[\mathbf{P}^T\mathbf{L}\mathbf{P} + \mu\left(\mathbf{P}^T - \mathbf{I}\right)\left(\mathbf{P} - \mathbf{I}\right)\right]\mathbf{Z}\right) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^T\mathbf{A}\mathbf{Z}\right)$$

where $\mathbf{A}$ represents $\mathbf{A} = \mathbf{P}^T\mathbf{L}\mathbf{P} + \mu\left(\mathbf{P}^T - \mathbf{I}\right)\left(\mathbf{P} - \mathbf{I}\right)$. Note that $\mathbf{A}$ is symmetric if the graph is symmetrically built. The gradient of the above loss function can be derived and recovered with respect to $\mathbf{Z}$ as: $\frac{\partial Q}{\partial \mathbf{Z}} = \mathbf{A}\mathbf{Z} = \mathbf{A}\mathbf{V}\mathbf{Y}$. As described earlier, the gradient matrix can be searched to find the minimal element for updating the following equation:

$$(i^*, j^*) = \arg\min_{x \in X_u, 1 \le j \le c} \nabla_{\mathbf{Z}_{ij}} Q$$

The label matrix can be updated by setting $\mathbf{Y}_{i^*j^*} = 1$. Because of the binary nature of $\mathbf{Y}$, $\mathbf{Y}_{i^*j^*}$ can be set to equal 1 instead of using a continuous gradient approach. Accordingly, after each iteration, the node regularizer can be recalculated using the updated label matrix.

The updated $\mathbf{Y}$ in accordance with certain disclosed embodiments is greedy and could therefore oscillate and backtrack from predicted labeling in previous iterations without convergence guarantees. To guarantee convergence and avoid backtracking, inconsistency or unstable oscillation in the greedy propagation of labels, in preferred embodiments, once an unlabeled point has been labeled, its labeling can no longer be changed. In other words, the most recently labeled point (i^*, j^*) is removed from future consideration and the algorithm only searches for the minimal gradient entries corresponding to the remaining unlabeled samples. Thus, to avoid changing the labeling of previous predictions, the new labeled node x_i can be removed from $\mathbf{X}_u$ and added to $\mathbf{X}_l$.

The following equations summarize the updating rules from step t to $t+1$ in certain embodiments of the scheme of graph transduction via alternative minimization (GTAM). Although the optimal $\mathbf{F}^*$ can be computed in each iteration, it does not need to explicitly be updated. Instead, it can be implicitly used to directly updated $\mathbf{Y}$:

$$
\begin{aligned}
\nabla_Z Q^t &= \mathbf{A} \cdot \mathbf{V}^t \mathbf{Y}^t \\
(i^*, j^*) &= \arg\min_{x_i \in X_u, 1 \le j \le c} \nabla_{Z_{ij}} Q^t \\
\mathbf{Y}_{i^* j^*}^{t+1} &= 1 \\
v_{ii}^{t+1} &= d_i / \sum_{k=1}^{l} d_k \mathbf{Y}^{t+1}{}_{kj} \\
\mathbf{X}_U^{t+1} &\leftarrow \mathbf{X}_L^t + x_{i^*}; \mathbf{X}_U^{t+1} \leftarrow \mathbf{X}_U^t - x_{i^*}
\end{aligned}
$$

The procedure above can repeat until all points have been labeled in connection with the label propagation of the disclosed subject matter.

To handle errors in a label set, embodiments of the disclosed methods and systems can be extended to formulate a graph transduction procedure with the ability to handle mislabeled instances. A bidirectional greedy search approach can be used to simultaneously drive wrong label correction and new label inferencing. This mechanism can allow for automatic pruning of incorrect labels and maintain a set of consistent and informative labels. Modified embodiments of the systems and methods disclosed earlier can be equipped to more effectively deal with mislabeled samples and develop new "Label Diagnosis through Self Tuning" (LDST) systems and methods. In an exemplary embodiment of these systems and methods, a set of initial

labels is acquired. They can be acquired, for example, either by user annotation or from another resource, such as text-based multimedia search results. The gradient of the cost function with respect to label variable is computed based on the current label set is computed, and a label is added from said unlabeled data set based on the greedy search, i.e., finding the unlabeled sample with minimum gradient value. A label is then removed from said label set based on the greedy search, i.e., finding the labeled sample with maximum gradient value. The last two can be performed in reverse order without losing generalization, and can be executed a variable number of times (e.g., several new labels can be added after removing an existing label). Certain embodiments of the disclosed systems and methods update the computed gradients based on the new label set and repeat the last two parts of the procedure to retrieve a refined label set.

Embodiments of the disclosed LDST systems and methods can be used to improve the results of text based image search results. In one embodiment, top-ranked images can be truncated to create a set of pseudo-positive labels, while lower-ranked images can be treated as unlabeled samples. LDST systems and methods can then be applied to tune the imperfect labels and further refine the rank list. Additional embodiments can be used on a variety of data set types, including text classification on webpages and to correctly identify handwritten data samples.

The human vision components of the presently disclosed subject matter (including the brain-computer interface) are now described.

As previously mentioned, in recent years, there has been substantial interest in decoding the human brain state. There have been a variety of neural signals which have been targeted for decoding, ranging from spike trains collected via invasive recordings to hemodynamic changes measured via non-invasive fMRI. Systems and methods in accordance with the disclosed subject matter use EEG as a non-invasive measure to relate brain state to events correlated with the detection of "interesting" visual objects and images. One example of a robust signal using EEG is the P300. The P300 reflects a perceptual orienting response or shift of attention which can be driven by the content of the sensory input stream. Oscillatory brain activity often found during resting state (10Hz oscillations known as "alpha" activity) as well as transient oscillations sometimes associated with perceptual processing (30Hz and higher known as "gamma" activity) can also be indicative of a subject's attention state.

Certain systems and methods in accordance with the disclosed subject matter distinguish between two distinct brain states: (+) positive states in which the subject sees something of interest in an image, versus (-) negative states for which the image contains nothing of particular interest. This distinction is not to deduce from the brain signal what the exact content is or what the subject sees in the image, but instead, to utilize the high temporal resolution of EEG to detect individual recognition events from just a short segment of EEG data. For individual images that are presented to a subject for as little as 100ms, exemplary embodiments of the disclosed systems and methods can detect the brain signals elicited by positive examples, and distinguish them from the brain activity generated by negative example images. A task for exemplary disclosed analysis systems and methods is to classify the signal between two possible alternatives.

In exemplary systems and methods in accordance with the disclosed subject matter, a fast image sequences is presented to a user via a process known as rapid serial visual presentation (RSVP). In exemplary embodiments of RSVP, images can be presented very rapidly, for example, at rates of 5 to 10 images per second. To classify brain activity elicited by these images, certain exemplary methods analyze 1s of data, recorded with multiple surface electrodes, following the presentation of an image.

Systems and methods in accordance with the disclosed subject matter can be used to measure linear variations in EEG measurement signals. By averaging over EEG measurements with appropriate coefficients (positive or negative with magnitudes corresponding to how discriminant each electrode is) it is possible to obtain a weighted average of the electrical potentials that can be used to differentiate positive from negative examples as represented below:

$$y_t = \sum_i w_i x_{it}$$

Here x_{it} represents the electrical potential measured at time t for electrode i on the scalp surface, while w_i represents the spatial weights which have to be chosen appropriately. A goal of this summation is to combine voltages linearly such that the sum y is maximally different between two conditions. This can be thought of as computing a neuronal current source y_t that differs most between times samples $t+$ following positive examples and the times $t-$ following negative

examples, $y_{t+} > y_{t-}$. Label '+' indicates that the expression is evaluated with a signal x_{it} recorded following positive examples and label '-' indicates the same for negative examples. There are a number of algorithms available to find the optimal coefficients w_i in such a binary linear classification problem, e.g., Fisher Linear Discriminants (FLD), Penalized Logistic Regression (PLR), and Support Vectors Machines (SVM).

In certain methods and systems in accordance with the disclosed subject matter, optimal weight vectors, w_{ki} are calculated for a time window following the presentation of the data (index k labels the time window):

$$y_{kt} = \sum_i w_{ki} x_{it}, \quad t = (k-1)T \ldots kT$$

These different current sources y_{kt} can then be combined in an average over time to provide the optimal discriminant activity over the relevant time period:

$$y = \sum_t \sum_k v_k y_{tk};$$

For an efficient on-line implementation of this method, FLD can be used to train coefficients w_{ik} within each window of time, i.e., w_{ik} is trained such that $y_{kt+} > y_{kt-}$. The coefficients v_k can be used learned using PLR after all exemplars have been observed such that $y_+ > y_-$. Because of the two-part process of first combining activity in space, and then again in time, this algorithm can be referred to as a "Hierarchical Discriminant Component Analysis"

Note that the first part in the exemplary embodiments described above does not average over time samples within each window. Instead, each time sample provides a separate exemplar that is used when training the FLD. For instance, a system with 50 training exemplars and 10 samples per window results in 500 training samples for a classification algorithm that can need to find 64 spatial weighting coefficients w_{ik} for the kth window. These multiple samples within a time window will correspond to a single examplar image and are therefore not independent. They do, however, provide valuable information on the noise-statistic: variations in the signal within the time window are assumed to reflect non-discriminant "noise." In other words, one can assume that spatial correlation in the high-frequency activity ($f > 1/T$) is shared by the low-frequency discriminant activity. In addition, by training the spatial weight separately for each window one assume that the discriminant activity is not correlated in time beyond the time window time scale. Both these assumptions

contribute to a system's ability to combine thousands of dimensions optimally despite the small number of known training images.

The method described above combines activity linearly. This is motivated by the notion that a linear combination of voltages corresponds to a current source, presumably of neuronal origin within the skull. Thus, this type of linear analysis is sometimes called source-space analysis ("beam-forming" is a common misnomer for the same). A general form of combining voltages linearly in space and time can be represented as:

$$y = \sum_t \sum_i w_{it} x_{it}$$

However, the number of free parameters w_{it} in this general form is the full set of dimensions – 6,400 for certain embodiments, for example – with only a handful of positive exemplars to choose their values. To limit the degrees of freedom one can restrict the matrix w_{it} to be of lower rank, e.g., K. The linear summation can then be written as:

$$y = \sum_t \sum_i w_{it} x_{it} ,$$

where is a low-rank bilinear representation of the full parameter space.

This bilinear model assumes that discriminant current sources are static in space with their magnitude (and possibly polarity) changing in time. The model allows for K such components with their spatial distribution captured by u_{ik} and their temporal trajectory integrated with weights v_{tk}. Again, a goal is to find coefficient u_{ik}, v_{tk} such that the bilinear projection is larger for positive examples than for negative examples, i.e., $y_+ > y_-$. Notably, the x values referenced above need not be a time-domain signal, but could also be in the frequency domain. The linear integration could be performed in either domain.

The algorithms presented so far capture a type of activity that is often referred to as event related potentials (ERP). This term, ERP, refers to activity that is evoked in a fixed temporal relationship to an external event, that is, positive and negative deflections occur at the same time relative to the event – for example, the time of image presentation. In addition to this type of evoked response activity the EEG often shows variations in the strength of oscillatory activity. Observable events

can change the magnitude of ongoing oscillatory activity or can induce oscillations in the EEG. To capture the strength of an oscillation, irrespective of polarity, it is common to measure the "power," or the square of the signal, typically after it has been filtered in a specific frequency band. Instead of a linear combination, to capture power, one has to allow for a quadratic combination of the electrical potentials.

Once an interest score, y, is calculated, it can be converted to an interest label for use in a multimedia analysis/computer vision system as described above. More specifically, a binarization function $g(\cdot)$ can be applied to convert interest scores to multimedia labels as $\mathbf{y} = g(\mathbf{e})$, where $yi \in \{1,0\}$ and $yi = 1$ for $ei > \varepsilon$, otherwise $yi = 0$. The value ε can be referred to as an interest level for discretizing the EEG scores

Using the principles of computer vision systems (such as TAG) and brain-computer interfaces (EEG-based) as set forth above, systems and methods according to the disclosed subject matter herein can be implemented.

In accordance with the disclosed subject matter, it is possible to implement computer vision followed by EEG-RSVP systems and methods described above. Given prior information of a target type which can be instantiated in a TAG-based model, including contextual cues, TAG-based processing can operate on a dataset so as to eliminate regions of very low target probability and also provide an initial ordering of regions having high target probability. In addition, TAG can center image chips of potential regions of interest (ROIs) in a large image or set of images, which improves detection since potential targets are foveated when images are presented to subject. The top M images of the reordered dataset, in which sensitivity is high but specificity can be low, can be sampled and presented to the subject for EEG-RSVP analysis. The brain-computer interface processing can be tuned to produce high sensitivity and low specificity, with the EEG-RSVP mode can be used to increase specificity while maintaining sensitivity.

Figure 6 illustrates the hardware components of a particular embodiment of a subsystem for brain data acquisition in accordance with the disclosed subject matter. Such a subsystem can include EEG electrodes **610**, which may be passive or active electrodes. The electrodes are connected to an EEG amplifier **620**, which processes and amplifies the EEG signals for further analysis. An analog-to-digital converter **630** is then used to input the data received from the

amplifier into a computer **650**. Interface **640** between the A-D converter **630** and the computer **650** may be a wire interface connected via USB or other standard, or a wireless connection via Bluetooth or other standard, or any other known mechanism for data transfer. In certain embodiments, the system hardware implementation can use multiple computers **650**, such as three personal computers (laptop, desktop, handheld, or any other personal computing device), two used for the RSVP and EEG recording and classification, and one for image processing, or the functionality of all modules could be performed from a single computer. One of ordinary skill in the art would understand a variety of different configurations of such a system, including a general purpose personal computer programmed with software sufficient to enable the methods of the disclosed subject matter described herein.

In one exemplary embodiment, the analysis system utilizes a 64 electrode EEG recording system in a standard montage. EEG can be recorded at, for example, a 1 kHz sampling rate. While the EEG is being recorded, the RSVP display module can uses a dedicated interface to display blocks of images at the specified frame rate. In certain embodiments, blocks are typically 100 images long. The frame rate can be set to 5 or 10Hz depending on the target/imagery types, and the human observer's response to preliminary presentations. The interface draws from a pool of potential target chips and a pool of "distracters." One role of the distracters is to achieve a desired prevalence of target chips, that will maintain the human observer engaged in the presentation: if the prevalence is too low or too high, the observer can not keep an adequate focus and can more easily miss detections. Given that the computer vision outputs include some false positives, the number of distracters used depends in fact on the expected number of true target chips from the computer vision module.

The exemplary EEG analysis module can receive a list of image chips and detection details from the computer vision module, that includes pixel locations and detection confidence scores, and uses this input to generate the RSVP image sequences that will be used for presentation to the subject and analysis. It then performs several tasks: it acquires and records the EEG signals, using for example the hardware identified in Figure 6, orchestrates the RSVP, matches the EEG recordings with the presented images, trains an underlying classifier using training sequences, and uses the classifier with new generated image sequences.

In certain embodiments, a classification module relies on a hierarchical discriminant component analysis algorithm. At the first level, the classifier can use multiple temporal linear discriminators, each trained on a different time window relative to the image onset, to estimate EEG signatures of target detection. At a second level, the classifier can estimate a set of spatial coefficients that will optimally combine the outputs of the temporal discriminators to yield the final classification outcomes. The classification module is used in two different stages: training and actual usage with new imagery. The training can include a presentation of blocks with a set number of known targets in each block. The training sequences need not be related to the test sequences, in terms of their content, as the premise of the approach is that it detects objects of interest, but is not sensitive to the signatures of specific objects, and can therefore maintain its detection performance from one type of imagery to another.

Once an exemplary human imaging module has analyzed the relevant data, it can generate a list of images or image chips and their associated classification confidences, which can be used to prioritize the visualization of the corresponding images or image locations. The visualization interface permits the visualization of the prioritized locations in an adequate software environment for the task or user at hand. For example, for image analysts, certain embodiments use an interface to RemoteView, an imagery exploitation software application often used in the GeoIntelligence community. The interface provides a play control like toolbar that lets the analyst jump from one prioritized location to the next, while the analyst still retains access to all of RemoteView's functionality.

It is also possible and useful in connection with the disclosed subject matter to implement EEG-RSVP analysis systems and methods followed by computer vision (such as TAG) systems and methods as part of a combined system and method. In certain exemplary embodiments, in the absence of prior knowledge of the target type or a model of what an "interesting" image is, the EEG-RSVP is first run on samples of D_i, which can result in an image reordering in which images are ranked based on how they attracted the human subject's attention. This reordering can be used to generate labels for a computer vision based learning system which, given a partial labeling of D_i, propagates these labels and re-orders the database. In this embodiment, EEG scores are numbers with more positive scores indicating that the subject was interested in or strongly attending to the presented multimedia data. The

scores are sorted and the multimedia data associated with the top N scores are considered positives and labeled as such (given label +1) and used as training data for the TAG system. N can be chosen to be fixed (e.g., top 20 scores) or can be selected based on the requirements of a certain precision (e.g., the N scores where at least X% are true positives and 100-X% are false positives). In another embodiment, the lowest M ranked EEG scores are considered negatives and labeled as such (label = -1) so that N positive examples and M negative examples are provided to the TAG system. In another embodiment, the real-number values of the EEG scores are used to weight the strength of the training examples. For instance a EEG score of 0.3 for a EEG labeled image would result in a training label that is three times stronger than an EEG labeled image with a score of 0.1. The EEG-RSVP is designed to identify a small number "interesting" images which are then used by a semi-supervised computer vision system to reorder the entire image database.

Exemplary systems in which computer vision analysis follows EEG analysis can be similar to the alternative computer vision followed by EEG systems, in that they can use the same type of components, such as a computer vision module, an EEG analysis module, and a visualization/review module. However, in these systems, the EEG analysis module precedes the computer vision TAG components. Additionally, the number of examples provided by the EEG analysis can be insufficient to train conventional supervised learning algorithms, and there can be inaccuracies in the EEG outputs due to typically lower sensitivity of the approach. Therefore, a computer vision TAG module underpinned by a semi-supervised learning algorithm can be used to improve the EEG output. In certain embodiments, the outputs of the EEG systems and methods are a set of positive and negative examples (as determined by a suitable EEG confidence threshold), that can serve as labeled inputs to a graph-based classifier to predict the labels of remaining unlabeled examples in a database. TAG can then incorporate its automated graph-based label propagation methods and in real or near-real time generate refined labels for all remaining unlabeled data in the collection.

Further, as previously mentioned, it is also possible to have tightly coupled EEG systems and computer vision systems and methods in accordance with the disclosed subject matter. In certain embodiments, both EEG analysis and computer visions are run in parallel and coupled either at the feature space level, or at the output (or confidence measure) level, leading to a single combined confidence

measure that can serve as a priority indicator. This mode can require prior information on the target type. These modes can also potentially include feedback or multiple iterations within a closed-loop system.

Additionally, in certain embodiments of the disclosed subject matter, data is first analyzed using a EEG system and the results are used as the basis for input to a computer vision system such as TAG. The computer vision system can then be used to refine the data to be presented to a human once again as part of further EEG-based analysis. Among other things, a closed-loop implementation such as this can allow for more refined results and more efficient analysis of large sets of data.

Figure 7 is a diagram illustrating exemplary aspects of a combined human-computer/multimedia-processing system in accordance with the presently disclosed subject matter. In this particular embodiment, EEG-based generic interest detector **710** includes an interest object detector **710** which performs calculations resulting in an initial label (annotation) set **730**. This set can be used as an input to a computer vision system **740** including, among other things, a visual similarity graph **750** and a label refinement module **760** which can be used to generate a refined label/annotation set **770**.

Figure 8 is a flow chart illustrating a combined EEG-based interest detection method coupled to a TAG labeling propagation method in accordance with an exemplary implementation of the presently disclosed subject matter. At **810**, a system presents multimedia data to a user. At **820**, the system receives user response data based on the user's brain signal response to the presented data. At **830**, the system determines user interest in the presented data based on the EEG response data. **820** and **830** involve using hardware such as the system of Figure 6 to, more specifically, receive EEG signals and amplify, decode, and process them in real time as the human subject is viewing a rapid succession of images being presented on a display. At **840**, the system can extract relevant features from image data and generates an initial label set. At **850**, the similarity or association relations between data samples are computed or acquired to construct an affinity graph. At **860**, some graph quantities, including a propagation matrix and gradient coefficient matrix, are computed based on the affinity graph. At **870**, an initial label or score set over a subset of graph data is acquired from the system component which generated the initial label set based on user response data. **850** and **860** can be performed before or after **870**. At **875**, which is optional, one or more new labels are selected and added to

the label set. **880** is optional, wherein one or more unreliable labels are selected and removed from the existing label set. At **890**, cleaned label set are obtained and a node regularization matrix is updated to handle the unbalanced class size problem of label data set. Note that **875**, **880** and **890** can be repeated, if necessary, until a certain number of iterations or some stop criteria are met. At **895**, the final classification function and prediction scores over the data samples are computed. The output of **895** can be used to select and arrange certain presentations of images to users as input to **810** to complete a looped integration system.

Certain embodiments of the disclosed systems and methods can also be used for web search improvements. Images on such web sharing sites often are already associated with textual tags, assigned by users who upload the images. However, it is well known to those skilled in the art that such manually assigned tags are erratic and inaccurate. Discrepancies can be due, for example, to the ambiguity of labels or lack of control of the labeling process. Embodiments of the disclosed systems and methods can be used to efficiently refine the accuracy of the labels and improve the overall usefulness of search results from these types of internet websites, and more generally, to improve the usefulness and accuracy of internet multimedia searches overall.

Because the disclosed systems and methods are scalable in terms of feature representation, other application specified features can also be utilized to improve the graph propagation.

In another embodiment, a system and method for collaborative search can be employed. In such an embodiment, multiple EEG scores can be received from multiple human observers simultaneously, with the human observers each being presented the same multimedia data. The multiple scores can be processed and used to construct labels for the displayed multimedia data.

Further, computer vision systems and methods for use with brain-computer interfaces and methods as described herein are not limited to TAG or LDST systems and methods. Other back-end systems that may be used to process the brain-computer interface data may include any type of graphical probabilistic/generative models, including clustering, support vector machines, belief networks, and kernel-based systems and methods. Any computer vision label propagation component may be utilized. Ultimately, the described brain-computer interface can be used in

conjunction with a number of different computer analysis systems to achieve the principles of the disclosed subject matter.

The foregoing merely illustrates the principles of the disclosed subject matter. Various modifications and alterations to the described embodiments will be apparent to those skilled in the art in view of the teachings herein. Further, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes, and can not have been selected to delineate or circumscribe the inventive subject matter. Accordingly, the disclosure herein is intended to be illustrative, but not limiting, of the scope of the disclosed subject matter, which is set forth in the following claims.

CLAIMS

1. A computer-based method for labeling multimedia objects comprising:

storing in one or more memories multimedia data;

presenting selected multimedia data to a first user;

determining user interest in said selected multimedia data based upon brain signal responses of the first user;

generating selected multimedia label data corresponding to the selected multimedia data based on the determination of user interest;

using a processor and based on at least said generated selected multimedia label data, performing at least one of refining multimedia label data associated with said selected multimedia data or predicting new multimedia label data pertaining to said stored multimedia data by calculating a classification function.

2. The method of claim 1, further comprising

storing a multimedia affinity graph in said one or more memories, wherein said affinity graph represents multimedia data samples as nodes and comprises edges measuring relatedness among data samples; and

calculating a classification function based on at least the selected multimedia label data using a processor associated with said one or more memories, wherein calculating said classification function comprises iteratively performing at least updating selected multimedia label data relating to selected multimedia data or predicting new multimedia label data for stored multimedia data using said processor.

3. The method of claim 1, wherein generating selected multimedia label data comprises decoding the brain signal responses of the user and generating at least one interest score for said selected multimedia data.

4. The method of claim 3, further comprising refining said interest score and generating an updated interest measurement.

5. The method of claim 3, further comprising converting said at least one interest score to at least one interest label.

6. The method of claim 5, wherein converting said at least one interest score to at least one interest label comprises using a binarization function.

7. The method of claim 1, wherein the multimedia data comprises image data.

8. The method of claim 1, wherein the multimedia data comprises video data.

9. The method of claim 1, wherein the multimedia data comprises audio data.

10. The method of claim 1, wherein predicting new multimedia label data comprises automatically selecting a most informative data sample, predicting its corresponding class and labeling the corresponding data sample.

11. The method of claim 1, wherein refining at least a portion of said multimedia label data comprises performing a greedy search among the gradient direction of the classification function.

12. The method of claim 2, wherein at least one multimedia label data sample is further normalized based on a regularization matrix calculated using members of a corresponding class and connectivity degrees of the corresponding nodes in the graph.

13. The method of claim 2, wherein calculating a classification function comprises incremental calculation using graph superposition, wherein a newly added label is incorporated incrementally without calculating a classification function using all labels.

14. The method of claim 1 wherein noisy label data is replaced.

15. The method of claim 14, wherein replacing noisy label data comprises adding a multimedia label data sample for every multimedia label data sample that is removed.

16. The method of claim 2 wherein noisy label data is replaced.

17. The method of claim 16, wherein replacing noisy label data comprises adding a multimedia label data sample for every multimedia label

18. The method of claim 16, wherein replacing noisy label data or predicting new multimedia label data comprise updating a node regularization matrix.

19. The method of claim 16, wherein replacing noisy label data or predicting new multimedia label data comprises minimizing an objective function.

20. The method of claim 1, comprising presenting said selected multimedia data to said first user one or more additional times and determining user interest in said selected multimedia data one or more additional times to further refine said multimedia label data.

21. The method of claim 1, further comprising:

presenting said selected multimedia data to a second user,

determining user interest in said selected multimedia data based on at least brain signal responses of the first user and brain signal responses of the second user.

22. The method of claim 1, further comprising using said refined or predicted multimedia label data to perform a search of said stored multimedia data.

23. A system for labeling multimedia data comprising:

one or more memories storing multimedia data;

one or more processors coupled to said one or more memories, wherein said one or more processors are configured to:

present selected multimedia data to a first user;

determine user interest in said selected multimedia data based upon brain signal responses of the first user;

generate selected multimedia label data corresponding to the selected multimedia data based on the determination of user interest;

based on at least said generated selected multimedia label data, perform at least one of refining multimedia label data associated with said selected multimedia data or predicting new multimedia label data pertaining to said stored multimedia data by calculating a classification function.

24. The system of claim 23, wherein said one or more processors are further configured to:

store a multimedia affinity graph in said one or more memories, wherein said affinity graph represents multimedia data samples as nodes and comprises edges measuring relatedness among data samples; and

calculate a classification function based on at least the selected multimedia label data, wherein calculating said classification function comprises iteratively performing at least updating selected multimedia label data relating to selected multimedia data or predicting new multimedia label data for stored multimedia data.

25. The system of claim 23, wherein generating selected multimedia label data comprises decoding the brain signal responses of the user and generating at least one interest score for said selected multimedia data.

26. The system of claim 25, wherein said one or more processors are further configured to refine said interest score and generate an updated interest measurement.

27. The system of claim 25, wherein said one or more processors are further configured to convert said at least one interest score to at least one interest label.

28. The system of claim 27, wherein converting said at least one interest score to at least one interest label comprises using a binarization function.

29. The system of claim 23, wherein the multimedia data comprises image data.

30. The system of claim 23, wherein the multimedia data comprises video data.

31. The system of claim 23, wherein the multimedia data comprises audio data.

32. The system of claim 23, wherein predicting new multimedia label data comprises automatically selecting a most informative data sample, predicting its corresponding class and labeling the corresponding data sample.

33. The system of claim 23, wherein refining at least a portion of said multimedia label data comprises performing a greedy search among the gradient direction of the classification function.

34. The system of claim 24, wherein said one or more processors are further configured to normalize at least one multimedia label data sample on a regularization matrix calculated using members of a corresponding class and connectivity degrees of the corresponding nodes in the graph.

35. The system of claim 24, wherein calculating a classification function comprises incremental calculation using graph superposition, wherein a newly added label is incorporated incrementally without calculating a classification function using all labels.

36. The system of claim 23 wherein noisy label data is replaced.

37. The method of claim 14, wherein replacing noisy label data comprises adding a multimedia label data sample for every multimedia label data sample that is removed.

38. The method of claim 2 wherein noisy label data is replaced.

39. The system of claim 38, wherein replacing noisy label data comprises adding a multimedia label data sample for every multimedia label

40. The system of claim 38, wherein replacing noisy label data or predicting new multimedia label data comprise updating a node regularization matrix.

41. The system of claim 38, wherein replacing noisy label data or predicting new multimedia label data comprises minimizing an objective function.

42. The system of claim 23, wherein said one or more processors are further configured to present said selected multimedia data to said first user one or

more additional times and determine user interest in said selected multimedia data one or more additional times to further refine said multimedia label data.

43. The system of claim 23, wherein said one or more processors are further configured to:

present said selected multimedia data to a second user,

determine user interest in said selected multimedia data based on at least brain signal responses of the first user and brain signal responses of the second user.

44. The system of claim 23, wherein said one or more processors are further configured to use said refined or predicted multimedia label data to perform a search of said stored multimedia data.

45. A computer readable medium containing digital information which when executed cause a processor or processors to:

store in one or more memories multimedia data;

present selected multimedia data to a first user;

determine user interest in said selected multimedia data based upon brain signal responses of the first user;

generate selected multimedia label data corresponding to the selected multimedia data based on the determination of user interest;

based on at least said generated selected multimedia label data, perform at least one of refining multimedia label data associated with said selected multimedia data or predicting new multimedia label data pertaining to said stored multimedia data by calculating a classification function.

46. The computer readable medium of claim 45 containing digital information which when executed further causes the processor or processors to:

store a multimedia affinity graph in said one or more memories, wherein said affinity graph represents multimedia data samples as nodes and comprises edges measuring relatedness among data samples; and

calculate a classification function based on at least the selected multimedia label data, wherein calculating said classification function comprises iteratively performing at least updating selected multimedia label data relating to

selected multimedia data or predicting new multimedia label data for stored multimedia data.

47. The computer readable medium of claim 45, wherein generating selected multimedia label data comprises decoding the brain signal responses of the user and generating at least one interest score for said selected multimedia data.

48. The computer readable medium of claim 47 containing digital information which when executed further causes the processor or processors to refine said interest score and generate an updated interest measurement.

49. The computer readable medium of claim 47 containing digital information which when executed further causes the processor or processors to convert said at least one interest score to at least one interest label.

50. The computer readable medium of claim 49, wherein converting said at least one interest score to at least one interest label comprises using a binarization function.

51. The computer readable medium of claim 45, wherein the multimedia data comprises image data.

52. The computer readable medium of claim 45, wherein the multimedia data comprises video data.

53. The computer readable medium of claim 45, wherein the multimedia data comprises audio data.

54. The computer readable medium of claim 45, wherein predicting new multimedia label data comprises automatically selecting a most informative data sample, predicting its corresponding class and labeling the corresponding data sample.

55. The computer readable medium of claim 45, wherein refining at least a portion of said multimedia label data comprises performing a greedy search among the gradient direction of the classification function.

56. The computer readable medium of claim 46, wherein at least one multimedia label data sample is further normalized based on a regularization matrix calculated using members of a corresponding class and connectivity degrees of the corresponding nodes in the graph.

57. The computer readable medium of claim 46, wherein calculating a classification function comprises incremental calculation using graph superposition, wherein a newly added label is incorporated incrementally without calculating a classification function using all labels.

58. The computer readable medium of claim 45 wherein noisy label data is replaced.

59. The method of claim 14, wherein replacing noisy label data comprises adding a multimedia label data sample for every multimedia label data sample that is removed.

60. The computer readable medium of claim 46 wherein noisy label data is replaced.

61. The computer readable medium of claim 60, wherein replacing noisy label data comprises adding a multimedia label data sample for every multimedia label

62. The computer readable medium of claim 60, wherein replacing noisy label data or predicting new multimedia label data comprise updating a node regularization matrix.

63. The computer readable medium of claim 60, wherein replacing noisy label data or predicting new multimedia label data comprises minimizing an objective function.

64. The computer readable medium of claim 45 containing digital information which when executed further causes the processor or processors to present said selected multimedia data to said first user one or more additional times and determine user interest in said selected multimedia data one or more additional times to further refine said multimedia label data.

65. The computer readable medium of claim 45 containing digital information which when executed further causes the processor or processors to:

present said selected multimedia data to a second user,

determine user interest in said selected multimedia data based on at least brain signal responses of the first user and brain signal responses of the second user.

66. The computer readable medium of claim 45 containing digital information which when executed further causes the processor or processors to use said refined or predicted multimedia label data to perform a search of said stored multimedia data.

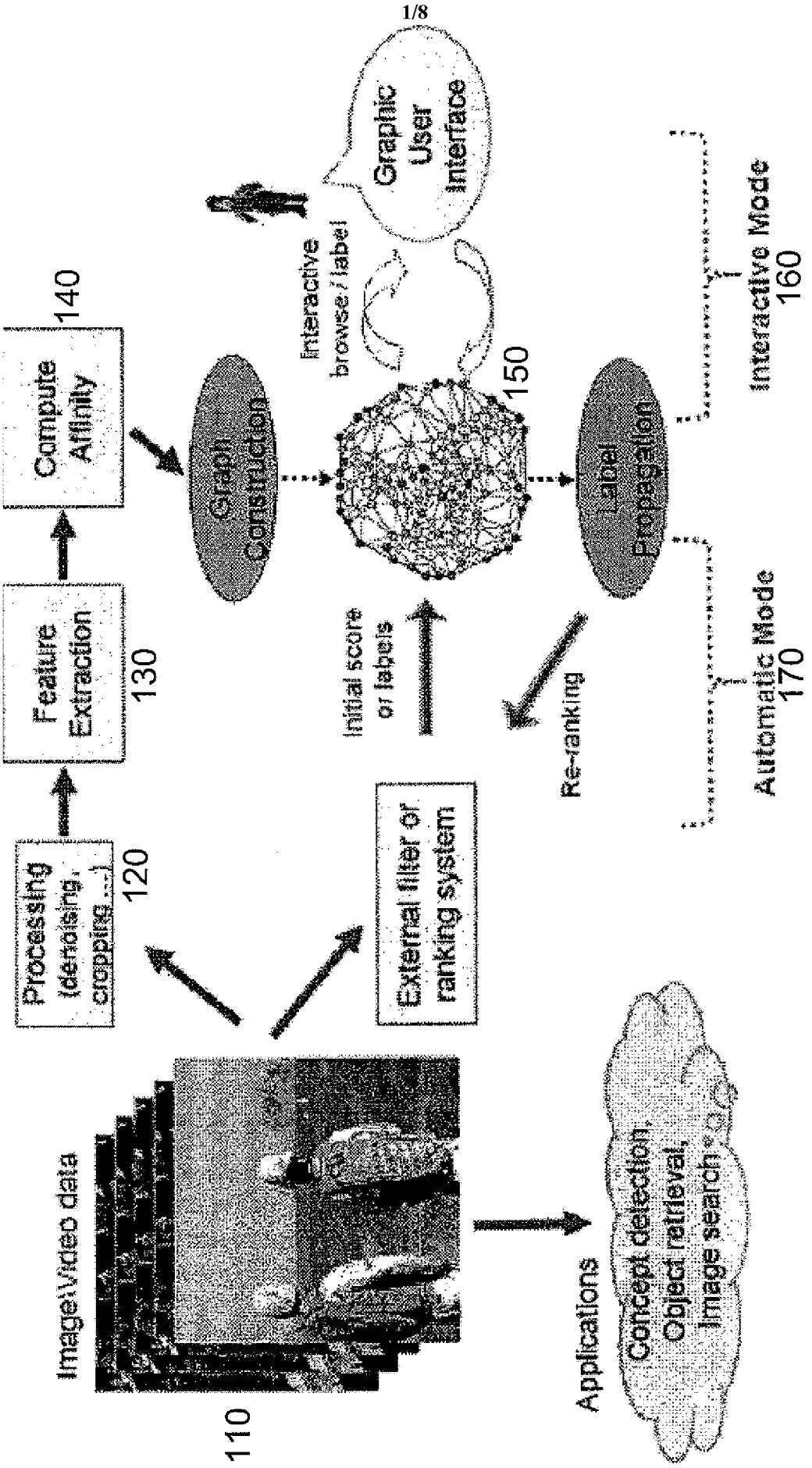
Image\Video data
110
Processing (denoising, cropping ...)
120
Feature Extraction
130
Compute Affinity
140
External filter or ranking system
Graph Construction
Interactive browse / label
Graphic User Interface
150
Initial score or labels
Re-ranking
Label Propagation
Interactive Mode
160
Automatic Mode
170
Applications
Concept detection,
Object retrieval,
Image search ...

FIGURE 1

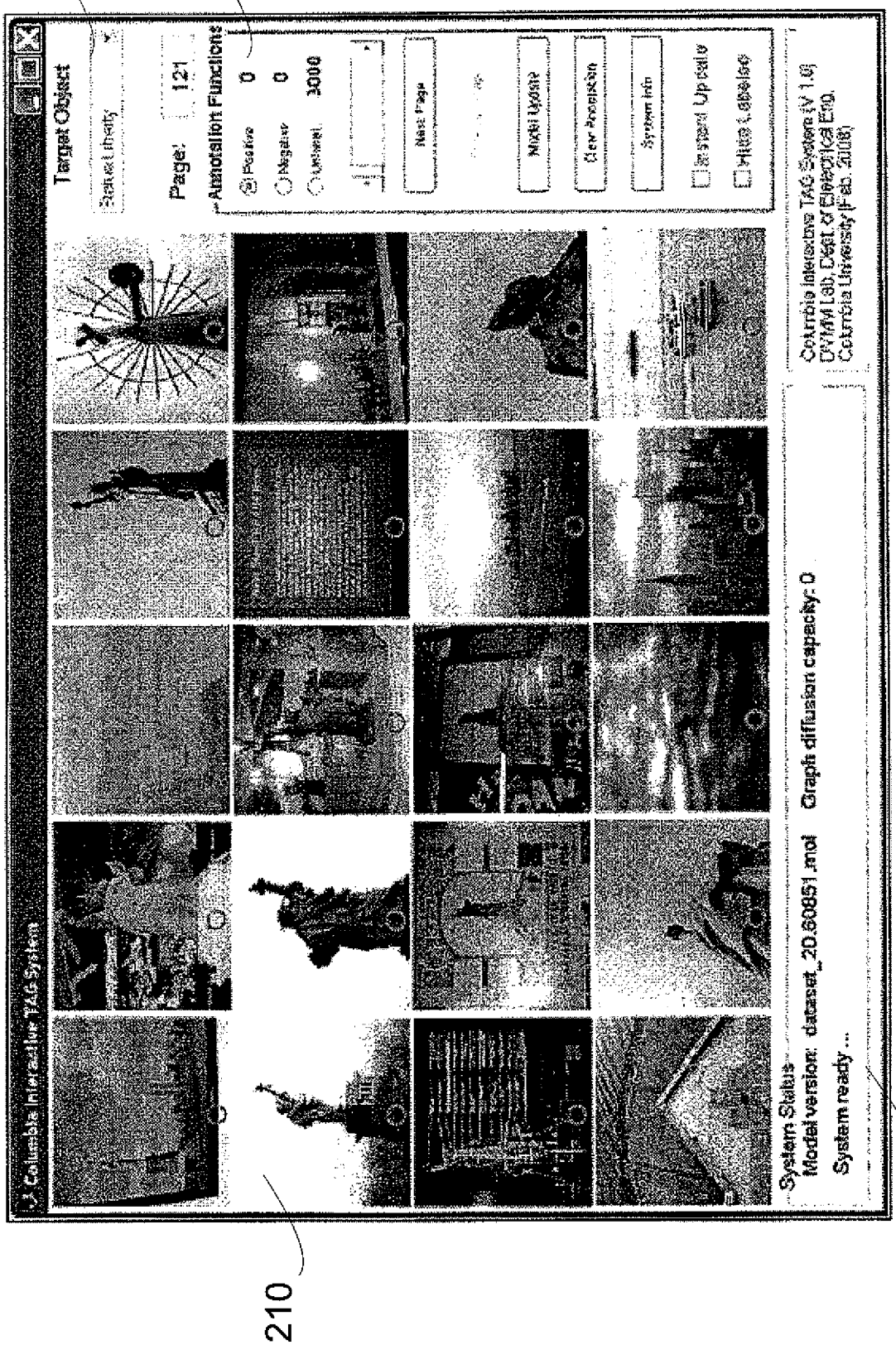
230
240
210
220
Target Object
Page: 121
Annotation Functions
Positive 0
Negative 0
Unlabeled 3000
Next Page
Model Update
Clear Annotation
System Info
Columbia Interactive TAG System (V 1.0)
DVMM Lab, Dept. of Electrical Eng.
Columbia University (Feb. 2008)
System Status
Model version: dataset_20.60851.mod
Graph diffusion capacity: 0
System ready ...

FIGURE 2

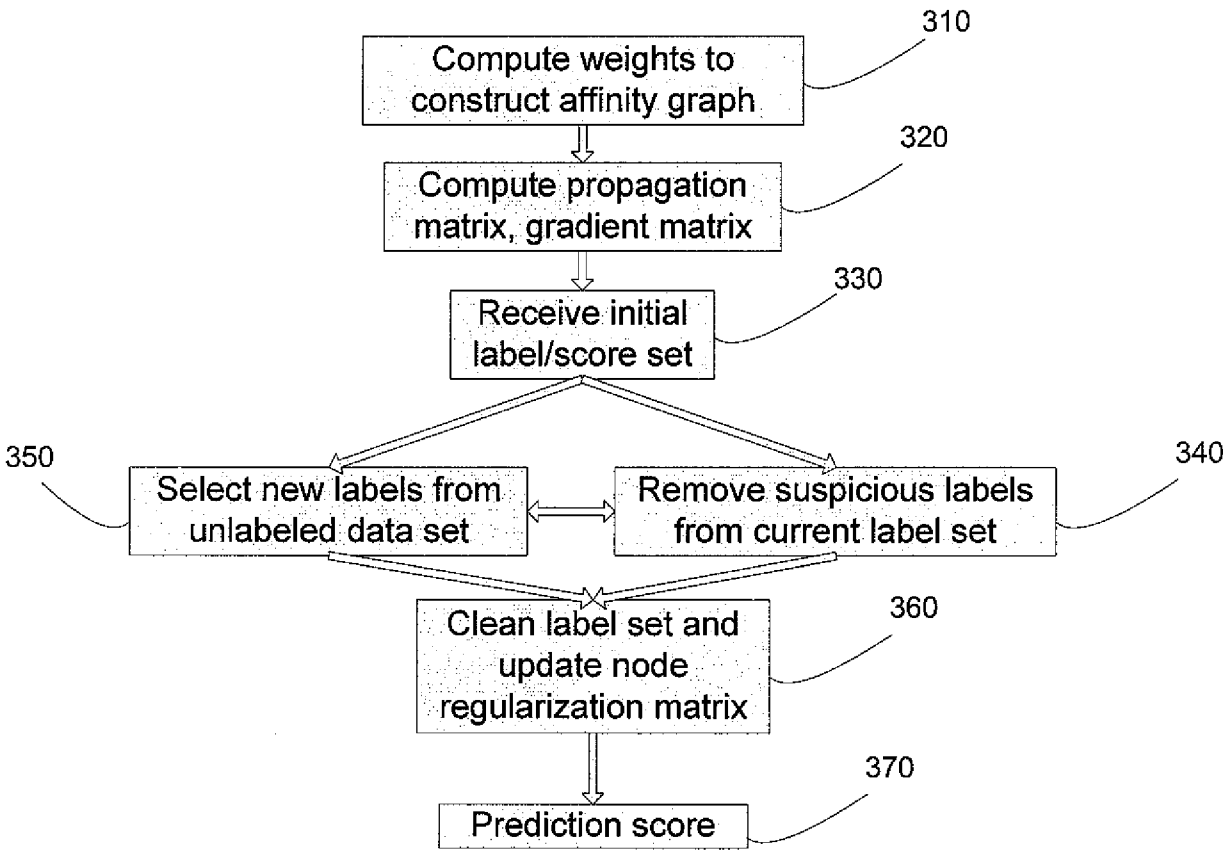
Compute weights to construct affinity graph
310
Compute propagation matrix, gradient matrix
320
Receive initial label/score set
330
Select new labels from unlabeled data set
350
Remove suspicious labels from current label set
340
Clean label set and update node regularization matrix
360
Prediction score
370

FIGURE 3

W =
Weight matrix
D =
Node Degree matrix
Y =
Label matrix
V =
Node regularizer
A fraction of constructed graph
x_1
x_2
x_3
x_4
x_5

FIGURE 4

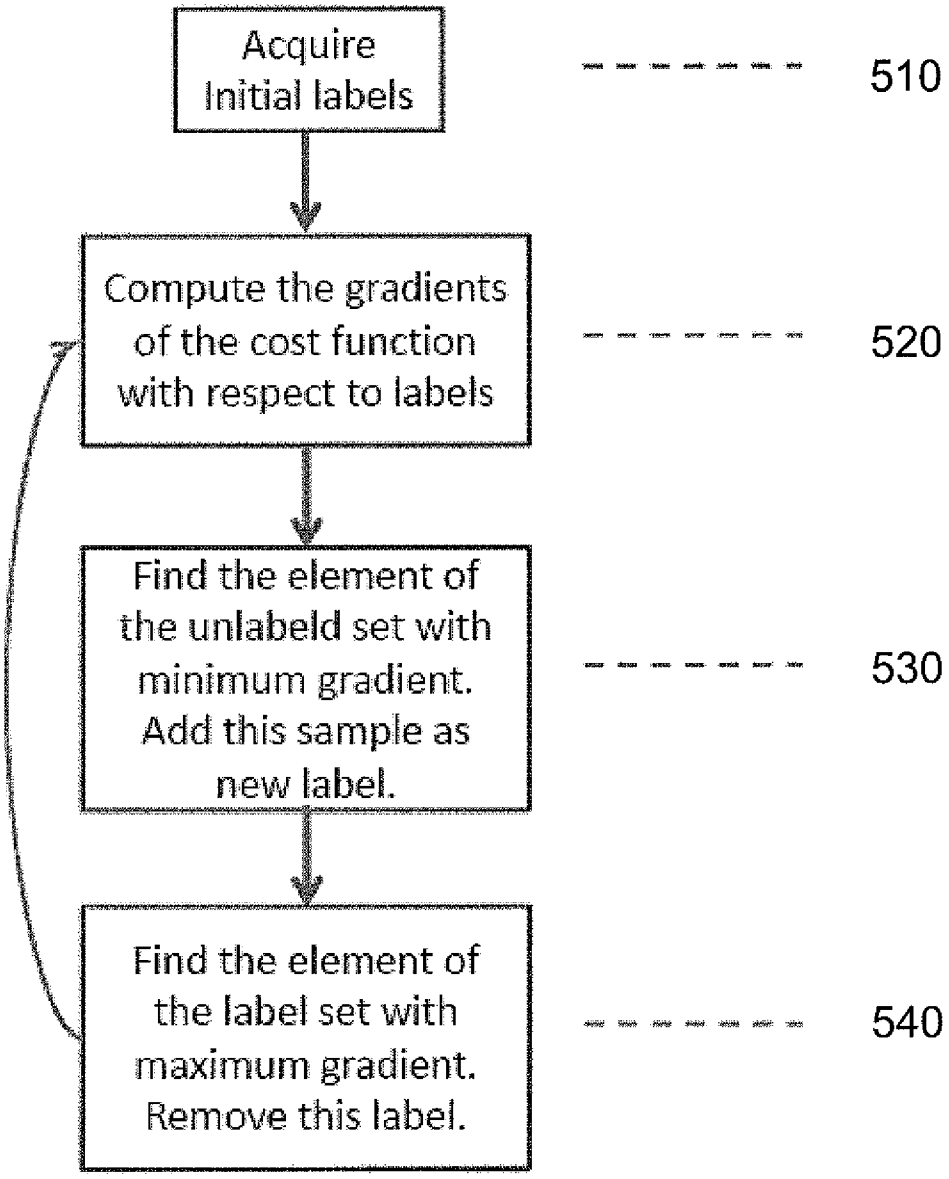
Acquire Initial labels
510
Compute the gradients of the cost function with respect to labels
520
Find the element of the unlabeld set with minimum gradient. Add this sample as new label.
530
Find the element of the label set with maximum gradient. Remove this label.
540

FIGURE 5

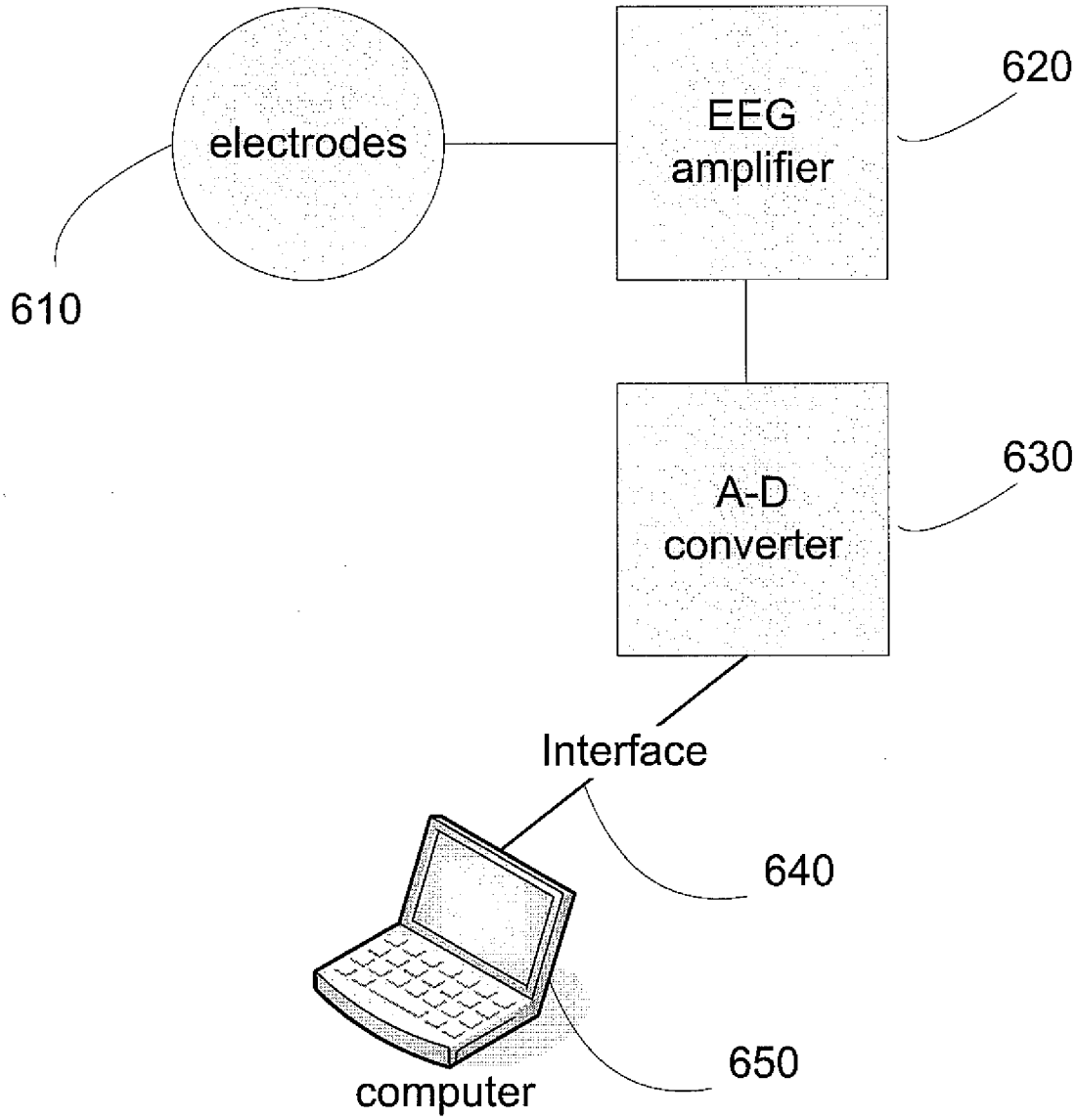
electrodes
610
EEG amplifier
620
A-D converter
630
Interface
640
computer
650

FIGURE 6

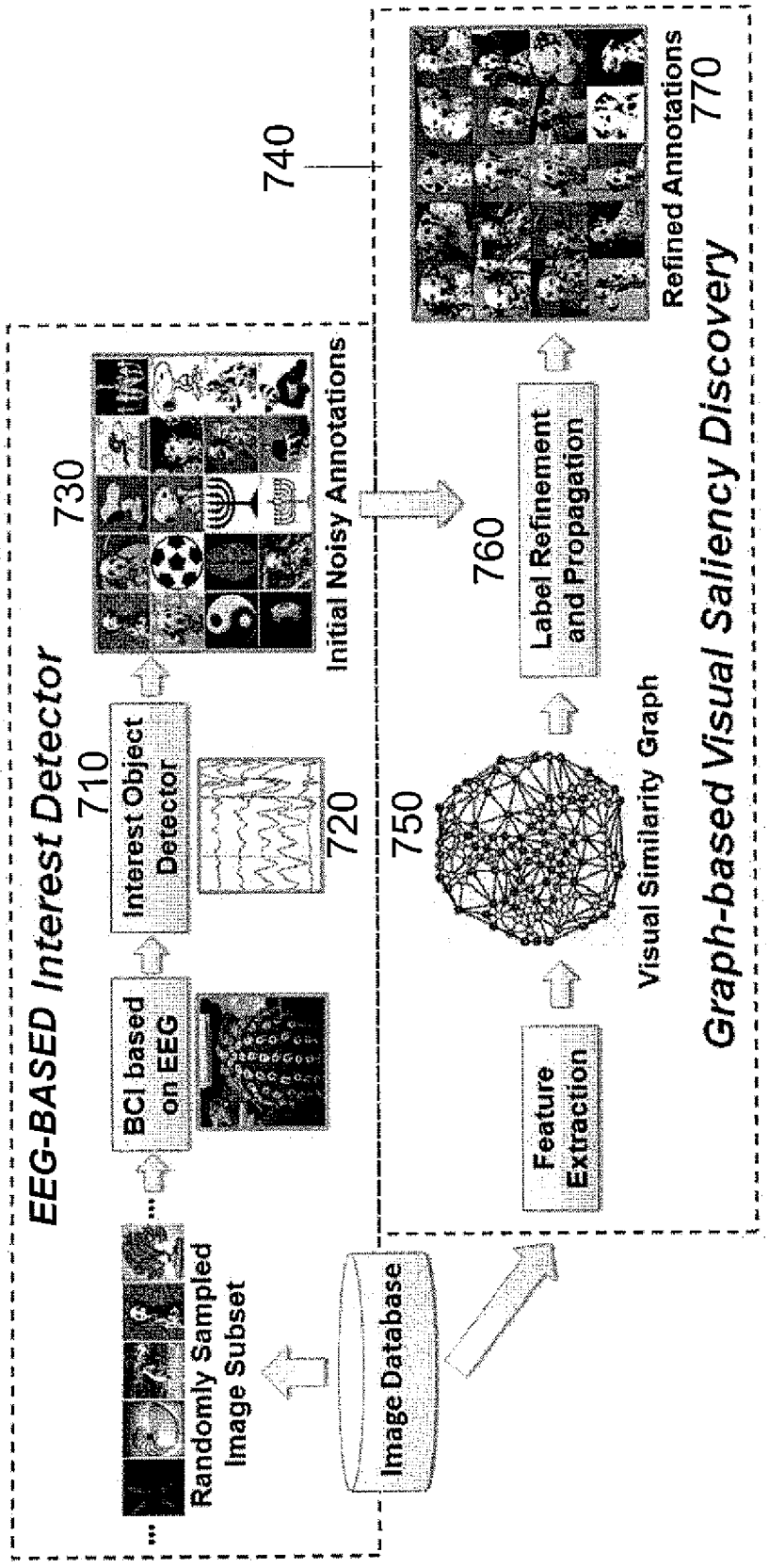
EEG-BASED Interest Detector
Graph-based Visual Saliency Discovery
Randomly Sampled Image Subset
Image Database
BCI based on EEG
Interest Object Detector
710
720
730
Initial Noisy Annotations
740
Feature Extraction
750
Visual Similarity Graph
Label Refinement and Propagation
760
Refined Annotations
770

FIGURE 7

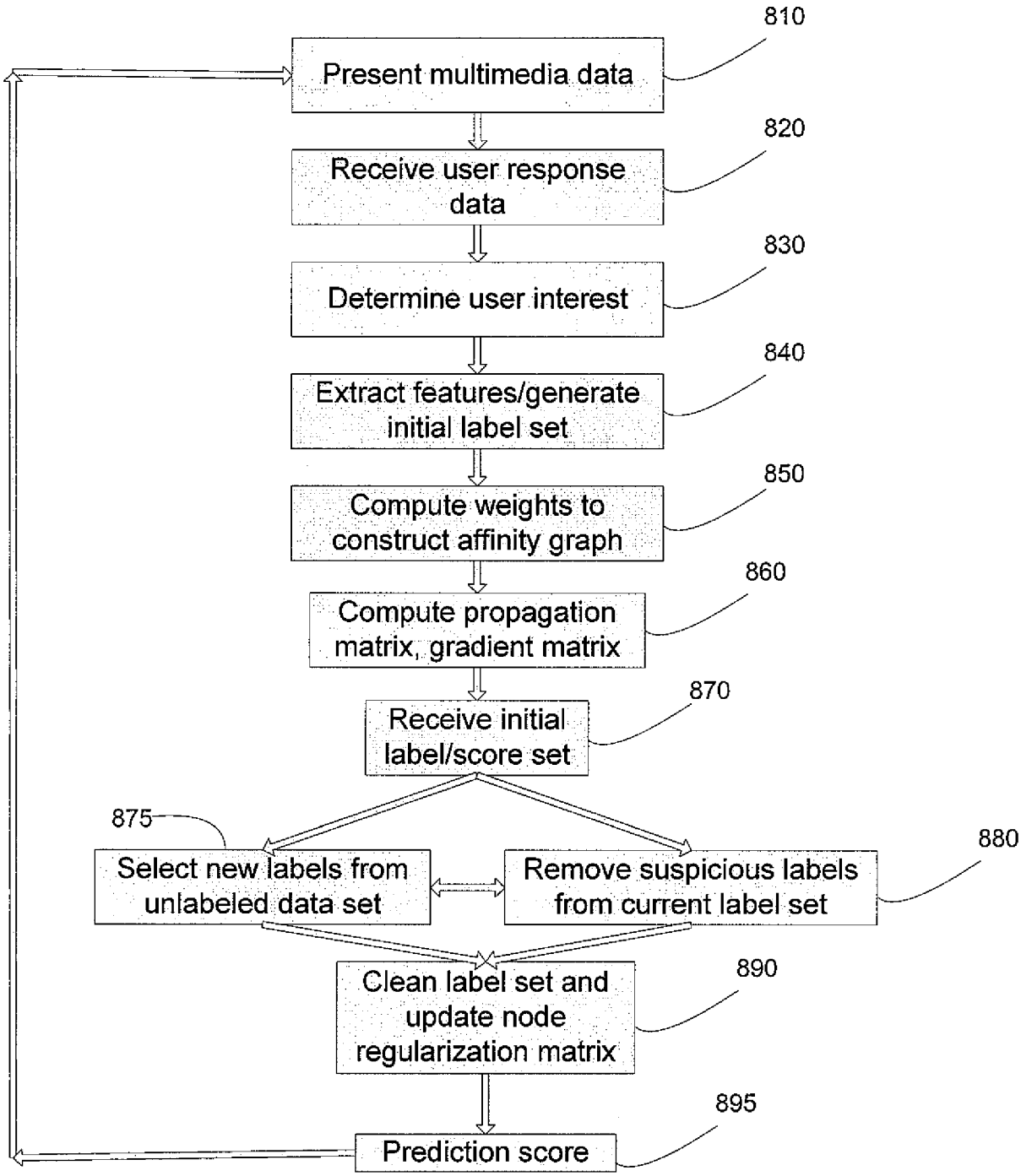
810
Present multimedia data
820
Receive user response data
830
Determine user interest
840
Extract features/generate initial label set
850
Compute weights to construct affinity graph
860
Compute propagation matrix, gradient matrix
870
Receive initial label/score set
875
Select new labels from unlabeled data set
880
Remove suspicious labels from current label set
890
Clean label set and update node regularization matrix
895
Prediction score

FIGURE 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/23494

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - H04H 60/33 (2010.01)
USPC - 725/10
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC - 725/10

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 434/322, 350; 725/10, 12

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Dialog Classic (Chinese Pat Abstr; Derwent Index, EPFT, French Pat, Jap Abstr, USPFT, WIPO/PCT PFT); Google Scholar; Terms searched: 20080222670, AFFINIT, AFFINITY GRAPH, ANOTHER, AUDIO, BINAR, BRAIN, CEREB, CLASS, CLASSIF, CLASSIFICATION, CLASSIFICATION FUNCTION, CLASSIFIES, CLASSIFY, COMPUTER VISION, DATABASE, DEGREE, EDGE...

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	SAJDA, et al "In a blink of an eye and a switch of a transistor: Cortically-coupled computer vision", Journal of Latex Class Files, Vol 6, No. 1 (Sajda et al.), January 2007 (01.2007); entire document, especially: p 2, col 2, para 3; p 4, fig 1 and col 1, para 4; p. 5, col 1, para 2-4; p 6, fig 2; p 7, col 2, para 2 and 3; p 8, col 1, para 1 to p 8, col 2, para 1; p 8, col 1, para 4 to p 9, col 1, para 2; p 9, col 2, para 3 to p. 10, col 1, para 1; p, 12, col 1, para 1 and 2; p 12, col 2, para 1	1-12, 14-18, 20-34, 36-40, 42-56, 58-62, 64-66 --- 13, 19, 35, 41, 57, 63
Y	WANG, et al "Columbia TAG System - Transductive Annotation by Graph Version 1.0", Columbia University ADVENT Technical Report #225-2008-3 (Wang et al.) 15 October 2008 (15.10.2008); entire document, especially: p 4, para 2 and 4; p 6, para 5	13, 35, 57
Y	US 2003/0046018 A1 (Kohlmorgen et al.) 06 March 2003 (06.03.2003); entire document, especially: para [0040]	19, 41, 63
A	US 2006/0293921 A1 (McCarthy et al.) 28 December 2006 (28.12.2006); entire document	1-66
A	US 2008/0222670 A1 (Lee et al.) 11 September 2008 (11.09.2008); entire document	1-66

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 March 2010 (16.03.2010)	01 APR 2010
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)